



UNIVERSIDADE FEDERAL DO ACRE
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

**MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO DE CASO APLICADO A
RETENÇÃO E EVASÃO DO CURSO DE BACHARELADO EM SISTEMAS DE
INFORMAÇÃO DA UFAC**

RIO BRANCO

2021

ANDRIELLE DE LIMA BEZERRA

**MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO DE CASO APLICADO A
RETENÇÃO E EVASÃO DO CURSO DE BACHARELADO EM SISTEMAS DE
INFORMAÇÃO DA UFAC**

Monografia apresentada como exigência
parcial para obtenção do grau de
bacharel em Sistemas de Informação da
Universidade Federal do Acre.

Orientador: Manoel Limeira de Lima
Júnior

RIO BRANCO

2021

TERMO DE APROVAÇÃO

ANDRIELLE DE LIMA BEZERRA

MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO DE CASO APLICADO A RETENÇÃO E EVASÃO DO CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO DA UFAC

Esta monografia foi apresentada como trabalho de conclusão de Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, sendo aprovada pela banca constituída pelo professor orientador e membros abaixo mencionados.

Compuseram a banca:



Prof. Manoel Limeira de Lima Júnior, Doutor
Curso de Bacharelado em Sistemas de Informação

Prof. Daricélio Moreira Soares, Doutor
Curso de Bacharelado em Sistemas de Informação

Prof. Claudionor Alencar do Nascimento, Mestre
Curso de Bacharelado em Sistemas de Informação

Rio Branco - AC, 29 de junho de 2021

*Dedico este trabalho a Deus e à minha
família, em especial minha mãe.*

AGRADECIMENTOS

Sou grata primeiramente a Deus, por sua infinita misericórdia em minha vida, por sempre mostrar graciosidade, mesmo não sendo merecedora e por todas as incontáveis bênçãos, sendo uma delas a oportunidade de cursar o ensino superior.

Em segundo lugar, sou grata a minha mãe, Maria Ornelia Campos de Lima por sempre me incentivar em todos os meus sonhos e me apoiar independente da circunstância. Também sou grata ao restante da minha amada família, juntamente com minha família em Cristo a Igreja Batista Regular Maranata, agradeço a ambas por todo apoio que direto e indiretamente, contribuíram de alguma forma para que conseguisse concluir mais uma etapa da minha jornada.

Da mesma forma sou grata pelos amigos que tive o privilégio de conhecer e reencontrar na graduação, especialmente os da minha turma de 2017, esses anos foram muito mais agradáveis porque tive a oportunidade de estar rodeada por bons amigos.

Agradeço também a todos os meus professores, por todo conhecimento repassado e por contribuírem diretamente para minha formação acadêmica, e agradeço de forma especial ao meu orientador Manoel Limeira de Lima Júnior, por toda orientação e ajuda para o desenvolvimento deste trabalho. E por fim, agradeço ao Núcleo de Tecnologia da Informação (NTI) por fornecer os dados necessários para realização desta pesquisa.

“Tudo quanto fizerdes, fazei-o de todo o coração, como para o Senhor e não para homens.”

Colossenses 3:23

RESUMO

Nos últimos 10 anos, o curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre apresentou uma média anual de 3,6% de alunos que concluíram no tempo ideal (4 anos) e de 25,8% de alunos que concluíram com atraso (no total, 29,4% alunos formados), isso mostra um grande índice de evasão e de retenção. Esta pesquisa utilizou a Mineração de Dados Educacionais para analisar a problemática relativa à retenção e à evasão do curso de Bacharelado em Sistemas de Informação. O objetivo central deste trabalho foi construir modelos utilizando algoritmos de classificação para previsão da retenção e da evasão dos estudantes, a partir da base de dados de histórico de alunos fornecida pelo Núcleo de Tecnologia da Informação da UFAC, com dados referentes aos anos de 2008 até 2019. A metodologia utilizada baseou-se nas etapas do processo de descoberta de conhecimento, também conhecido como *Knowledge Discovery in Databases (KDD)*, que consiste nas etapas de seleção, pré-processamento, transformação, mineração de dados e interpretação e avaliação. Os algoritmos *Naive Bayes*, *SMO*, *IBK*, *J48* e *Random Forest* foram aplicados e os resultados obtidos mostraram que o algoritmo que obteve os melhores resultados foi o *Random Forest*, pois alcançou a média de acurácia mais alta em todos experimentos realizados e pôde ser melhorado por meio da técnica de seleção de atributos. Os experimentos realizados mostraram resultados com uma alta taxa de acurácia, nas previsões de retenção, o *Random Forest* alcançou a média de 87,11% e 71,21% e nas previsões sobre evasão obteve uma média de 87,12% e 82,57% de acurácia. Além disso, a utilização do método de seleção de atributos possibilitou uma melhoria na acurácia para 87,76% e 73,29% para os experimentos da retenção e 87,89% e 83,76% para os experimentos da evasão.

Palavras-chave: Mineração de Dados Educacionais. *Knowledge Discovery in Databases (KDD)*. Classificação. WEKA. Retenção. Evasão.

ABSTRACT

In the last 10 years, the Baccalaureate Course in Federal University of Acre Information Systems presented an annual average of 3.6% of students who concluded at ideal time (4 years) and 25.8% of students who concluded with a delay (in total , 29.4% graduated students), this shows a great index of evasion and retention. This research used the mining of educational data to analyze the problematic relative to the retention and evasion of the bachelor's degree in information systems. The central objective of this work was to build models using classification algorithms for estimating retention and student evasion, from the student history database provided by the UFAC Information Technology Center, with data for the years 2008 until 2019. The methodology used was based on the stages of the knowledge discovery process, also known as Knowledge Discovery in Databases (KDD), which consists of the selection steps, pre-processing, transformation, data mining and interpretation and evaluation. The Naive Bayes, SMO, IBK, J48 and Random Forest algorithms were applied and the results obtained showed that the algorithm that obtained the best results was Random Forest, because it reached the highest accuracy average in all experiments carried out and could be improved by The attribute selection technique. The experiments performed showed results with a high accuracy rate, in retention predictions, Random Forest reached an average of 87.11% and 71.21% and preview predictions obtained an average of 87.12% and 82.57 % of accuracy. In addition, the use of the attribute selection method allowed an accuracy improvement to 87.76% and 73.29% for retention experiments and 87.89% and 83.76% for evasion experiments.

Keywords: Educational Data Mining. *Knowledge Discovery in Databases (KDD)*. Classification. WEKA. Retention. Dropout.

LISTA DE FIGURAS

Figura 1 - Relação de Concluintes e Retidos/Evadidos do Curso de Sistemas de Informação por ano.....	16
Figura 2 - Metodologia do trabalho.....	21
Figura 3 - Etapas do processo do KDD	24
Figura 4 - Divisão da base de dados utilizando validação cruzada	27
Figura 5 - Principais áreas da mineração de dados educacionais.....	33
Figura 6 - Possíveis caminhos do estudante na graduação.....	36
Figura 7 - Divisão da base de dados.....	48
Figura 8 - Instâncias de um estudante em cada período.....	49
Figura 9 - Instâncias em cada período do experimento da retenção e evasão..	49
Figura 10 - Instâncias das disciplinas referente a um estudante.....	50
Figura 11 - Instâncias em cada período do experimento da situação da disciplina.....	50

LISTAS DE QUADROS

Quadro 1 - Matriz de confusão.....	28
Quadro 2 - Mapeamentos dos atributos dos trabalhos relacionados.....	38
Quadro 3 - Atributos disponibilizados.....	41
Quadro 4 - Todos os atributos da base de dados.....	43
Quadro 5 - Atributos selecionados.....	46
Quadro 6 - Atributos mais relevantes e frequentes por experimento.....	68

LISTAS DE TABELAS

Tabela 1 - Despesa total por aluno da Universidade Federal do Acre.....	18
Tabela 2 - Estatística de ingresso e egresso por turma.....	20
Tabela 3 - TGS do curso de Sistemas de Informação.....	35
Tabela 4 - Acurácia dos algoritmos do 1º experimento.....	52
Tabela 5 - Seleção de atributos do 1º experimento.....	53
Tabela 6 - Acurácia dos algoritmos do 2º experimento.....	55
Tabela 7 - Seleção de atributos do 2º experimento.....	56
Tabela 8 - Acurácia dos algoritmos do 3º experimento.....	58
Tabela 9 - Seleção de atributos do 3º experimento.....	59
Tabela 10 - Acurácia dos algoritmos do 4º experimento.....	61
Tabela 11 - Seleção de atributos do 4º experimento.....	62
Tabela 12 - Acurácia dos algoritmos do 5º experimento.....	64
Tabela 13 - Seleção de atributos do 5º experimento.....	65
Tabela 14 - Médias dos resultados da seleção de atributos do <i>Random Forest</i>	68

SUMÁRIO

1 INTRODUÇÃO	14
1.2 OBJETIVOS DA PESQUISA	17
1.2.1 Objetivo geral	17
1.2.2 Objetivos específicos	17
1.3 JUSTIFICATIVA DA PESQUISA	18
1.4 METODOLOGIA	20
1.5 ORGANIZAÇÃO DO ESTUDO	22
2 FUNDAMENTAÇÃO TEÓRICA	23
2.1 KNOWLEDGE DISCOVERY IN DATABASES	23
2.1.1 Classificação	25
2.1.2 Métricas de avaliação da classificação	27
2.1.3 Algoritmos de classificação	30
2.2 SOFTWARE PARA MINERAÇÃO DE DADOS	31
2.3 MINERAÇÃO DE DADOS EDUCACIONAIS	32
2.3.1 Diplomação	34
2.3.2 Retenção	35
2.3.3 Evasão	35
2.4 TRABALHOS RELACIONADOS	36
3 PREVISÃO DE EVASÃO E RETENÇÃO NO CURSO DE SISTEMAS DE INFORMAÇÃO	40
3.1 PLANEJAMENTO DO ESTUDO	40
3.1.2 Seleção e pré-processamento dos dados	41
3.1.2 Transformação dos dados	42
3.1.2 Atributos utilizados	46
3.1.2 Divisão dos experimentos	47

3.2 DISCUSSÃO E AVALIAÇÃO DOS RESULTADOS	51
3.2.1 Retenção	51
3.2.2 Evasão	57
3.2.2 Situação da disciplina	64
3.3 CONSIDERAÇÕES SOBRE O CAPÍTULO	67
4 CONSIDERAÇÕES FINAIS	70
4.1 CONTRIBUIÇÕES	70
4.2 LIMITAÇÕES DO ESTUDO	72
4.3 RECOMENDAÇÕES	72
REFERÊNCIAS	73

1 INTRODUÇÃO

A Universidade Federal do Acre (UFAC), assim como as demais Instituições Federais de Ensino Superior (IES), desenvolvem um papel importante no contexto regional, tendo como atribuições o fortalecimento da pesquisa e preparação dos estudantes, a fim de inserir profissionais de diversas áreas no mercado, com mão de obra qualificada para que possam atuar e contribuir para o desenvolvimento socioeconômico, principalmente do estado do Acre.

No entanto, a retenção e a evasão dos estudantes no ensino superior têm sido um dos fatores que interferem em um melhor resultado na preparação destes estudantes ao mercado de trabalho. A retenção é caracterizada como a permanência prolongada da graduação, aumentando o tempo médio para finalizar o curso, ou seja, o atraso do período de conclusão do curso (UNIVERSIDADE FEDERAL FLUMINENSE, 2016). Já a evasão é caracterizada, resumidamente, como a saída do estudante antecipadamente do curso, sem a devida integralização do mesmo.

Dessa maneira, com o crescente aumento do uso da tecnologia e a possibilidade de armazenar grandes quantidades de dados dos alunos, é factível a utilização da Mineração de Dados Educacionais (MDE) para melhorar os processos de ensino-aprendizagem. Neste contexto, com o objetivo de auxiliar os tomadores de decisão, tem-se como possibilidade a utilização da Mineração de Dados

Educacionais como ferramenta para extrair informações úteis a partir de dados dos estudantes para realizar a análise de retenção e evasão.

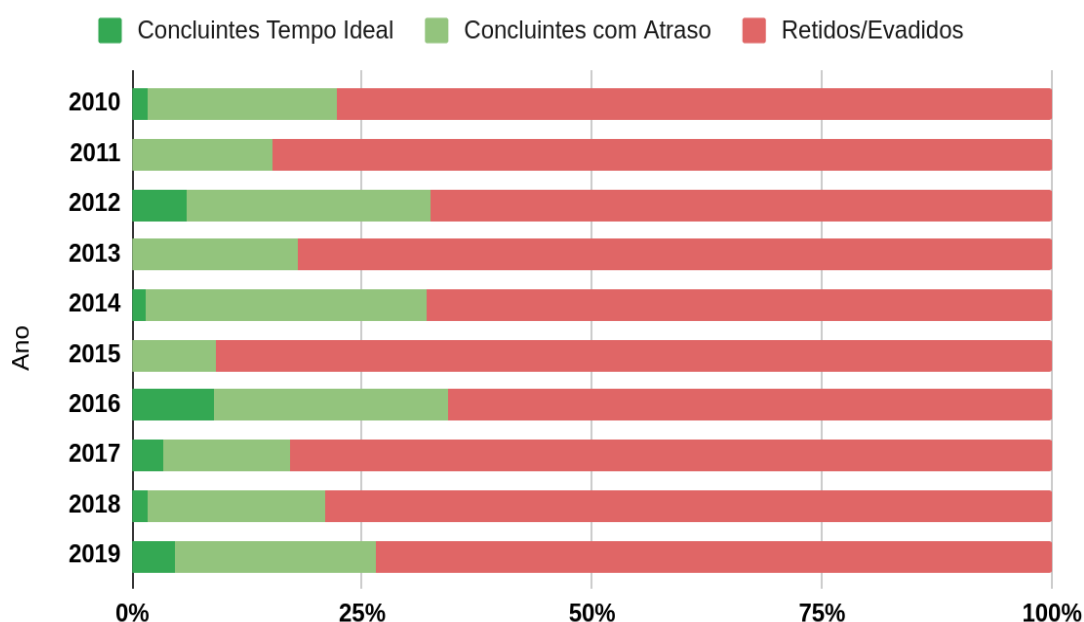
1.1 PROBLEMA DA PESQUISA

Segundo PROPLAN (2019), em uma compilação dos dados divulgados em um documento intitulado “UFAC em Números”, no ano de 2019, a Universidade Federal do Acre matriculou ao total 7.766 alunos na graduação, referente aos 48 cursos ofertados. Nesse mesmo ano, teve 2.512 ingressantes e apenas 1.226 estudantes concluíram sua graduação. O insucesso para se finalizar a graduação pode estar relacionado à desistência do curso ou à dificuldade em conseguir o rendimento acadêmico necessário para obter créditos suficientes para fazer jus à diplomação (MANHÃES, 2015).

A Figura 1, de acordo com dados cedidos pela coordenação, mostra a relação de concluintes e retidos/evadidos por ano do curso de Bacharelado em Sistemas de Informação da referida instituição de ensino, expostos na . Os concluintes foram subdivididos em duas categorias os **concluintes no tempo ideal**, caracterizados como alunos que obtiveram êxito em se graduar no tempo previsto para o curso (em 4 anos), e os **concluintes com atraso**, que são os demais alunos que se formaram naquele mesmo ano, mas que são alunos oriundos de turmas anteriores.

O estudo proposto neste trabalho está relacionado à última categoria da Figura 1, os **retidos/evadidos**, sendo classificados em uma única categoria, pois ambos são referentes aos alunos que não concluíram o curso. Como abordado anteriormente, os retidos são os alunos que não conseguiram se graduar no tempo estipulado, mas permanecem no curso, enquanto os evadidos são aqueles que não fazem mais parte do curso.

Figura 1 - Relação de Concluintes e Retidos/Evadidos do Curso de Sistemas de Informação por ano.



Fonte: Elaboração própria, 2020.

De acordo com o gráfico apresentado na Figura 1, é perceptível que a representação da categoria **retidos/evadidos** é a maior parcela em todos os anos analisados (de 2010 a 2019). Sendo assim, com base nos dados mostrados, é possível afirmar que, nos últimos 10 anos, o curso de Bacharelado em Sistemas de Informação apresentou uma média anual de 3,6% de alunos que concluíram no tempo ideal e 25,8% de alunos que concluíram com atraso. Somando-se as duas porcentagens, têm-se em média de 29,4% alunos formados por ano.

Considerando o restante dos alunos, o curso apresentou em média 70,6% de retenção e de evasão por ano, ou seja, conforme mostram os dados é possível inferir que o curso possui um elevado índice de retenção e de evasão de discentes.

Portanto, o presente trabalho buscou analisar esta problemática considerando o elevado índice de retenção e evasão do curso de Bacharelado em Sistemas de Informação: é possível construir modelos de classificação que possam prever, com base no desempenho acadêmico, a retenção e a evasão dos alunos no curso Bacharelado em Sistemas de Informação?

1.2 OBJETIVOS DA PESQUISA

Nesta seção estão definidos os objetivos desta pesquisa, tanto de forma geral quanto de forma específica.

1.2.1 Objetivo geral

Construir modelos utilizando técnicas e algoritmos de Mineração de Dados que possam prever a evasão e a retenção dos discentes, com uma boa taxa de acerto, identificando as características que INFLUENCIAM do tempo de conclusão ou até mesmo saída do curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre.

1.2.2 Objetivos específicos

- a. Extrair dados a partir da base de dados acadêmicos presentes no Sistemas de Informação para Ensino (SIE), que foi disponibilizada pelo Núcleo de Tecnologia da Informação da UFAC.
- b. Aplicar técnicas e algoritmos de mineração de dados para propor modelos de previsão.
- c. Avaliar os modelos a partir da taxa de acerto.
- d. Identificar as características mais importantes que influenciam nas previsões.

1.3 JUSTIFICATIVA DA PESQUISA

O estudo sobre a retenção e a evasão no ensino superior é de extrema relevância, porque os problemas causados pela mesma podem ser diversos, afetando, principalmente, ao estudante que fica retido ou no pior dos casos se torna evadido. Além disso, diversos outros prejuízos podem ser listados, incluindo os financeiros, tais como o tempo de investimento por cada aluno para a universidade e o governo federal, os recursos de infraestrutura, a remuneração de docentes, de servidores e os custos com os demais serviços necessários oferecidos em prol dos estudantes.

A Tabela 1, segundo dados da apuração realizada sobre o custo gerados das Universidades Federais, evidencia as despesas da Universidade Federal do Acre referente aos anos de 2009 até 2016, destacando que no ano de 2016, foi gasto um total de R\$ 29.421,10 por aluno, como destacado em negrito na Tabela 1 (PATRÍCIO, 2018), e, neste mesmo período, segundo dados divulgados por PROPLAN (2016), a UFAC obteve 2.616 ingressantes e apenas 1.226 concluintes.

Tabela 1 - Despesa total por aluno da Universidade Federal do Acre.

DESPESA	2009	2010	2011	2012	2013	2014	2015	2016
Pessoal Ativo	7,080.9	6,090.4	6,291.6	7,711.9	12,028.0	12,031.6	14,067.7	15,159.1
Contribuição da União ao PSS	1,364.8 1	1,208.1	1,250.7	1,656.1	2,297.9	2,265.7	2,638.6	2,568.0
Benefícios a Servidores	232.3	365.7	376.1	495.6	865.0	782.4	852.5	1,162.3
OCC (1)	807.5	495.3	418.0	408.0	1,517.1	1,938.0	1,905.7	2,531.0
Residência Médica	146.0	112.2	144.5	151.2	281.5	332.3	423.0	500.5
Assistência Estudantil	146.7	127.5	118.3	236.0	591.7	710.6	724.1	998.5
REUNI e Emendas	402.5	421.1	584.8	1,023.2	1,050.4	1,513.2	1,045.4	46.0
PARFOR	-	-	-	-	45.8	123.6	67.7	97.3
Pessoal Inativo	3,516.8	3,057.2	2,889.9	3,342.7	5,312.7	5,050.7	5,886.8	6,083.2
Precatórios Pessoal	288.8	925.1	4.9	44.9	-	-	-	187.4
Despesas c/ Receitas Próprias	65.0	94.5	39.1	23.0	50.8	51.7	97.7	78.8
TOTAL	14,051.3	12,897.2	12,117.9	15,092.5	24,041.0	24,799.7	27,709.1	29,412.1

Fonte: Patrício (2018, p. 58).

Desta forma, é possível fazer um cálculo baseado nestas informações disponíveis para se obter conhecimento do gasto para a UFAC dos alunos que ficam retidos e evadidos. De acordo com os dados do ano de 2016, e exemplificando com o cenário em que os alunos não concluintes permanecem retidos na instituição, tem-se o seguinte: subtraindo dos **2.616 ingressantes** os **1.226 concluintes** resultam em **1.390 alunos não concluintes** (1.2). Observando o gasto anual por aluno de **29.421,1 reais** e multiplicando pelos **1.390 alunos não concluintes**, tem-se um custo total de mais de **40 milhões de reais** com despesa por ano (2.2). Considerando também que o cenário ideal seria que o número de ingressantes fossem proporcionais ao número de concluintes, mas não ocorre por diversos fatores que fogem do escopo desta pesquisa, este cenário serviu apenas para exemplificar as despesas que podem ser geradas pelos alunos retidos e evadidos.

$$\text{não concluintes} = \text{ingressantes} - \text{concluintes} \quad (1.1)$$

$$\text{não concluintes} = 2.616 - 1.226 = 1.390 \quad (1.2)$$

$$\text{despesa anual} = \text{não concluintes} * \text{gasto por aluno anual} \quad (2.1)$$

$$\text{despesa anual} = 1.390 * 29.421,1 = 40.895.329 \text{ reais} \quad (2.2)$$

Levando-se em consideração estes dados, é possível fazer uma análise mais específica do índice de retenção e de evasão do curso de Sistemas de Informação, considerando o mesmo valor de **R\$ 29.421,10** gasto por aluno para UFAC no ano de 2016, e fazendo o mesmo comparativo em relação a quantidade de estudantes (3.1): dos **50 ingressantes**, apenas **6 concluíram** no ano de 2016, supondo que os **44 não concluintes** (3.2) permaneceram na instituição a despesa por ano (4.2) seria no valor de **R\$ 1.294.528,40** para esses alunos que ficaram retidos.

$$\text{não concluinte} = \text{ingressantes} - \text{concluintes} \quad (3.1)$$

$$\text{não concluinte} = 50 - 6 = 44 \quad (3.2)$$

$$\text{despesa anual} = \text{não concluintes} * \text{gasto por ano} \quad (4.1)$$

$$\text{despesa anual} = 4 * 29.421,1 = 1.294.528,4 \text{ reais} \quad (4.2)$$

Além disso, fazendo uma releitura dos dados informados, que estão presentes no Tabela 2, é possível visualizar o percentual de alunos de uma turma que concluíram a graduação no tempo ideal de 4 anos e o percentual de alunos retidos/evadidos daquela turma. A maior porcentagem de alunos formados no tempo ideal é referente à turma ingressante no ano de 2013, que alcançou 12% dos alunos

das 50 vagas ofertadas, que concluíram a graduação no ano de 2016. No entanto, nos demais anos, o índice apresentou um teto de 8%, alcançando o crítico valor de 100% de retenção nas turmas de 2008, 2010 e 2012.

Tabela 2 - Estatística de ingresso e egresso por turma.

Ano de Ingresso	Ano de Egresso	Concluintes no tempo ideal	Não concluintes no tempo ideal
2007	2010	2%	98%
2008	2011	0%	100%
2009	2012	8%	92%
2010	2013	0%	100%
2011	2014	2%	98%
2012	2015	0%	100%
2013	2016	12%	88%
2014	2017	4%	96%
2015	2018	2%	98%
2016	2019	6%	94%

Fonte: Elaboração própria, 2020.

Como se trata de um problema de interesse público, o presente trabalho visa trazer benefícios para a sociedade e, principalmente, para a Universidade Federal do Acre, sendo mais relevante ao Curso Bacharelado em Sistemas de Informação.

Com a construção de modelos preditivos sobre a retenção e a evasão dos discentes, será possibilitado o auxílio à gestão em busca da adoção de estratégias com o objetivo de diminuir os casos de retenção e de evasão, contribuindo tanto para a comunidade acadêmica quanto para a sociedade geral.

1.4 METODOLOGIA

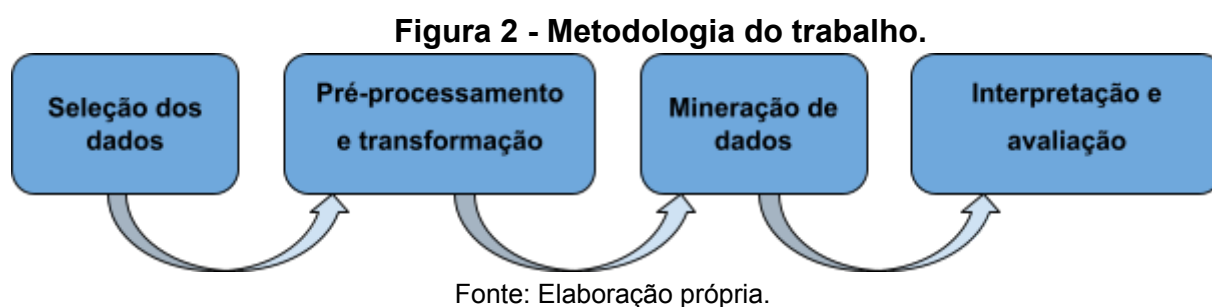
Segundo Wazlawick (2009), a metodologia consiste no sequenciamento dos passos necessários para evidenciar se os objetivos apresentados foram atingidos de fato. Desta maneira, se os passos estabelecidos no método foram realizados, consequentemente, os resultados deverão ser convincentes. Ainda de acordo com

Wazlawick (2009), foram definidas algumas classificações de tipos de pesquisa, na qual uma delas, e que foi utilizada neste projeto, é a "Apresentação de algo diferente", que consiste na apresentação de uma forma diferente para se solucionar um problema, sendo característicos de áreas emergentes.

O método científico utilizado nesta pesquisa é classificado como quantitativo, pois utiliza técnicas a partir de dados para quantificar informações para o estudo. A abordagem utilizada neste trabalho é um estudo de caso, onde o cenário de retenção e de evasão do curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre foi analisado em especial.

Além disso, em relação aos objetivos, é uma pesquisa exploratória, considerando que a utilização da Mineração de Dados na área educacional investigando a retenção e evasão é um fenômeno que ainda não foi aplicado no curso de Sistemas de Informação.

Para a realização deste trabalho, foram utilizados dados acadêmicos disponibilizados pelo Núcleo de Tecnologia da Informação (NTI) da referida instituição de ensino. Sendo assim, foram realizados os seguintes passos presentes na Figura 2, que estão associados às etapas apresentadas no processo de *Knowledge Discovery in Databases* (KDD):



1. **Seleção de dados:** realizou-se, inicialmente, a seleção de um conjunto de dados acadêmicos, que foram utilizados nas etapas seguintes.
2. **Pré-processamento e transformação dos dados:** o pré-processamento é a atividade responsável pela limpeza na base de dados, no qual são feitas correções de eventuais inconsistências e remoção de ruídos, em conjunto com a transformação, que reorganizou os dados adequadamente para que possam ser interpretados nas etapas seguintes.

3. **Mineração de dados:** Nesta etapa foi utilizada a tarefa de Classificação, sendo realizada a aplicação de algoritmos e técnica visando a criação de modelos que possam indicar a retenção e evasão do aluno. Também foi utilizado o método *InfoGainAttributeEval* para seleção dos melhores atributos.
4. **Interpretação e avaliação:** os modelos descritos na etapa anterior foram analisados, considerando sua acurácia, para que, a partir do desempenho acadêmico, possa ser prevista a evasão e retenção do estudante.

1.5 ORGANIZAÇÃO DO ESTUDO

O presente trabalho possui, além deste capítulo introdutório, mais 3 capítulos, que serão apresentados nesta seção.

O Capítulo 2 consiste na Fundamentação Teórica para entendimento da pesquisa, no qual são abordados conceitos essenciais para a construção da pesquisa, tais como: na Seção 2.1, o processo do KDD e a tarefa de classificação; na Seção 2.2, a ferramenta WEKA, que foi utilizada para a mineração de dados; a Seção 2.3, que aborda sobre a Mineração de Dados Educacionais; e a Seção 2.4, que apresenta e compara a pesquisa com os trabalhos relacionados.

O Capítulo 3 apresenta o processo experimental para se chegar aos resultados, que está dividido da seguinte maneira: planejamento do estudo, discussão e avaliação dos resultados e considerações do capítulo.

Por último, o Capítulo 4 descreve as considerações finais do trabalho, as limitações encontradas e as recomendações para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

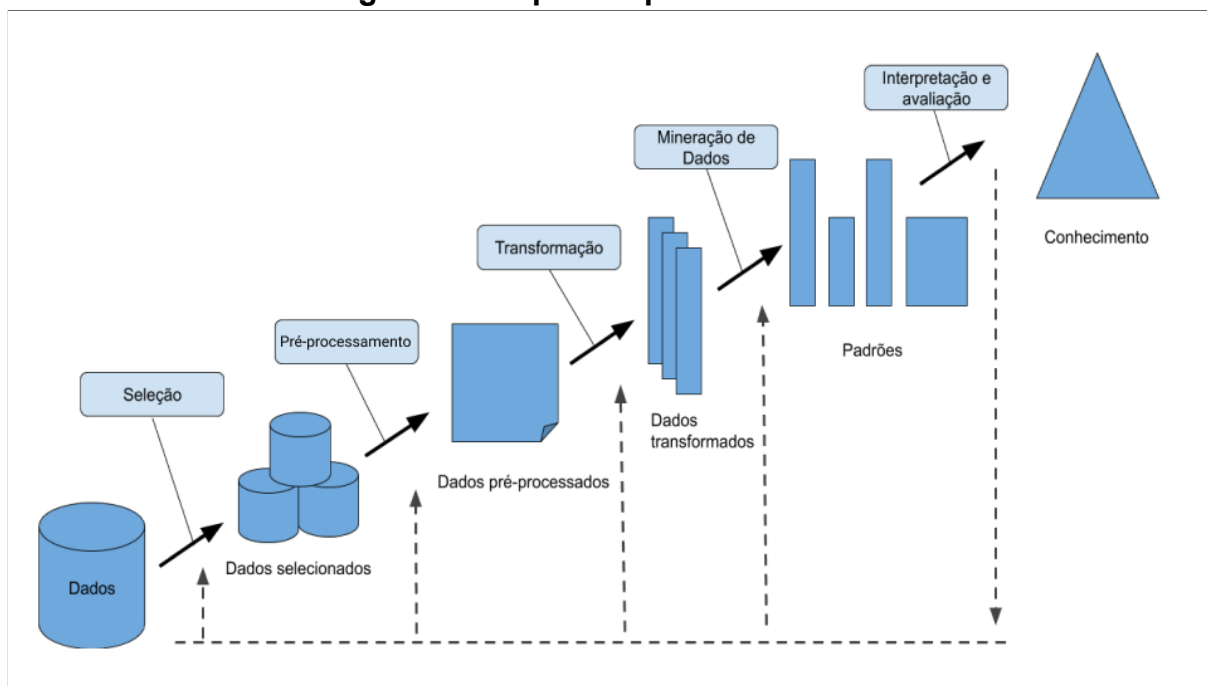
Neste capítulo são abordados alguns conceitos fundamentais para realização deste trabalho. A Seção 2.1 define processo *Knowledge Discovery in Databases* (KDD). Além disso, os conceitos complementares serão descritos nas subseções, abordando a tarefa de classificação, as métricas de avaliação e os algoritmos de classificação. Já na Seção 2.2 são abordados os temas Mineração de Dados Educacionais, Diplomação, Retenção e Evasão. A Seção 2.3 apresenta a ferramenta utilizada para o desenvolvimento do trabalho. Por último, na Seção 2.4, são abordados os trabalhos relacionados à pesquisa.

2.1 KNOWLEDGE DISCOVERY IN DATABASES

A noção de encontrar padrões úteis em dados recebeu historicamente diversos nomes, tais como mineração de dados, extração de conhecimento, descoberta de informações, coleta de informações, arqueologia de dados e processamento de padrões de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O *Knowledge Discovery in Databases* (KDD) é o processo de descoberta de conhecimento em banco de dados, que tem como principal objetivo encontrar maneiras automatizadas para explorar base de dados e reconhecer padrões existentes (BERRY, 2000).

A Figura 3 ilustra o processo de KDD, que inclui 5 etapas, que são elas: seleção dos dados, pré-processamento, transformação, mineração de dados e interpretação dos resultados. As etapas do processo são interativas e iterativas, como mostradas nas setas, sendo iterativas pois é possível retornar às etapas anteriores, caso seja necessário a repetição de uma ou mais etapas.

Figura 3 - Etapas do processo do KDD.



Fonte: Adaptado de Fayyad; Piatetsky-Shapiro; Smyth (1996).

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), as etapas do KDD podem ser conceituadas da seguinte maneira:

1. **Seleção:** etapa inicial do processo que é responsável pela seleção dos dados, onde é necessária a criação do conjunto de dados adicionando variáveis e amostras que devem ser incluídas para realização da descoberta de conhecimento.
2. **Pré-processamento:** a segunda etapa consiste na limpeza e o pré-processamento de dados, sendo que as operações básicas incluem eliminação de ruídos, coleta de informações necessárias para modelagem e determinação de estratégias para lidar com as variáveis.
3. **Transformação:** a terceira etapa é responsável pela organização dos dados com redução ou transformação da dimensionalidade, discretização dos dados, transformando valores contínuos em discretos, ou seja, um número

limitado de estados. Esse processo é importante para que na etapa seguinte os algoritmos possam ser executados corretamente.

4. **Mineração de dados:** a quarta etapa pesquisa os padrões de interesse representativo, podendo ser um específico ou um grupo de tais representações, incluindo regras de classificação ou árvores, regressão e agrupamento, envolvendo os modelos que podem ser ajustados de acordo com a observação dos dados a fim de determinar padrões.
5. **Interpretação e avaliação dos resultados:** a última etapa envolve a interpretação dos padrões e modelos que foram extraídos dos processos anteriores, incluindo a verificação e resolução dos conflitos de acordo com possível conhecimento extraído.

A quarta das etapas descritas, a mineração de dados, consiste em um problema de descoberta de tipos de padrões. Para isso, as principais tarefas de mineração de dados são: regressão (estimativa de uma valor numérico com base na nos demais atributos), regras de associação (consiste na identificação dos atributos que estão relacionados entre si), clusterização (identificação de valores similares para criar agrupamentos) e classificação (visa descobrir qual classe de um determinado registro) (CAMILLO; SILVA, 2009).

Este trabalho fará uso da tarefa de classificação para o estudo do problema relacionado à retenção e evasão dos estudantes e será melhor detalhada na próxima subseção.

2.1.1 Classificação

A classificação, que é uma das tarefas de Mineração de Dados mais estudadas do processo do KDD, consiste em um procedimento que atribui rótulos para os objetos de forma que os objetos de uma categoria sejam equivalentes aos objetos previamente rotulados de um conjunto de treinamento. Dessa forma, a classificação busca estudar o histórico dos registros (atributos) e, a partir deles,

desenvolver descrições de suas características em cada uma das classes (KAMSU-FOGUEM et al., 2013).

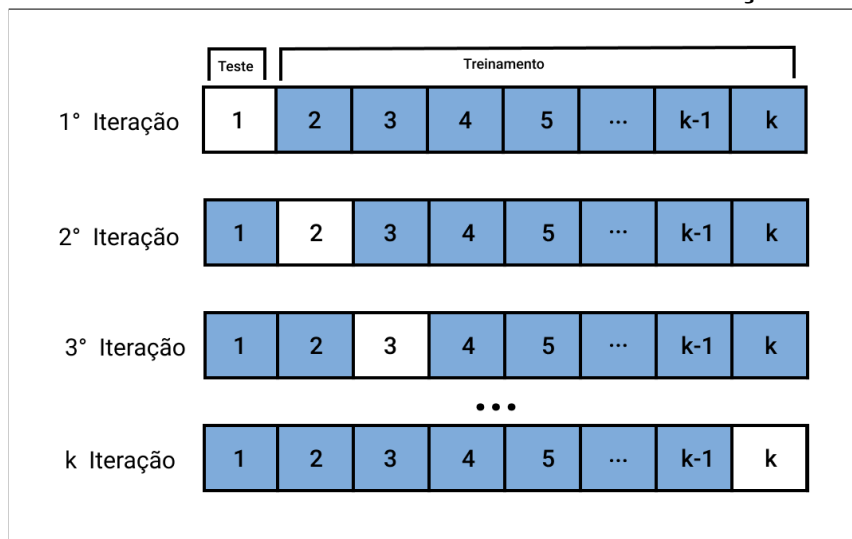
Ainda segundo Kamsu-Foguem et al. (2013), a classificação é tradicionalmente um tipo de problema de aprendizado supervisionado, pois a partir dos dados, tenta-se aprender uma função, que é extraída a partir de algoritmos supervisionados gerando uma previsão.

A tarefa de classificação é dividida em duas fases: treino e teste. Na fase de treinamento, é analisado um conjunto de registros previamente classificados de acordo com a finalidade desejada, para que, dessa forma, o algoritmo possa aprender com base nesses registros de entrada fornecidos e assim criar um modelo com esses padrões. Depois de concluir o treinamento, a fase de teste é executada em uma amostra do conjunto com novos registros. A partir do modelo gerado anteriormente, é possível classificar esses novos registros, sendo utilizado métodos específicos para avaliar sua capacidade de classificação (ALVES, 2020). Alguns desses métodos para avaliação dos algoritmos preditivos são o *holdout* e a validação cruzada.

O método *holdout* divide o conjunto de dados aleatoriamente ou preservando a ordem dos dados em dois conjuntos treino e teste, no qual, geralmente, $\frac{2}{3}$ dos dados vão para treinamento e $\frac{1}{3}$ para o teste. Primeiramente, o modelo preditivo é treinado com a base separada inicialmente para treinamento, buscando os melhores parâmetros para o modelo. Em seguida, o modelo é testado num conjunto de dados separado para o teste, onde seu desempenho é avaliado. No entanto, o *holdout* é recomendado para base de dados maiores, porque os resultados podem ser tendenciosos, se o conjunto de dados for pequeno (XIONG et al., 2020).

Já a validação cruzada (*cross-validation*) é caracterizada pela divisão dos dados em K-partições proporcionais, como mostrada na Figura 4. Logo depois, uma das K-partições é utilizada como o conjunto de teste e as demais K-1 partições são unidas para formar um conjunto de treinamento, e as iterações são repetidas até chegar em K, ou seja, na última iteração. Uma vantagem da validação cruzada é que a estimativa de erro é calculada com média do resultado de todas as K-iterações para obter a eficácia total do modelo (XIONG et al., 2020).

Figura 4 - Divisão da base de dados utilizando validação cruzada.



Fonte: Adaptado de Xiong et al. (2020).

Dessa forma, com intuito de selecionar o procedimento mais adequado para a pesquisa, foi escolhida a validação cruzada, pois consiste em um processo de reamostragem usado para avaliar modelos com uma amostra limitada de dados, que é o caso da pesquisa. A previsão gerada é avaliada de acordo com algumas métricas, que serão detalhadas na próxima subseção.

2.1.2 Métricas de avaliação da classificação

As métricas de avaliação são meios para medir o desempenho de um classificador, porque, a partir destas medidas, é possível analisar a performance do algoritmo com os dados que foram avaliados. Com essa visualização do desempenho, é possível buscar a adoção de estratégias para melhorar continuamente a capacidade de classificação e o aumento das possibilidades de acerto nas previsões.

Esta avaliação é feita de acordo com os números de registros classificados como corretos e incorretos. A origem das métricas é chamada matriz de confusão, que é um dos meios para avaliar a capacidade preditiva de um algoritmo (ALVES, 2020).

A matriz de confusão mostra quatro valores possíveis para caracterizar os elementos que servem para a obtenção das principais métricas de avaliação dos classificadores, considerando, por exemplo, um problema com classe binária. Os valores da diagonal principal indicam as previsões corretas, enquanto os valores fora da diagonal principal indicam previsões incorretas (THARWAT, 2020).

Quadro 1 - Matriz de confusão.

MATRIZ DE CONFUSÃO		CLASSE REAL	
		Positivo	Negativo
CLASSE PREVISTA	Positivo	VP Verdadeiro Positivo	FP Falso Positivo
	Negativo	FN Falso Negativo	VN Verdadeiro Negativo

Fonte: Adaptado de Tharwat (2020).

De acordo com Chicco e Jurman (2020), os resultados da matriz de confusão podem ser conceituados da seguinte maneira:

- **Verdadeiro Positivo (VP):** quando o valor de determinado registro é classificado corretamente como positivo.
- **Verdadeiro Negativo (VN):** quando o valor de um determinado registro é classificado corretamente como negativo.
- **Falso Positivo (FP):** quando o valor de um determinado registro é classificado incorretamente como positivo.
- **Falso Negativo (FN):** quando o valor de um determinado registro é classificado incorretamente como negativo.

Após a obtenção dos valores da matriz de confusão, para avaliar o desempenho do classificador, é necessário utilizar algumas fórmulas. Para isso, serão conceituadas algumas delas, tais como: **acurácia**, **taxa de erro**, **sensibilidade** e **especificidade**.

A **acurácia** (5) consiste na proporção da taxa de acertos, ou seja, levando em consideração todos os registros presentes na base de dados, é calculada a precisão dos registros que foram classificados corretamente. A acurácia representa a razão entre as instâncias preditas corretamente e todas as instâncias, como mostrado na equação (5) (ALVES, 2020; CHICCO; JURMAN, 2020).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (5)$$

Em contrapartida, a **taxa de erro** (6) é caracterizada como a proporção dos registros classificados incorretamente, sendo a razão dos registros preditos de forma errada e todos os resultados da matriz de confusão (HAN; KAMBER; PEI, 2011).

$$Taxa\ de\ erro = \frac{FP + FN}{VP + VN + FP + FN} \quad (6)$$

A **sensibilidade** (7), ou taxa de verdadeiro positivo, consiste na proporção de registros positivos que foram classificados de modo correto, considerando dessa forma a capacidade do modelo em avaliar corretamente os dados (HAN; KAMBER; PEI, 2011).

$$Sensibilidade = \frac{VP}{VP + FN} \quad (7)$$

Enquanto isso, a **especificidade** (8), ou taxa de verdadeiro negativo, representa a proporção de registros negativos que foram corretamente classificados como negativo, ou seja, é considerada a capacidade do modelo de identificação dos registros que não pertencem à classe que está sendo utilizada (HAN; KAMBER; PEI, 2011).

$$Especificidade = \frac{VN}{VN + FP} \quad (8)$$

Para o contexto desta pesquisa, foi utilizado a acurácia como o indicador para a avaliação do desempenho do classificador, com a medição dos modelos gerados a partir da taxa de acerto dos registros que foram classificados corretamente. Com isso, é possível avaliar e determinar quais modelos apresentam os melhores

resultados para o contexto. Os modelos foram gerados com a utilização dos algoritmos de classificação que serão apresentados na subseção 2.1.3.

2.1.3 Algoritmos de classificação

A tarefa de classificação utiliza aprendizagem supervisionada, encontrando uma função através de dados rotulados e possuindo diversos algoritmos que podem ser utilizados em conjunto ou separadamente para diferentes propósitos. Neste trabalho serão estudados e usados nos experimentos alguns deles, tais como: *K-Nearest Neighbor*, *Naive Bayes*, Máquina de Vetor de Suporte, J48 e *Random Forest*.

O algoritmo **K-vizinho mais próximo** (*K-Nearest Neighbor* ou KNN) faz parte da família de aprendizagem baseada em instância (*Instance-based Learning* ou IBL). Essa família conserva as instâncias de treinamento e, dessa forma, quando surge uma nova instância, o algoritmo classifica a partir das instâncias similares próximas armazenadas do conjunto de treinamento. No algoritmo KNN é atribuída a uma nova instância a classe mais frequente entre os vizinhos mais próximos (FARIA, 2016).

O **Naive Bayes (NB)**, que é baseado no teorema de probabilidade do matemático Thomas Bayes, é um algoritmo de aprendizagem máquina, sendo um dos algoritmos de mineração de dados mais populares, conhecido principalmente pela resolução de problemas de classificação. Este classificador é chamado ingênuo (*naive*), pois sua estrutura assume que os valores dos atributos são independentes uns dos outros, ou seja, em outras palavras, todas as variáveis folha têm apenas o seguinte relacionamento: A dependência da variável de classe que é considerada como a variável raiz (CHEN et al., 2020; ZEVAREX; SANTOS, 2020).

O algoritmo **Máquina de Vetor de Suporte** (*Support Vector Machine* ou SVM) é capaz de, a partir do treinamento e com o aprendizado gerado, obter a capacidade de generalização. Dessa forma, por exemplo, considerando um problema binário, o objetivo do SVM é dividir as instâncias das duas classes e

utilizar um classificador linear que maximiza a margem de separação entre as classes que estão divididas no hiperplano. O SVM tem como objetivo produzir um classificador que pode trabalhar com exemplos desconhecidos, ou seja, aplicar o modelo em instâncias que durante o processo de treinamento não foram utilizadas, obtendo-se assim a capacidade de prever novas entradas e saídas no futuro (OLIVEIRA JUNIOR, 2010).

O **J48** é uma implementação de um dos algoritmos de indução de árvores mais conhecidos, o C4.5. Uma árvore de decisão é constituída de uma cadeia de nós de decisão que tem como objetivo a existência de um atributo alvo, que é conectado através das ramificações desde a raiz até as folhas. O algoritmo J48 é normalmente utilizado, pois apresenta poucas restrições relacionadas às características dos atributos utilizados, sendo adequado para os processos que envolvem atributos qualitativos contínuos e discretos, não exigindo uma específica distribuição de probabilidade (VIEIRA et al., 2018).

O **Random Forest** é um algoritmo de aprendizado de máquina supervisionado que utiliza o resultado de várias árvores de decisão e que gera seu próprio resultado baseado nelas, diminuindo assim a probabilidade da variância e viés, pois o algoritmo alcança suas próprias conclusões através das folhas ou decisões finais de cada nó. Com isso, a precisão do modelo é aumentada, visto que está encontrando o valor médio, caso o atributo alvo seja numérico, e rótulo mais frequente, quando o alvo é discreto, através dos resultados de muitas árvores de decisão diferentes (LEITE; MORAES; LOPES, 2021).

2.2 SOFTWARE PARA MINERAÇÃO DE DADOS

A mineração de dados pode ser executada de diversas formas, como, por exemplo, através da implementação direta dos algoritmos de mineração, conforme foi mencionado na Seção 2.1.3. No entanto, existem linguagens de programação que fornecem diversas bibliotecas e ferramentas de mineração de dados com as opções, por exemplo, para se utilizar as tarefas de classificação. Dessa maneira,

alguns softwares disponibilizam algoritmos já implementados e que facilitam o uso e a aplicação da mineração de dados. O software WEKA é um exemplo de software dessa finalidade e foi utilizado nesta pesquisa.

WEKA (*Waikato Environment for Knowledge Analysis*) é um software de aprendizado de máquina, desenvolvido na linguagem de programação Java, pela Universidade de Waikato, na Nova Zelândia, sendo um software de código aberto e distribuído através da licença *GNU General Public License*.

WEKA possui um conjunto de algoritmos de aprendizado de máquina (*Machine Learning* ou ML) para tarefas de mineração de dados, possuindo também ferramentas para preparação, classificação, regressão, agrupamento, regras de associação de mineração e visualização de dados e integrando várias tecnologias de aprendizado de máquina padrão. Através do WEKA, é possível obter conhecimento útil, pois especialistas em um determinado campo podem usar o ML para obter conhecimento útil de um banco de dados que é muito grande para ser analisado manualmente (WITTEN et al., 2016).

2.3 MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados Educacionais (*Educational Data Mining* ou EDM) foi formalizada no evento chamado "*EDM: First International Conference on Educational Data Mining*", que foi realizado em Montreal, Canadá no ano de 2008. Posteriormente, o evento estabeleceu-se como uma conferência anual, sendo que, em 2009, foi criado um periódico específico para esta área nomeado de "*JEDM - Journal of Educational Data Mining*" (COSTA et al., 2013).

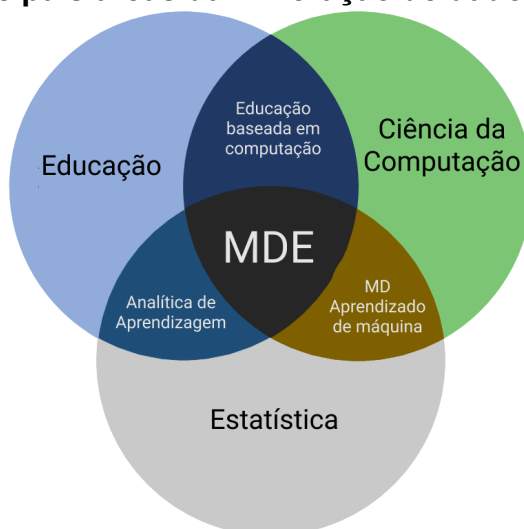
A Mineração de Dados Educacionais é definida como um campo de pesquisa que tem como foco principal a aplicação de métodos em ambientes educacionais para a exploração e investigação dos conjuntos de dados obtidos nos mesmos (BAKER; ISOTANI; CARVALHO, 2011).

Segundo Silva (2019), a Mineração de Dados Educacionais constitui-se um campo de pesquisa em crescimento que visa o desenvolvimento, a adaptação e a aplicação de métodos do KDD para descobrir modelos de conhecimento a partir de grandes bancos de dados contendo informações referentes à educação. As aplicações de EDM tentam descobrir novos conhecimentos para compreender melhor os alunos e o processo de aprendizagem para tentar resolver alguns problemas relacionados à área de educação. Por consequência, a partir desses processos, a EDM converte os dados originais do sistema educacional em informações que são pertinentes tanto para a pesquisa educacional quanto para a prática do processo educacional.

A EDM tem como objetivo principal utilizar a análise de dados através da exploração estatística, da aprendizagem de máquina e de algoritmos de mineração de dados aplicados sobre os diferentes tipos de dados da área de ensino com a finalidade de responder questões da área educacional (OLIVEIRA JÚNIOR, 2015).

Do mesmo modo, como retratado na Figura 5, a Mineração de Dados Educacionais consiste no estudo das áreas relacionadas a educação, a ciência da computação e a estatística, utilizando também a interseção entre elas, tais como: a educação baseada na computação, aprendizado de máquina e analítica de aprendizagem. O objetivo é utilizar a junção dessas áreas de conhecimento para investigar questões científicas relacionadas à área da educação.

Figura 5 - Principais áreas da mineração de dados educacionais.



Fonte: Adaptado de Romero e Ventura (2013).

Dentro de uma das áreas da Mineração de Dados Educacionais, mais especificamente, a educação relacionada a um programa de nível superior, existem conceitos a respeito do rendimento de um aluno, desses conceitos pode-se especificar a diplomação, a retenção e a evasão, que serão tratadas na subseção seguinte.

2.3.1 Diplomação

Segundo um relatório publicado pela comissão especial sobre os cursos de graduação do ensino superior públicas, a diplomação pode ser conceituada da seguinte maneira: “Diplomado: aluno que concluiu o curso de graduação dentro do prazo máximo de integralização curricular, fixado pelo CFE, contado a partir do ano/período-base de ingresso” (ANDIFES; ABRUEM; SESU/MEC, 1996, p. 59). Em outras palavras, os diplomados referem-se aos estudantes que concluíram o curso dentro do prazo estipulado para integralização das disciplinas.

O baixo número de diplomados é um dos fatores que mais interfere na queda da Taxa de Sucesso na Graduação (TSG). Este indicador de desempenho é utilizado na educação superior, sendo calculado como mostrado na equação (9), através da razão entre número total de diplomados e número total de ingressantes.

$$TSG = \frac{N^{\circ} \text{ de diplomados}}{N^{\circ} \text{ total de alunos ingressantes}} \quad (9)$$

Fonte: TCU; SESu/MEC; SFC (2004, p. 4).

O curso de Bacharelado em Sistemas de Informação da UFAC apresenta os seguintes resultados da Taxa de Sucesso na Graduação, mostrados na Tabela 3, de acordo com os últimos 5 anos, quanto mais próximo do 1 melhor é o desempenho da taxa.

Tabela 3 - TGS do curso de Sistemas de Informação.

Ano	TGS
2015	0,10
2016	0,46
2017	0,20
2018	0,26
2019	0,34

Fonte: Elaboração própria.

2.3.2 Retenção

A retenção no ensino superior constitui-se no processo de permanência maior para cumprimento da carga horária prevista na matriz curricular do estudante em um curso de graduação. Como este conceito é amplo, é possível fazer algumas inferências de elementos que fazem parte do processo de retenção tais como: reprovações, trancamentos ou até mesmo atraso voluntário por parte do aluno. Com a retenção, o aluno necessita de um tempo prolongado para a conclusão do curso, comprometendo a taxa de sucesso citada anteriormente, gerando ociosidade dos recursos financeiros, humanos e materiais, além da possibilidade de provocar a evasão do estudante (PEREIRA, 2013).

2.3.3 Evasão

A evasão do ensino superior é um problema que prejudica os resultados do sistema educacional. Os estudantes que começaram, mas não concluíram o curso causaram uma perda de níveis sociais, acadêmicos e econômicos. No setor público, são recursos públicos que foram investidos mas sem gerar o devido retorno. No setor privado, é uma perda significativa de receita. Tanto um como no outro caso, a

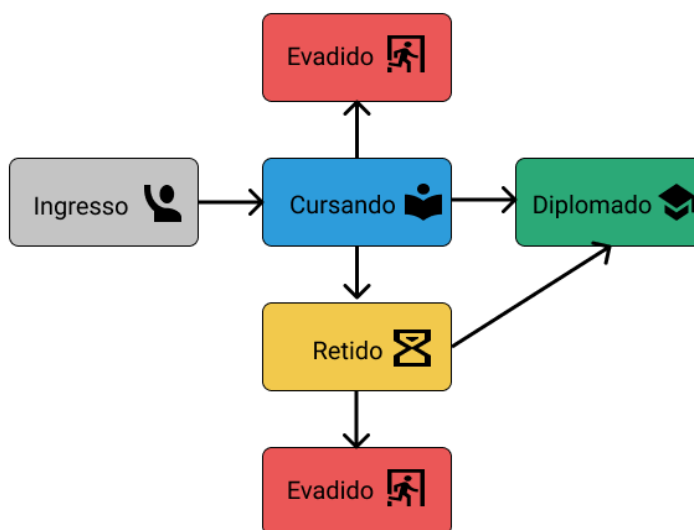
evasão ocasiona ociosidade de professores, funcionários, equipamentos e espaço físico (SILVA FILHO et al., 2007).

Segundo estudo realizado pela ANDIFES, ABRUEM, SESU/MEC (1996), a evasão pode ser classificada da seguinte maneira:

1. **Evasão de curso:** quando os alunos deixam seu curso de ensino superior de origem, sem concluí-lo em diferentes circunstâncias, como por exemplo, abandono (não realiza a matrícula inscrição), desistência (oficial), transferência ou reopção (mudança de curso) ou exclusão por normas institucionais como jubilamento.
2. **Evasão da instituição:** quando um aluno desliga-se da instituição de ensino superior da qual está vinculado mas continua o estudo em outra instituição de ensino superior.
3. **Evasão do sistema:** quando um aluno abandona o ensino superior permanente ou temporariamente.

A Figura 6 ilustra os conceitos descritos anteriormente exemplificando a partir do aluno **ingresso** os possíveis caminhos que o aluno pode seguir no decorrer da graduação depois do **cursando** tem-se a possibilidade para o aluno seguir para **diplomado**, **retido** ou **evadido**.

Figura 6 - Possíveis caminhos do estudante na graduação.



Fonte: Elaboração própria.

2.4 TRABALHOS RELACIONADOS

Nesta seção serão descritos alguns trabalhos que utilizaram a Mineração de Dados para responder questões relacionadas à área educacional e um trabalho que estudou a problemática de retenção e evasão do curso de Bacharelado em Sistemas de Informação da UFAC.

Oliveira Júnior (2015) desenvolveu uma pesquisa que buscou analisar a evasão de alunos de graduação em cursos presenciais para auxiliar tomadores de decisão em instituições de ensino utilizando uma metodologia de abordagem computacional para identificação de padrões. A tarefa de classificação foi utilizada para considerar as categorias “haverá evasão” e “não haverá evasão”, com base na seleção e criação de atributos. Além disso, também foi proposto um método de seleção dos melhores atributos e, desse modo, os resultados experimentais mostraram os atributos criados foram relevantes para a previsão da evasão, indicando a contribuição da criação de novos atributos nas tarefas de mineração de dados, possibilitando que essas inferências subsidiem a tomada de decisão dos gestores da educação.

Manhães (2015) estudou os alunos em risco de evasão e se o desempenho acadêmico desses alunos era diferente daqueles que conseguiram concluir o curso. Sob essa premissa, é possível determinar quais alunos têm maior probabilidade de concluir ou abandonar os cursos de graduação. Portanto, a mineração de dados pode ser usada para desenvolver estratégias computacionais para que possam prever quais alunos estão em risco de abandono e quais alunos podem se formar. Foi criada uma arquitetura modular, que incorpora técnicas de mineração de dados educacionais para prever os alunos ao final de cada semestre e apontar aqueles que estão em risco de evasão.

A pesquisa elaborada por Hoed (2016) apresentou um estudo sobre evasão em cursos de graduação da área da computação que tem como base dados do Instituto Nacional de Pesquisa e Pesquisa Educacional Anísio Teixeira (INEP) e da Universidade de Brasília (UnB). Na pesquisa quantitativa realizada foi obtida a taxa de abandono anual e aplicadas técnicas estatísticas de análise de sobrevivência e

mineração de regras de associação por meio do algoritmo Apriori. Também foi realizada uma análise qualitativa com base em questionário com alunos evadidos do curso superior da UnB para verificar os motivos da evasão. Por meio da análise de sobrevivência, pode-se considerar o tempo de fuga e evasão do aluno para analisar a situação de evasão do mesmo, sendo analisada a possibilidade de evasão para diferentes fatores. Juntamente com tecnologia de mineração de dados e a análise de sobrevivência, o conjunto pode fornecer um suporte importante para a política educacional das instituições de ensino superior.

Os trabalhos Oliveira Júnior (2015), Manhães (2015), Hoed (2016), utilizaram técnicas de mineração de dados aplicadas no contexto educacional, especialmente relacionado à análise dos dados referente ao ensino superior. Para isso, foi realizada a aplicação de algoritmos nas bases de dados de cada pesquisa. No Quadro 2 encontra-se o mapeamento dos atributos em comuns, dos trabalhos relacionados e a atual pesquisa do curso de Sistemas de Informação da UFAC na coluna “SI - UFAC (2021)”, com algumas adaptações dos nomes dos atributos que possuíam nomenclatura distintas mas tinham mesmo significado.

Quadro 2 - Mapeamentos dos atributos dos trabalhos relacionados.

Atributos em comum	Oliveira Júnior (2015)	Manhães (2015)	Hoed (2016)	SI - UFAC (2021)
Gênero	✓		✓	✓
Ano de ingresso		✓	✓	✓
Forma ingresso		✓	✓	✓
Idade início do curso	✓			✓
Código do curso		✓	✓	✓
Nome do curso		✓	✓	✓
Situação atual do aluno no curso		✓	✓	✓
Coeficiente rendimento	✓	✓		✓
Situação do aluno na disciplina		✓		✓
Nome na disciplina		✓		✓
Nota do aluno na disciplina		✓		✓
Ano da disciplina cursada pelo aluno		✓		✓
Créditos da disciplina		✓		✓

Fonte: Elaboração própria.

Já em relação ao cenário de retenção e de evasão no curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, Saldanha (2016) realizou uma pesquisa qualitativa com alunos evadidos entre 2006 a 2012 e os retidos entre 2008 a 2012. Foi feito um levantamento com entrevistas e aplicação de questionários a esses alunos para obter informações sobre seu relacionamento com o curso. Os principais fatores identificados que levam à retenção são o fracasso com as reprovações nas disciplinas consideradas mais difíceis em comparação ao ensino médio, também foi apontado que um dos motivos é a necessidade de trabalho e falta de motivação. Quanto à evasão, foi determinado a dificuldade com didática de ensino dos professores, falta de tempo suficiente devido ao trabalho e falta de afinidade com o curso.

3 PREVISÃO DE EVASÃO E RETENÇÃO NO CURSO DE SISTEMAS DE INFORMAÇÃO

Este capítulo apresenta a metodologia utilizada durante o desenvolvimento do trabalho e os resultados alcançados. A Seção 3.1 descreve o planejamento do estudo, desde a seleção dos atributos até a execução e interpretação dos resultados apresentados pelos algoritmos. A Seção 3.2 apresenta a mineração e avaliação dos dados e, por fim, a Seção 3.3 mostra a conclusão dos resultados da pesquisa.

3.1 PLANEJAMENTO DO ESTUDO

A pesquisa utilizou a metodologia apresentada na Seção 1.4, referente ao processo de *Knowledge Discovery in Databases* (KDD), onde foi usada a base de dados disponibilizada pelo Núcleo de Tecnologia da Informação (NTI) da Universidade Federal do Acre. O sigilo dos dados foi mantido, pois as informações utilizadas nos experimentos, não continham dados pessoais dos alunos que os identificassem diretamente, preservando o sigilo dos alunos no decorrer da pesquisa.

3.1.2 Seleção e pré-processamento dos dados

A seleção dos dados consistiu na análise inicial dos atributos existentes no banco de dados da UFAC que poderiam ser disponibilizados para o desenvolvimento da pesquisa. A base de dados disponibilizada possuía 16 atributos, conforme mostrado no Quadro 3.

Quadro 3 - Atributos disponibilizados.

Nº	Atributos	Tipo
1	genero	categórico
2	estado_civil	categórico
3	data_nascimento	numérico
4	codigo_disciplina	categórico
5	nome_disciplina	categórico
6	carga_horária	numérico
7	creditos	numérico
8	ano	numérico
9	semestre	categórico
10	situacao	categórico
11	media_final	numérico
12	numero_faltas	numérico
13	ano_ingresso	numérico
14	forma_ingresso	categórico
15	ano_evasao	numérico
16	forma_evasao	categórico

Fonte: Elaboração própria.

Ao total, a base possuía 40.503 instâncias referente aos anos de 1996 até 2020, totalizando 24 anos de histórico. Para realizar os experimentos preditivos com dados relativos à retenção e evasão, foi necessário realizar o pré-processamento com uma limpeza para diminuir as inconsistências e filtrar das 40.503 instâncias, apenas os dados relacionados à grade curricular vigente, que é a versão 4 do ano de 2008, excluindo, dessa forma, os dados relativos às versões anteriores da grade: versão 1 (1996), versão 2 (2004), versão 3 (2006). Ao final do pré-processamento,

resultou em 19.977 instâncias, que são relacionadas aos anos de 2008 até 2019, no qual cada instância representa uma matéria cursada pelo aluno. Desta forma, os experimentos foram conduzidos com dados de 11 anos do curso de Sistemas de Informação.

3.1.2 Transformação dos dados

Após a etapa do pré-processamento, iniciou-se a etapa de transformação dos dados, na qual constatou-se que os atributos já existentes no conjunto de dados poderiam derivar outros novos atributos, enriquecendo a base de dados. Os novos atributos criados também tiveram embasamento teórico, como mostrado na Seção 2.4, referentes a estudos já realizados na área de mineração de dados educacionais.

Como os dados fornecidos não possuíam um atributo para identificar o período, foi necessário criar atributos para essa identificação. Além disso, esses atributos tiveram um papel importante para a criação de outros novos atributos para a pesquisa. Os atributos criados para os períodos foram esses, que foram adaptados para cada experimento:

1. **Período porcentagem:** utiliza o cálculo de porcentagem da quantidade de total disciplinas aprovadas pelo o aluno no curso e a quantidade total de disciplinas da grade curricular para fazer uma estimativa do período daquele aluno no semestre.
2. **Período ideal:** faz uso da separação dos períodos de acordo com a ordem para se cursar as disciplinas, seguindo a divisão da grade curricular do curso

Após essa divisão de identificação dos períodos, também foram criados novos atributos utilizando a linguagem de programação Python, como, por exemplo, o atributo criado relativo ao coeficiente de rendimento do aluno. Para o cálculo deste atributo, foram utilizados os atributos de período citado anteriormente, calculando a média ponderada de dois indicadores: a média final e os créditos das disciplinas

como peso. Dessa forma, foi obtido o coeficiente de rendimento do aluno por período.

Para contribuir com o processo de previsão, também foram adicionados atributos relacionados ao histórico, que são referentes a dados dos períodos anteriores do aluno tais como: coeficiente de rendimento, quantidade de disciplinas cursadas, aprovadas e reprovadas e média das disciplinas aprovadas e reprovadas.

Com esses atributos, é possível analisar se o desempenho dos alunos nos períodos anteriores afetam de alguma forma seu desempenho no período posterior. Foram utilizados dois tipos históricos: o geral e o recente.

1. **Histórico geral:** é o histórico acumulado do aluno de todos os períodos anteriores, como, por exemplo, caso o aluno esteja no 3º período, o histórico geral contém dados do 1º e 2º período.
2. **Histórico recente:** é o histórico dos dados somente do período anterior, ou seja, utilizando o exemplo anterior, caso aluno esteja no 3º período, o histórico recente vai conter dados apenas do 2º período.

Esses dois conjuntos de atributos foram importantes, pois, através deles, é possível ter uma visão geral do aluno nos períodos, bem como o desempenho no período anterior, mostrando também se em determinado período ocorre um aumento ou declínio com relação ao histórico geral.

Finalizando a etapa de transformação dos dados, tem-se o total de 42 atributos, como mostrado no Quadro 4. Para facilitar a compreensão dos atributos, os mesmos foram categorizados por classe, de acordo com a primeira coluna. Na terceira coluna está o nome do atributo; na quarta coluna, a sua respectiva descrição ao lado e, na última coluna, a marcação dos atributos que foram criados.

Quadro 4 - Todos os atributos da base de dados.

Classe	Nº	Atributo	Descrição	Criado
Dados gerais	1	genero	Gênero do(a) estudante	Não
	2	estado_civil	Estado civil do(a) estudante	Não
	3	data_nascimento	Data de nascimento do(a) estudante	Não
	4	idade_inicio_curso	Idade do(a) estudante no primeiro ano do curso	Sim

	5	duracao_curso	Duração do curso baseado no ano_ingresso (entrada) e ano_evasão (saída) do(a) estudante	Sim
Ingresso	6	forma_ingresso	Forma de entrada do(a) estudante	Não
	7	ano_ingresso	Ano de entrada do(a) estudante na UFAC	Não
Evasão	8	forma_evasao	Forma de saída do(a) estudante da UFAC	Não
	9	ano_evasao	Ano de saída do(a) estudante da UFAC	Não
Dados da disciplina	10	semestre	Semestre que o(a) estudante cursou a disciplina	Não
	11	ano	Ano que o(a) estudante cursou a disciplina	Não
	12	codigo_disciplina	Código de identificação da disciplina	Não
	13	nome_disciplina	Nome da identificação da disciplina	Não
	14	carga_horaria	Carga horária total da disciplina	Não
	15	numero_faltas	Número de faltas do(a) estudante na disciplina	Não
	16	media_final	Nota final do(a) estudante na disciplina	Não
	17	periodo_porcentagem	Cálculo da estimativa do período de acordo com a quantidade de disciplinas aprovadas	Sim
	18	periodo_ideal	Período ideal para cursar a disciplina de acordo com a grade curricular	Sim
Créditos	19	creditos	Créditos total da disciplina	Não
	20	credito_teorico	Créditos teórico da disciplina	Sim
	21	credito_pratico	Créditos prático da disciplina	Sim
	22	credito_campo	Créditos campo da disciplina	Sim
Frequência	23	percentual_frequencia_periodo	Resultado do cálculo de frequência utilizando o número_faltas e carga_horária para cada disciplina durante o período	Sim
	24	percentual_frequencia_disciplina	Resultado do cálculo de frequência utilizando o número_faltas e carga_horária para cada disciplina	Sim
Período	25	media_periodo	Média aritmética do(a) estudante referente às matérias que deveriam ser cursadas no período ideal	Sim
	26	coeficiente_rendimento_periodo	Média ponderada da média_final de cada disciplina e os créditos da mesma pelo período_ideal	Sim
	21	total_disciplinas_cursada	Quantidade total de disciplinas cursadas no período	Sim
Histórico geral	22	historico_geral_coeficiente_rendimento	Média ponderada da media_final de cada disciplina e os créditos acumulado dos períodos anteriores	Sim
	23	historico_geral_total_cursadas	total de disciplinas cursadas acumulado dos períodos anteriores	Sim

	24	historico_geral_aprovadas_quantidade	Quantidade de disciplinas aprovadas acumulado dos períodos anteriores	Sim
	25	historico_geral_aprovadas_media	Média das disciplinas aprovadas acumulado dos períodos anteriores	Sim
	26	historico_geral_reprovadas_quantidade	Quantidade de disciplinas reprovadas acumuladas dos períodos anteriores	Sim
	27	historico_geral_reprovadas_media	Média das disciplinas reprovadas acumuladas dos períodos anteriores	Sim
	28	historico_geral_percentual_frequencia	Utilizando o atributo percentual_frequencia acumulado dos períodos anteriores	Sim
Histórico recente	23	historico_recente_coeficiente_rendimento	Média ponderada da media_final de cada disciplina e os créditos da mesma pelo período anterior	Sim
	29	historico_recente_total_cursadas	Total de disciplinas cursadas acumuladas no período anterior	Sim
	30	historico_recente_aprovadas_quantidade	Quantidade de disciplinas aprovadas no período anterior	Sim
	31	historico_recente_aprovadas_media	Média das disciplinas aprovadas no período anterior	Sim
	32	historico_recente_reprovadas_quantidade	Quantidade de disciplinas reprovadas no período anterior	Sim
	33	historico_recente_reprovadas_media	Média das disciplinas reprovadas no período anterior	Sim
	34	historico_recente_percentual_frequencia	Utilizando o atributo percentual_frequencia do período anterior	Sim
	35	historico_recente_classe_rendimento	Categorização do rendimento de acordo com as aprovações com relação ao total de disciplinas cursadas	Sim
	36	historico_recente_classe_desempenho	Categorização do desempenho de acordo com a média das disciplinas cursadas	Sim
	37	historico_recente_classe_disciplinas_cursadas	Categorização da quantidade média das disciplinas cursadas	Sim
Alvo	38	retido	Classificar sim ou não para a retenção do aluno	Sim
	39	categoria_retencao	Categorização com relação ao tempo do aluno para concluir o curso	Sim
	40	evadido	Classificar sim ou não para a evasão do aluno	Sim
	41	categoria_evasao	Categorização da forma de evasão do aluno do curso	Não
	42	situacao_disciplina	Situação do(a) estudante referente a disciplina cursada	Não

Fonte: Elaboração própria.

3.1.2 Atributos utilizados

Dentre os atributos descritos no Quadro 4, alguns não foram usados na mineração de dados, pois não agregaram valor na execução dos algoritmos, servindo apenas como atributo intermediário para a criação dos novos atributos. Por outro lado, os atributos referentes à média final, número de faltas e percentual de frequência da disciplina foram excluídos, porque caracterizam o alvo da previsão, gerando uma alta taxa de acurácia, e acabavam identificando as classes e enviesando a capacidade de predição do modelo.

O Quadro 5 mostra os atributos selecionados para execução dos experimentos. Como foram realizados diferentes tipos de experimentos, alguns atributos não foram usados em todos, pois não eram aplicáveis ao escopo da análise, como, por exemplo, os atributos com ID de 6 a 11 foram aplicados em apenas um dos cinco experimentos realizados.

Quadro 5 - Atributos selecionados.

Id	Atributo	Tipo
1	genero	categórico
2	estado_civil	categórico
3	idade_inicio_curso	numérico
4	duracao_curso	numérico
5	forma_ingresso	categórico
6	nome_disciplina	categórico
7	carga_horaria	numérico
8	credito_teorico	numérico
9	credito_pratico	numérico
10	credito_campo	numérico
11	historico_recente_cr	numérico
12	historico_recente_percentual_frequencia	numérico
13	historico_recente_media	numérico
14	historico_recente_aprovadas_quantidade	numérico
15	historico_recente_aprovadas_media	numérico

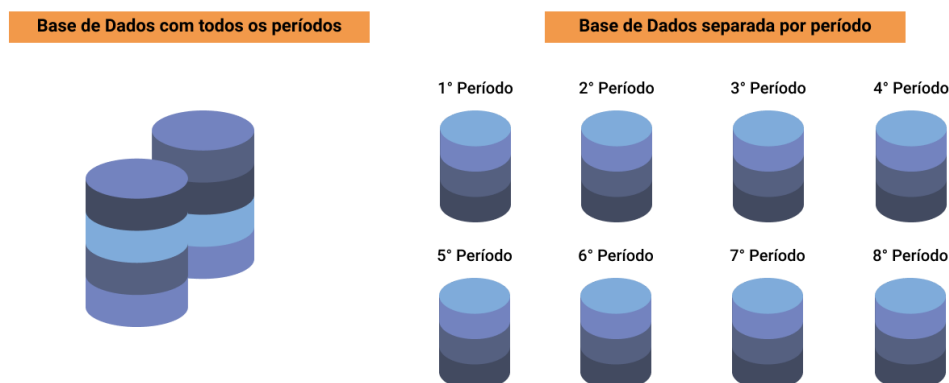
16	historico_recente_reprovadas_quantidade	numérico
17	historico_recente_reprovadas_media	numérico
18	historico_recente_cursadas	numérico
19	historico_recente_classe_rendimento	categórico
20	historico_recente_classe_desempenho	categórico
21	historico_recente_classe_disciplinas_cursadas	categórico
22	historico_geral_cr	numérico
23	historico_geral_percentual_frequencia	numérico
24	historico_geral_media	numérico
25	historico_geral_aprovadas_quantidade	numérico
26	historico_geral_aprovadas_media	numérico
27	historico_geral_reprovadas_quantidade	numérico
28	historico_geral_reprovadas_media	numérico
29	historico_geral_cursadas	numérico
30	evadido	categórico
31	categoria_evasao	categórico
32	retido	categórico
33	categoria_retencao	categórico
34	situacao_disciplina	categórico

Fonte: Elaboração própria.

3.1.2 Divisão dos experimentos

Para execução dos experimentos, foi necessário a divisão da base de dados por período, ou seja, a base de dados principal foi separada em 8 bases de dados secundárias, do 1º período até o 8º período, como mostrado na Figura 7, sendo aplicadas separadamente em cada uma dessas bases os algoritmos da tarefa de classificação.

Figura 7 - Divisão da base de dados.



Fonte: Elaboração própria.

Nesta pesquisa, ao total, foram realizados 5 experimentos baseados em atributo-alvo e explicados, resumidamente, da seguinte maneira, sendo melhor detalhados na Seção 3.1.2:

1. **Retenção em binário** ('RETIDO'): atributo binário da retenção que mostra se o aluno ficou retido (sim ou não).
2. **Retenção em categorias** ('CATEGORIA_RETENCAO'): atributo que classifica a retenção em níveis (sem retenção, baixa retenção, média retenção e alta retenção).
3. **Evasão em binário** ('EVADIDO'): atributo binário da evasão que mostra se o aluno se evadiu ou não.
4. **Evasão em categorias** ('CATEGORIA_EVASAO'): atributo que classifica a evasão em grupos (formado, desistência, jubramento, transferência e transferência interna).
5. **Situação na disciplina**: atributo que indica a situação de um aluno em uma disciplina como aprovado, reprovado e reprovado por frequência;


Para cada experimento realizado, foi necessário uma separação diferente da base, tanto para o estudo da **retenção** quanto para o da **evasão** utilizou-se o atributo 'PERIODO_PORCENTAGEM'.

Para os experimentos aplicados na retenção, foram consideradas todas as instâncias do conjunto de dados, ao total 3.950 instâncias analisadas. Já para o

experimento referentes a evasão, não foram considerados os alunos que ainda estão cursando, restando, dessa maneira, um total de 2.712 instâncias.

Em ambos os experimentos, cada instância representa um aluno e os dados deles referente àquele período, como exemplificado na Figura 8. Cada linha da tabela é relativa a um período do aluno, como, por exemplo, um estudante que está no 4º período terá 4 linhas com informações dos 4 períodos que ele cursou, tais como: quantidade de disciplinas cursada, coeficiente de rendimento e os demais atributos que foram mostrados no Quadro 5.

Figura 8 - Instâncias de um estudante em cada período.

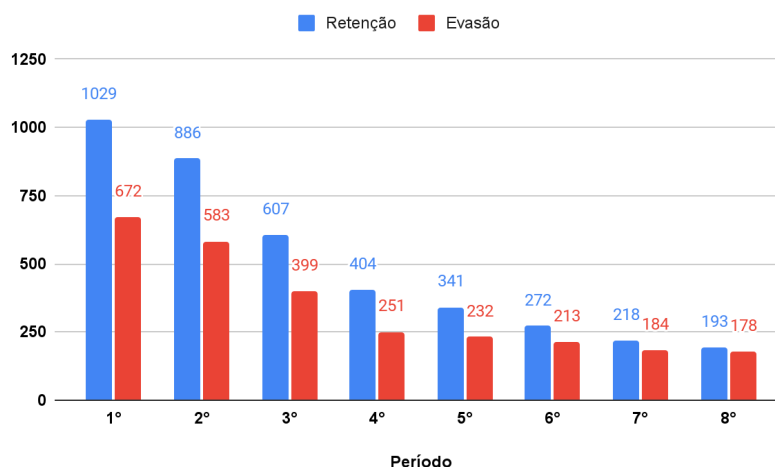


Período	Gênero	Disciplinas cursadas	Percentual de frequência	Coeficiente de rendimento
1	M	8	92%	8,52
2	M	5	85%	8,15
3	M	7	78%	7,25
4	M	6	89%	8,24

Fonte: Elaboração própria.

A Figura 9 mostra um gráfico de barras com quantidade de instâncias para retenção e evasão em cada período do curso. O eixo vertical representa a quantidade de instâncias, enquanto que o eixo horizontal representa os períodos do 1º até 8º. A linha com círculos mostra a distribuição das 3.950 instâncias da Retenção nos 8º períodos. De igual modo, a linha com os losangos mostra a distribuição das 2.712 instâncias da Evasão nos 8 períodos.

Figura 09 - Instâncias em cada período do experimento da retenção e evasão.

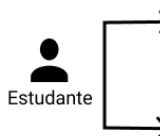


Fonte: Elaboração própria.

Já para o experimento realizado para a **situação da disciplina** dos estudantes, diferentemente da anterior, foi realizado um filtro para ordenar todas as disciplinas da grade, separando-as de acordo com o atributo 'PERIODO_IDEAL' para se cursar seguindo a grade curricular. Por exemplo, as disciplinas que devem ser cursadas no 1º período foram agrupadas para a base de dados do 1º período.

Ao contrário dos experimentos anteriores, cada instância da base de dados é referente a uma disciplina cursada pelo aluno durante a graduação, por isso a quantidade de instâncias é superior. Por exemplo, se o aluno cursar 24 disciplinas, ele possuirá 24 linhas na base de dados relacionadas a ele, no qual os atributos que mudam são aqueles relacionados à disciplina, como a média final e o número de faltas, como mostrado na Figura 10 de forma resumida.

Figura 10 - Instâncias das disciplinas referente a um estudante.

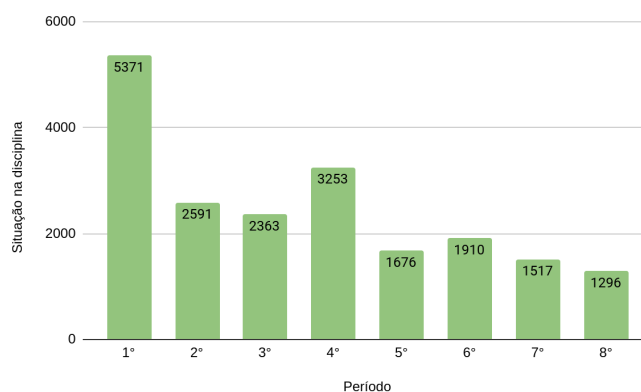


Período	Gênero	Nome disciplina	Número faltas	Situação disciplina
1	F	Lógica para Computação	5	Aprovado
1	F	Matemática Discreta	1	Aprovado
5	F	Pesquisa Operacional	8	Reprovado
4	F	Estrutura de Dados	2	Aprovado

Fonte: Elaboração própria.

Ao total, foram analisadas 19.977 instâncias. Como mostrada no gráfico de linhas da Figura 11, a linha vertical representa a quantidade de instâncias, enquanto que a linha horizontal representa os 8 períodos do curso, sendo possível observar a quantidade de instâncias presentes em cada período.

Figura 11 - Instâncias em cada período do experimento da situação da disciplina.



Fonte: Elaboração própria.

3.2 DISCUSSÃO E AVALIAÇÃO DOS RESULTADOS

Nesta subseção será discutida a última etapa da metodologia, que consiste na interpretação e na avaliação dos resultados obtidos no decorrer da pesquisa referente a mineração de dados.

Como mostrado na seção anterior, a base de dados foi dividida por períodos, ou seja, as instâncias foram distribuídas em 8 partes. Nos experimentos realizados para a retenção e a evasão, a base de dados foi dividida de acordo com atributo 'PERIODO_PORCENTAGEM'. Já no experimento referente à situação do aluno na disciplina, foi usado o 'PERIODO_IDEAL', onde a diferença dos períodos utilizados estão conceituadas no Quadro 5.

Por fim, para cada período, foram avaliados 6 algoritmos da tarefa de classificação, com a aplicação dos conceitos presentes na Seção 2.1.3.

3.2.1 Retenção

O **primeiro experimento** realizado para o estudo inicial da retenção classificou as 3.950 instâncias disponíveis em duas classes 'SIM' e 'NÃO', para o atributo criado 'RETIDO'. Essa classificação foi realizada de acordo com o atributo 'DURACAO_CURSO', onde, caso o aluno tenha mais de 4 anos de duração no curso, é rotulado como 'SIM' no atributo alvo e, caso não seja, será rotulado como 'NÃO'.

A Tabela 4 mostra o resultado da acurácia para cada algoritmo avaliado. A primeira coluna mostra o período de divisão, a segunda coluna mostra a classe majoritária obtida através do algoritmo *ZeroR* e sua porcentagem e as últimas

colunas mostram a acurácia para cada um dos algoritmos. As últimas duas linhas da Tabela 4 retratam as duas métricas gerais: a média aplicada às acurácias de cada algoritmo e a média das posições do *ranking* do desempenho dos algoritmos para cada um dos períodos. Além disso, os melhores resultados nas médias e em cada período estão destacados em negrito.

Tabela 4 - Acurácia dos algoritmos para previsão da retenção.

Atributo alvo: RETIDO (Sim, Não)							
Período	Classe Majoritária (%)	Acurácia dos Algoritmos (%)					
		Naive Bayes	SMO	IBK	PART	J48	Random Forest
1°	NAO (56,46)	65,01	73,08	73,08	75,22	77,36	80,08
2°	SIM (66,25)	57,67	76,19	74,49	77,43	79,91	78,67
3°	SIM (75,78)	61,94	79,74	76,94	80,72	80,89	83,20
4°	SIM (75,00)	62,87	79,95	69,55	77,97	79,46	81,93
5°	SIM (85,04)	70,97	91,20	89,44	88,56	88,86	92,08
6°	SIM (84,56)	77,94	88,97	86,76	87,50	88,60	91,54
7°	SIM (90,83)	85,32	94,04	92,20	94,95	93,58	94,04
8°	SIM (90,67)	90,16	96,89	96,89	92,75	93,26	95,34
Média (%)	SIM(7%)	71,49	85,01	82,42	84,39	85,24	87,11
Média do Ranking	-----	5,63	2,63	4,00	3,38	2,63	1,25

Fonte: Elaboração própria.

Com relação à média dos algoritmos e à média do *ranking*, o *Random Forest* apresentou o melhor desempenho com 87,11% de acurácia, em média, e consequentemente, a melhor média no *ranking* com 1,25. No entanto, o algoritmo *SMO* e *IBK* apresentaram o melhor resultado individual, chegando a 96,89% de acurácia no 8° período.

A segunda coluna da Tabela 4 mostra o resultado do algoritmo *ZeroR* que retrata a classe majoritária para cada período. É possível analisar que a partir do 2° período ocorre a predominância da retenção, com a porcentagem do rótulo 'SIM' aumentando gradualmente nos demais períodos, começando com 66,25% e chegando em 90,67% no 8° período.

Como o algoritmo *Random Forest* apresentou a melhor média geral de acurácia, foi utilizado o método implementado na ferramenta WEKA, *InfoGainAttributeEval*, que serve para avaliar o valor de um atributo, medindo o ganho de informação dos atributos em relação à classe por meio de um *ranking*. Desse modo, foram realizados testes excluindo os últimos atributos do *ranking* para avaliar se a acurácia do modelo poderia ser melhorada.

A Tabela 5 mostra o resultado da seleção de atributos para todos os 8 períodos. Na segunda coluna está a acurácia do *Random Forest* inicial e, na terceira coluna, está a acurácia melhorada desse algoritmo com a seleção de atributos obtidos através do *InfoGainAttributeEval*, aplicando a exclusão dos atributos menos relevantes. Já a quarta coluna da Tabela 5 mostra a porcentagem da melhoria obtida com a seleção de atributos com base no máximo que era possível ser melhorado. O cálculo é feito com a subtração entre a acurácia melhorada e a acurácia inicial dividida pela quantidade que falta para acurácia original chegar a 100, como mostrado na equação(10). A mesma equação foi aplicada para os demais experimentos.

$$Melhoria = \frac{(acurácia\ melhorada - acurácia\ inicial) * 100}{100 - acurácia\ inicial} \quad (10)$$

A quinta coluna mostra os atributos usados para melhoria da acurácia (a numeração dos atributos refere-se ao ID do Quadro 5), ou seja, aqueles que permaneceram sem serem excluídos, sendo ordenados de acordo com o ranking do *InfoGainAttributeEval*.

Tabela 5 - Seleção de atributos do 1º para previsão da retenção.

Atributo alvo: RETIDO (Sim, Não)				
Período	RF (%)	RF com seleção de atributos (%)	Melhoria (%)	Atributos utilizados (Id)
1º	80,08	80,08	0,00	5, 26, 28, 29, 23, 25, 24, 12, 14, 20, 13, 27, 30, 21, 15, 17, 3, 22, 2, 16, 18, 19
2º	78,67	80,47	8,44	5, 30, 28, 25, 15, 23, 24, 16, 19, 12, 14, 21, 20, 26, 3, 22, 2, 27, 18
3º	83,20	85,01	10,77	5, 28, 29, 25, 23, 30, 15, 24, 21, 14
4º	81,93	82,67	4,10	5, 25, 28, 24, 29, 23, 30, 14, 12, 27, 15, 16, 17, 21, 22, 19, 20, 3, 2, 26

5°	92,08	92,38	3,79	5, 28, 30, 23, 29, 25, 12, 24, 21, 16, 14, 20, 17, 15, 13, 27, 19, 3, 2
6°	91,54	91,54	0,00	25, 23, 5, 14, 12, 29, 28, 21, 15, 30, 17, 16, 24, 20, 27, 13, 18, 19, 22, 3, 2, 26
7°	94,04	94,04	0,00	28, 25, 23, 27, 29, 30, 15, 14, 12, 24, 21, 19, 13, 16, 5, 17, 22, 18, 20, 3, 2, 26
8°	95,34	95,85	10,94	19, 28, 29, 15, 25, 23, 22, 27, 30, 12, 14, 16, 24, 5, 13, 21, 20, 17, 3, 2, 18
Média (%)	87,11	87,76	4,75	-----

Fonte: Elaboração própria.

Como mostra a Tabela 5, dentre os 8 períodos, foi possível obter uma melhoria de acurácia em 5 deles, com o melhor resultado no último período obtendo uma melhoria de 10,94% da acurácia.

O **segundo experimento** realizado ainda no contexto da retenção buscou categorizar o experimento anterior, com as mesmas 3.950 instâncias. Para essa categorização, também foi utilizada o atributo 'DURACAO_CURSO', para a criação do atributo alvo 'CATEGORIA_RETENCAO', dividido da seguinte maneira:

1. **Sem retenção:** alunos que ainda possuem o 'DURACAO_CURSO' menor ou igual a 4 anos.
2. **Baixa retenção:** alunos que possuem 'DURACAO_CURSO' igual a 5 anos.
3. **Média retenção:** alunos que possuem 'DURACAO_CURSO' igual a 6 ou 7 anos.
4. **Alta retenção:** alunos que possuem 'DURACAO_CURSO' maior que 7 anos.

Da mesma forma que o experimento anterior, a Tabela 6 mostra as acurácias obtidas pelos 6 algoritmos utilizados para cada período do curso. Além disso, nas últimas linhas é possível ver as médias de acurácia e de *ranking* para cada algoritmo e destacados em negrito as acurácias mais alta dos algoritmos por período.

Tabela 6 - Acurácia dos algoritmos para previsão da categoria da retenção.

Atributo Alvo: CATEGORIA_RETENCAO (Sem Retenção, Alta Retenção, Média Retenção, Baixa Retenção)							
Período	Classe Majoritária (%)	Acurácia dos Algoritmos (%)					
		Naive Bayes	SMO	IBK	PART	J48	Random Forest
1°	SEM RETENÇÃO (56,46)	47,04	61,32	59,67	64,04	65,79	69,39
2°	SEM RETENÇÃO (33,75)	40,18	51,81	50,56	51,92	59,14	58,92
3°	ALTA RETENÇÃO (29,65)	41,52	52,22	49,75	57,50	57,83	64,91
4°	ALTA RETENÇÃO (27,72)	42,57	50,99	39,36	54,21	56,68	61,14
5°	ALTA RETENÇÃO (38,71)	51,91	67,74	62,76	70,09	69,79	75,07
6°	ALTA RETENÇÃO (45,59)	52,57	67,65	65,07	67,28	71,32	75,37
7°	ALTA RETENÇÃO (43,12)	55,96	71,10	60,55	71,10	73,39	78,90
8°	ALTA RETENÇÃO (43,01)	77,20	79,79	75,13	82,38	80,83	86,01
Média (%)	ALTA RETENÇÃO (6/8)	51,12	62,83	57,86	64,82	66,85	71,21
Média do Ranking	-----	5,63	2,63	4,00	3,38	2,63	1,25

Fonte: Elaboração própria.

A Tabela 6 mostra que o *Random Forest* continuou apresentando a melhor média geral, com 71,21% de acurácia, e, conseqüentemente, a melhor média do *ranking*, com 1,25. Este algoritmo também obteve o melhor resultado individual, com 86,01% de acurácia no 8° período.

Já a classe majoritária mostra que, nos dois primeiros períodos do curso, é o rótulo ‘SEM RETENÇÃO’ que predomina. No entanto, vale ressaltar que as demais classes correspondem a uma subdivisão da categoria do primeiro experimento ‘RETIDO’ (BAIXA RETENÇÃO, MÉDIA RETENÇÃO, ALTA RETENÇÃO) e por isso o ‘SEM RETENÇÃO’ foi predominante no início. Contudo, a partir do 3° período, a classe ‘ALTA RETENÇÃO’ ultrapassou a predominância com 29,65% e apresentou uma queda no 4° período para 27,72% e com a maior porcentagem de 45,59% no 5° período do curso.

Do mesmo modo que o experimento anterior, como mostrado na Tabela 6, o algoritmo que obteve a melhor acurácia foi *Random Forest (RF)* e por isso foi aplicado um teste para melhorar essa acurácia em todos os períodos, como mostrado na Tabela 7.

A segunda coluna da Tabela 7 mostra a acurácia inicial do *Random Forest*, enquanto a terceira coluna mostra o *Random Forest* com a seleção de atributos é a acurácia melhorada obtida através do *InfoGainAttributeEval* e aplicando a exclusão dos atributos menos relevantes. A última coluna mostra os atributos que foram utilizados para alcançar essa acurácia (a numeração dos atributos refere-se ao ID do Quadro 5).

Tabela 7 - Seleção de atributos para previsão categoria da retenção.

Atributo Alvo: CATEGORIA_RETENCAO (Sem Retenção, Alta Retenção, Média Retenção, Baixa Retenção)				
Período	RF (%)	RF com seleção de atributos (%)	Melhoria (%)	Atributos utilizados
1°	69,39	70,65	4,12	5, 26, 12, 24, 28, 14, 25, 27, 23, 20, 21, 13, 3, 15, 16, 22, 30, 29, 19
2°	58,92	58,92	0,00	5, 23, 30, 25, 28, 21, 15, 14, 12, 16, 27, 20, 3, 13, 19, 2, 22, 24, 26, 17, 18, 29
3°	64,91	66,72	5,16	5, 28, 25, 23, 29, 30, 27, 24, 15, 21, 14, 20, 3, 12, 16, 17, 22, 2, 13, 18
4°	61,14	64,36	8,29	5, 28, 29, 25, 30, 23, 14, 12, 21, 24, 27, 15, 17, 16, 22, 19, 20, 3, 2, 26, 13
5°	75,07	76,25	4,73	5, 23, 25, 28, 30, 29, 27, 24, 12, 21, 14, 20, 17, 15, 13, 18, 16, 19, 22, 3, 2
6°	75,37	78,31	11,94	23, 25, 28, 30, 5, 14, 12, 21, 29, 15, 16, 20, 17, 24, 27, 13, 19, 18, 22, 3, 2
7°	78,9	78,90	0,00	28, 25, 23, 30, 27, 15, 5, 24, 19, 29, 21, 22, 14, 12, 16, 13, 17, 20, 18, 3, 2, 26
8°	86,01	92,23	44,46	25, 28, 23, 30, 27, 15, 19, 29, 22, 24, 5, 14
Média (%)	71,21	73,29	9,84	-----

Fonte: Elaboração própria.

Vale ressaltar que, com essa aplicação da seleção de atributos, foi possível obter melhoria em 7 dos 8 períodos, como mostrado na quarta coluna da Tabela 7. O melhor resultado alcançado foi de 44,46% no 8° período, mostrando a capacidade de melhorar significativamente a acurácia do modelo preditivo.

Em conclusão às análises, o **primeiro** e o **segundo experimento**, relativos a retenção, apresentaram resultados que, a partir do primeiro ano do curso, principalmente com relação ao 1° período, a taxa de retenção é inferior aos demais

períodos. Dessa maneira, pode-se levantar a seguinte hipótese: os atrasos obtidos pelos alunos nos primeiros períodos são refletidos nos períodos consecutivos.

Como um exemplo da hipótese, boa parte das disciplinas referente ao primeiro período são pré-requisitos para os demais, ou seja, caso o aluno reprove em uma disciplina pré-requisito, conseqüentemente, já estaria retido por no mínimo um ano, pois teoricamente a disciplina só seria ofertada novamente no ano seguinte.

De acordo com os resultados da seleção de atributos presentes nas Tabelas 5 e 7, os atributos que aparecem mais frequentemente ocupando as posições mais importantes do *ranking* são: 'HISTORICO_GERAL_REPROVADAS_QUANTIDADE', 'FORMA_INGRESSO', 'HISTORICO_GERAL_REPROVADAS_MEDIA', 'HISTORICO_GERAL_MEDIA', 'HISTORICO_GERAL_CR', 'HISTORICO_GERAL_CURSADAS' com a descrição de cada atributo mostrado no Quadro 4.

Os algoritmos apresentaram uma melhora na acurácia, no qual todos alcançaram acima de 75% no 8º período. Um dos fatores que podem estar relacionados a essa acurácia, são os atributos do histórico do desempenho do aluno geral e recente (conceituados no Quadro 4) e como mostrado na Tabela 5 e 7, eles permaneceram após a seleção de atributos mostrando-se relevantes para os modelos de previsão, principalmente o histórico geral que utiliza dados cumulativos dos períodos anteriores.

Dessa forma, após o primeiro ano da graduação, como mostrado na Tabela 6, começando no 3º período, a alta taxa de retenção é predominante, ou seja, os alunos ficam altamente retidos nos demais períodos, postergando a conclusão do curso.

3.2.2 Evasão

O **terceiro experimento** realizado foi referente à análise da evasão do curso. Neste experimento não foram consideradas instâncias com alunos que ainda estão

cursando e, dessa forma, o número de instâncias utilizadas foi de 2.712 instâncias. Para realizar a classificação, foi criado um novo atributo baseado no atributo 'FORMA_EVASAO', que já estava presente na base de dados original, mas voltado para identificar casos que saíram do curso. A partir disso, os alunos que tinham o rótulo 'FORMADO' foram classificados como 'NÃO' no atributo criado 'EVADIDO' e os demais foram automaticamente rotulados como 'SIM'.

Conforme mostrado na Tabela 8, tem-se como resultado as acurácias obtidas dos 6 algoritmos utilizados para cada período do curso. A segunda coluna mostra o resultado do algoritmo *ZeroR* (classe majoritária e porcentagem) e as demais colunas mostram a acurácia para cada um dos algoritmos testados. Nas últimas linhas, encontram-se as duas métricas obtidas das acurácias e destacados em negrito as acurácias mais alta dos algoritmos em cada período.

Tabela 8 - Acurácia dos algoritmos para previsão da evasão.

Atributo alvo: EVADIDO (Sim, Não)

Período	Classe Majoritária (%)	Acurácia dos Algoritmos (%)					
		Naive Bayes	SMO	IBK	PART	J48	Random Forest
1°	SIM (90,03)	81,55	90,92	87,05	88,99	90,63	90,63
2°	SIM (74,44)	77,36	83,88	78,9	84,73	83,88	85,59
3°	SIM (67,42)	78,70	83,96	74,94	84,21	80,45	84,21
4°	SIM (51,79)	76,49	78,09	74,10	78,88	78,09	80,08
5°	NÃO (52,59)	81,47	82,33	76,29	78,45	78,45	81,03
6°	NÃO (62,91)	81,22	86,38	80,28	84,04	84,98	86,38
7°	NÃO (88,59)	79,35	88,59	88,59	91,85	96,74	91,85
8°	NÃO (90,45)	85,96	92,13	92,7	94,38	93,26	97,19
Média (%)	SIM (¼)	80,26	85,79	81,61	85,69	85,81	87,12
Média do Ranking	-----	5,63	2,63	4,00	3,38	2,63	1,25

Fonte: Elaboração própria.

O algoritmo que alcançou a melhor média geral e a média do *ranking* continuou sendo o *Random Forest* com 87,12% e 1,25%, respectivamente. Da mesma forma, ele também obteve a melhor média individual para o 7° período (97,19% de acurácia).

Além disso, para os resultados do algoritmo *ZeroR*, verifica-se que, do 1º período até o 4º período, no atributo 'EVADIDO', a classe predominante é 'SIM', mas possui uma queda gradativa da porcentagem, na qual inicia em 90,03% no 1º período e diminui para 51,79% no 4º período. Logo em seguida, a partir do 5º período, no atributo 'EVADIDO', a classe majoritária é o 'NÃO', com 52,59%, mas alcançando 90,45% de predominância no 8º período.

Conforme mostrado na Tabela 9, também foi aplicado o método para verificar a possibilidade de melhoria da acurácia para o algoritmo que alcançou a melhor média geral, no caso o *Random Forest*. A coluna RF é a acurácia inicial utilizando todos os atributos, já a coluna RF com seleção de atributos é a precisão aprimorada obtida por meio da exclusão de atributos menos relevantes obtidos através do *InfoGainAttributeEval*. A quarta coluna é a porcentagem de melhoria em relação ao algoritmo inicial, com a porcentagem do aumento entre as duas acurácias. A última coluna apresenta os atributos utilizados para se obter essa acurácia, ordenados de acordo com o *ranking*, no qual estão organizados de acordo com a numeração dos IDs do Quadro 5).

Tabela 9 - Seleção de atributos para previsão da evasão.

Atributo alvo: EVADIDO (Sim, Não)				
Período	RF (%)	RF com seleção de atributos (%)	Melhoria (%)	Atributos utilizados
1º	90,63	91,82	12,70	25, 28, 12, 14, 20, 23, 15, 17, 21, 24, 16, 26, 13, 29, 27, 30, 19, 22, 5, 3
2º	85,59	86,45	5,97	23, 25, 28, 14, 12, 15, 21, 17, 29, 20, 30, 24, 13, 16, 27, 19, 22, 18, 5, 3
3º	84,21	85,21	6,33	23, 28, 25, 30, 12, 24, 14, 20, 15, 17, 21, 13, 29, 16, 27, 5, 19, 3, 2, 22
4º	80,08	80,08	0,00	28, 23, 25, 12, 14, 30, 15, 21, 20, 24, 17, 13, 29, 16, 19, 22, 27, 5, 18, 26, 3, 2
5º	81,03	81,03	0,00	28, 30, 12, 23, 25, 15, 24, 20, 21, 14, 13, 17, 16, 29, 19, 27, 18, 5, 22, 3, 2, 26
6º	86,38	87,32	6,90	23, 28, 30, 12, 14, 25, 24, 15, 16, 20, 21, 13, 29, 17, 5, 27, 19, 22, 26, 2, 3, 18
7º	91,85	94,02	26,63	28, 16, 5, 27, 25, 13, 30, 23, 21
8º	97,19	97,19	0,00	12, 14, 20, 13, 17, 23, 21, 25, 28, 27, 30, 24, 15, 16, 26, 5, 19, 22, 2, 3, 29, 18
Média (%)	87,12	87,89	7,32	-----

Fonte: Elaboração própria.

Com essa seleção de atributos, foi possível alcançar uma melhoria na acurácia em 5 dos 8 períodos, no qual o 7º período mostrou a porcentagem mais alta, com uma melhoria de 26,63% de acurácia.

O **quarto experimento** realizado, ainda no contexto da evasão, utilizou a mesma base de dados com 2712 instâncias, e também utilizou o atributo 'FORMA_EVASAO'. Entretanto, neste experimento foram utilizados os rótulos que já estavam definidos na base de dados original, que estavam categorizados da seguinte maneira:

1. **Formado:** aluno que concluiu e finalizou todas as atividades curriculares do curso.
2. **Desistência:** aluno que abandonou o curso antes de concluir as atividades curriculares.
3. **Jubilamento:** quando o aluno passou do limite máximo de tempo para finalizar todas as atividades curriculares do curso ou não renovou a matrícula por 2 semestres seguidos.
4. **Transferido:** aluno realizou a transferência para outra instituição de ensino.
5. **Transferência interna:** aluno que foi transferido para outro curso dentro da mesma instituição de ensino.

Os resultados das acurácias para cada período dos algoritmos utilizados estão disponibilizados na Tabela 10. A segunda coluna mostra a classe majoritária para cada período e a respectiva porcentagem e, como nos demais experimentos, as últimas linhas correspondem à média de acurácia e de *ranking* para cada algoritmo, respectivamente e destacados em negrito as acurácias mais alta dos algoritmos por período.

Tabela 10 - Acurácia dos algoritmos para previsão da categoria da evasão.

Atributo alvo: CATEGORIA_EVASAO							
(Formado, Desistência, Jubilamento, Transferido, Transferência Interna)							
Período	Classe Majoritária (%)	Acurácia dos Algoritmos (%)					
		Naive Bayes	SMO	IBK	PART	J48	Random Forest
1°	JUBILADO (62,35)	54,32	73,51	64,58	75,45	76,79	76,04
2°	JUBILADO (49,91)	57,12	68,95	61,58	73,76	77,53	77,87
3°	JUBILADO (51,63)	64,91	73,18	68,67	80,95	78,70	79,95
4°	FORMADO (48,21)	63,35	68,13	64,14	71,31	66,14	73,31
5°	FORMADO (52,59)	73,71	76,72	72,84	75,00	81,47	78,02
6°	FORMADO (62,91)	81,22	84,98	79,34	85,45	83,57	86,85
7°	FORMADO (88,59)	79,89	88,59	88,04	90,22	93,48	91,30
8°	FORMADO (90,45)	87,64	91,57	92,13	95,51	93,26	97,19
Média (%)	FORMADO(5%)	70,27	78,20	73,92	80,96	81,37	82,57
Média do Ranking	-----	5,63	2,63	4,00	3,38	2,63	1,25

Fonte: Elaboração própria.

A média geral e a média do *ranking* mais alta de acurácia continuou sendo do algoritmo *Random Forest* com 82,57% e 1,25 respectivamente e também a com a média individual mais alta de 97,19% no 8° período.

Do 1° período até o 3° período, o rótulo predominante do atributo 'CATEGORIA_EVASAO' é o 'JUBILAMENTO', chegando em 62,35% de predominância no 1° período. Em seguida, do 4° período até o 8° período, o rótulo predominante foi o 'FORMADO', atingindo 90,45% no último período do curso, ou seja, para aqueles que permanecem nos períodos finais do curso, nota-se que predomina uma maior porcentagem de formados.

Conforme os dados mostrados na Tabela 11, com todos os 8 períodos, foi aplicada uma estratégia de seleção de atributos para tentar melhorar a acurácia do algoritmo que obteve a melhor média geral (*Random Forest*). A segunda coluna mostra a acurácia inicial do algoritmo, e, na terceira coluna, o *RF* com seleção de atributos mostra a acurácia obtida ao excluir atributos menos relevantes, conforme resultado obtido pelo *InfoGainAttributeEval*. Por fim, a quarta coluna mostra a porcentagem da melhoria obtida, com base no cálculo explicado na equação (10) e a

quinta coluna mostra a sequência de atributos que permaneceram, com base nos IDs mostrados no Quadro 5.

Tabela 11 - Seleção de atributos para previsão da categoria da evasão.

Atributo alvo: CATEGORIA_EVASAO (Formado, Desistência, Jubilamento, Transferido, Transferência Interna)				
Período	RF (%)	RF com seleção de atributos (%)	Melhoria (%)	Atributos utilizados
1°	76,04	77,98	8,10	25, 23, 14, 12, 20, 28, 15, 24, 17, 29, 21, 27, 16, 13, 26, 5, 30, 3, 19, 22, 2, 18
2°	77,87	77,87	0,00	25, 23, 28, 12, 29, 14, 21, 20, 15, 17, 24, 30, 27, 13, 5, 16, 19, 22, 18, 2, 3, 26
3°	79,95	80,45	2,49	28, 25, 23, 24, 30, 12, 14, 20, 15, 21, 17, 27, 29, 5, 13, 16, 19, 3, 2, 22, 26
4°	73,31	73,71	1,50	28, 23, 12, 14, 15, 25, 21, 24, 20, 17, 30, 29, 13, 16, 22, 19, 27, 5, 2, 3, 26, 18
5°	78,02	81,03	13,69	28, 30, 23, 25, 12, 24, 29, 15, 21, 20, 14, 13, 27, 17, 16
6°	86,85	87,79	7,15	23, 28, 12, 14, 25, 30, 15, 16, 20, 21, 24, 13, 29, 17, 5, 19, 27, 22, 2, 3
7°	91,3	92,93	18,74	27, 28, 16, 14, 5, 21, 30, 25, 20, 22, 2, 3, 29, 26, 17, 24, 12, 13, 19
8°	97,19	98,31	39,86	27, 23, 29, 12
Média (%)	82,57	83,76	11,44	-----

Fonte: Elaboração própria.

Como visto na Tabela 11, com a exclusão dos atributos menos relevantes, foi possível alcançar uma melhoria em 7 dos 8 períodos da base, com o 8° período alcançando uma melhoria de 39,86%.

Em conclusão às análises, o **terceiro** e o **quarto experimento**, que são referentes a evasão, apontaram que, diferentemente da retenção, a evasão ocorre predominantemente nos primeiros 2 anos do curso. Como mostrado na Tabela 11, até o 3° período, a evasão por jubilamento é a classe majoritária, podendo ser levantadas as seguintes hipóteses: o período de integralização do curso foi excedido pelo aluno ou a matrícula não foi renovada por dois semestres consecutivos.

A primeira hipótese corresponde aos casos em que os estudantes possuem dificuldade com as disciplinas iniciais da graduação e permanecem retidos por um

longo tempo. Com isso, eles excedem o prazo máximo de integralização do curso e isso resulta em um jubilamento.

Já na segunda hipótese, o estudante pode ter abandonado o curso sem solicitação formal, permanecendo com a matrícula mas não a renovando por dois semestres consecutivos. De semelhante modo, o estudante também pode ter realizado o trancamento por dois semestres consecutivos. Em ambos os casos, ocorre o jubilamento.

Também pode-se deduzir que, a partir dos resultados da Tabela 11, caso o aluno conclua o 3º período, as chances dele finalizar o curso são maiores, pois o números apontam que o rótulo 'FORMADO', a partir do 4º período representam valores acima de 48%, ultrapassando cada categoria relacionada a evasão. No entanto, em valores absolutos, os formados no 8º período (90,45%) correspondem à mesma quantidade dos formados no 4º período (48,21%), ou seja, isso indica que o rótulo 'FORMADO' começa a predominar a partir do segundo ano de curso.

Além dessas hipóteses, com base nos resultados da seleção de atributos das Tabelas 9 e 11, os atributos que aparecem mais frequentemente nas posições iniciais do *ranking* são: 'HISTORICO_GERAL_REPROVADAS_QUANTIDADE', 'HISTORICO_RECENTE_CR', 'HISTORICO_GERAL_MEDIA', 'HISTORICO_RECENTE_MEDIA', com a descrição de cada atributo mostrado no Quadro 4.

Os modelos gerados pelos algoritmos de classificação obtiveram uma alta taxa de acurácia principalmente nos períodos finais do curso. Em ambos os experimentos foi possível obter uma média de acurácia acima de 70% em todos os algoritmos. Destacando que o melhor algoritmo, o *Random Forest*, alcançou a média de acurácia de 87,12% no experimento 'EVADIDO' e 82,57% e no experimento 'CATEGORIA_EVASAO'. Além disso, foi possível melhorar 7,32% e 11,44% respectivamente a média das acurácias, do algoritmo *Random Forest*, utilizando a seleção de atributos com o método *InfoGainAttributeEval*.

3.2.2 Situação da disciplina

O **quinto experimento** está relacionado à situação das disciplinas cursadas pelos estudantes no decorrer do curso. Para esse experimento, a base possui 19.977 instâncias, pois cada linha é referente a uma disciplina cursada por um aluno. O atributo alvo utilizado foi 'SITUACAO_DISCIPLINA', que já estava disponível na base de dados original. Os rótulos desse atributo que foram considerados para a pesquisa foram: 'APROVADO', 'REPROVADO' e 'REPROVADO_FREQUÊNCIA'. A divisão da base foi feita de acordo com o 'PERIODO_IDEAL', que é referente à divisão das disciplinas de acordo com a grade curricular do curso.

A Tabela 12 mostra a acurácia obtida por cada algoritmo utilizado nos 8º períodos do curso. A primeira coluna mostra o período da divisão, a segunda coluna indica a classe majoritária e a porcentagem de acurácia e as demais colunas mostram a acurácia para cada um dos algoritmos aplicados. Por fim, as últimas linhas mostram a média das acurácias e do *ranking* para cada algoritmo, respectivamente e destacados em negrito as acurácias mais alta dos algoritmos por período.

Tabela 12 - Acurácia dos algoritmos para previsão de aprovação na disciplina.

Atributo alvo: SITUACAO_DISCIPLINA (Aprovado, Reprovado, Reprovado por frequência)							
Período	Classe Majoritária (%)	Acurácia dos Algoritmos (%)					
		Naive Bayes	SMO	IBK	PART	J48	Random Forest
1º	APROVADO (55,20)	70,60	78,96	74,75	76,09	78,14	79,97
2º	APROVADO (58,43)	74,84	80,86	77,34	77,69	80,47	82,63
3º	APROVADO (59,29)	72,37	80,15	75,96	78,12	81,34	82,35
4º	APROVADO (58,96)	75,28	80,73	76,08	76,33	78,27	80,08
5º	APROVADO (63,31)	76,07	80,67	76,73	77,74	76,67	78,88
6º	APROVADO (75,45)	81,99	88,22	84,66	84,76	86,81	87,91
7º	APROVADO (77,32)	82,07	88,40	86,75	86,62	88,73	88,53
8º	APROVADO (74,92)	82,25	88,66	84,57	85,73	88,35	88,19
Média (%)	APROVADO (8/8)	76,93	83,33	79,61	80,39	82,35	83,57

Média do Ranking	-----	6,00	1,75	4,75	4,00	2,75	1,75
-------------------------	-------	------	-------------	------	------	------	-------------

Fonte: Elaboração própria.

Analisando as médias mostradas na Tabela 12, observa-se que a média de acurácia geral mais alta foi do algoritmo *Random Forest* (88,19%). Já na média do *ranking*, houve um empate entre os algoritmos SMO e o *Random Forest* com 1,75, mas, na média individual, a mais alta foi do algoritmo SMO no 8º período .

Já ao analisar a segunda coluna da Tabela 12, compreende-se que a classe majoritária foi com o rótulo 'APROVADO', iniciando em 55,20% para o 1º período e alcançando o valor máximo de 77,32% para o 8º período. Apesar disso, vale ressaltar que, embora os aprovados sejam a maioria, o índice de reprovação ainda é altíssimo, porque, se considerar que no 1º período as reprovações são em torno de 44,80%, chega-se à conclusão de que quase metade das disciplinas cursadas resultam em reprovação.

Como o algoritmo *Random Forest* obteve a melhor acurácia em média geral, a estratégia de seleção de atributos *InfoGainAttributeEval* foi aplicada e a Tabela 13 apresenta a melhoria da acurácia do algoritmo. A primeira coluna representa o período, a segunda coluna apresenta o resultado do *Random Forest* inicialmente, a terceira coluna mostra o resultado do algoritmo obtido com a remoção dos atributos menos relevantes, a quarta coluna indica a melhoria da nova execução do algoritmo e, por último, a quinta coluna informa os atributos de maior relevância que foram mantidos.

Tabela 13 - Seleção de atributos para previsão de aprovação na disciplina.

Atributo alvo: SITUACAO_DISCIPLINA (Aprovado, Reprovado, Reprovado por frequência)				
Período	RF (%)	RF com seleção de atributos (%)	Melhoria (%)	Atributos utilizados
1º	79,97	80,60	3,15	12, 14, 15, 23, 25, 20, 13, 17, 28, 24, 21, 16, 26, 27, 29, 18, 30, 19, 22, 5, 3, 9
2º	82,63	83,48	4,89	12, 14, 13, 20, 15, 17, 23, 25, 24, 21, 28, 16, 29, 26, 18, 30, 27, 19, 9
3º	82,35	83,66	7,42	12, 14, 15, 20, 13, 17, 21, 16, 25, 23, 24, 28, 29, 18, 30, 26, 9, 19, 27, 5
4º	80,08	80,42	1,71	14, 12, 15, 13, 20, 17, 21, 25, 24, 23, 28,

				16, 29, 26, 18, 30, 19, 22, 27, 5
5°	78,88	78,88	0,00	14, 12, 13, 15, 20, 17, 21, 16, 23, 25, 24, 28, 29, 26, 18, 27, 30, 19, 5, 22, 9, 3, 2
6°	87,91	88,64	6,04	14, 12, 15, 13, 20, 17, 21, 16, 24, 25, 23, 28, 29, 26, 18, 19, 22, 30, 27, 5
7°	88,53	89,52	8,63	14, 12, 13, 20, 15, 17, 16, 21, 25, 28, 23, 24, 29, 18, 26, 19, 27, 22, 30, 5, 9, 3, 2
8°	88,19	88,66	3,98	14, 12, 13, 15, 20, 21, 17, 16, 24, 28, 25, 23, 29, 18, 26, 27, 19, 30, 22, 9
Média (%)	83,57	84,23	4,48	-----

Fonte: Elaboração própria.

A Tabela 13 mostra que foi possível alcançar uma melhoria na acurácia em 7 dos 8 períodos, comprovando a eficácia da remoção de atributos de menor relevância classificados através do método *InfoGainAttributeEval*. De forma individual, o maior aumento de acurácia no 7º período, onde foi alcançado um aumento de 8,63% em relação à execução com todos os atributos da base de dados.

Em conclusão às análises, o **5º experimento** mostrou que o número de aprovações na base de dados são maiores que as reprovações. No entanto, é importante ressaltar que esse resultado complementa os experimentos anteriores, pois, como mostrado no experimento da retenção, o número de retidos no curso é alto em quase todos os períodos.

Em hipótese relacionada à retenção, pode-se inferir que os alunos retidos também podem cursar menos disciplinas devido a eventuais reprovações em pré-requisitos, que impedem com que o aluno curse algumas disciplinas do período seguinte. Além disso, os alunos retidos ainda podem optar por cursar menos disciplinas devido a alguma estratégia, como, por exemplo, para contornar um problema familiar ou de falta de tempo devido a trabalho, conforme apontado em Saldanha (2016).

Com base nos resultados da seleção de atributos da Tabela 13, os atributos que aparecem mais frequentemente nas posições iniciais do *ranking* são: 'HISTORICO_RECENTE_CR', 'HISTORICO_RECENTE_MEDIA', 'HISTORICO_RECENTE_APROVADAS_QUANTIDADE', 'HISTORICO_RECENTE_PERCENTUAL_

FREQUENCIA', "HISTORICO_RECENTE_CLASSE_RENDIMENTO', com a descrição de cada atributo mostrado no Quadro 4.

Além disso, embora a taxa de aprovação nos 4 primeiros períodos do curso sejam maiores que 50%, esse valor ainda indica uma alta taxa de reprovação, que pode ser associada à grande quantidade de jubilamentos nesses períodos iniciais da graduação, conforme apresentado na Seção 3.2.2. Com isso, mesmo que as aprovações predominem, as reprovações podem contribuir para que ocorram as hipóteses levantadas na Seção 3.2.2 relacionadas a esses casos de evasão.

3.3 CONSIDERAÇÕES SOBRE O CAPÍTULO

Em virtude dos dados apresentados no decorrer do capítulo, de acordo com o processo experimental realizado, incluindo desde a seleção dos atributos até as análise dos experimentos, foi finalizada com sucesso a aplicação da tarefa de classificação para análise da retenção e evasão, alcançado uma boa taxa de acurácia nos modelos gerados pelos algoritmos *Naive Bayes*, *SMO*, *IBK*, *J48* e *Random Forest*. Além disso, os resultados obtidos foram melhores que as classes majoritárias dos experimentos e também foi possível utilizar a seleção de atributos para melhorar os resultados dos algoritmos que dessa forma alcançaram acurácias mais altas.

Em todos os experimentos, o *Random Forest* mostrou-se o algoritmo com a média de acurácia mais alta e esse resultado se dá pelo funcionamento do algoritmo, que utiliza a combinação de diversas árvores de decisão de forma aleatória, obtendo uma predição que evita o sobreajuste (*overfitting*) com uma acurácia mais estável.

A Tabela 14 mostra a média dos resultados do *Random Forest* nos 5 experimentos, na segunda coluna se encontra a média das acurácias iniciais alcançadas em cada um dos experimentos e ao lado a média da acurácia obtida

com a seleção de atributos aplicando o método InfoGain Attribute Eval e na última coluna a porcentagem de melhoria entre as duas médias.

Tabela 14 - Médias dos resultados da seleção de atributos do *Random Forest* (RF).

Experimento	RF (%)	RF com seleção de atributos (%)	Melhoria (%)
1° EXPERIMENTO - 'RETIDO'	87,11	87,76	4,75
2° EXPERIMENTO - 'CATEGORIA_RETENCAO'	71,21	73,29	9,84
3° EXPERIMENTO - 'EVADIDO'	87,12	87,89	7,32
4° EXPERIMENTO - 'CATEGORIA_EVASAO'	82,57	83,76	11,44
5° EXPERIMENTO - 'SITUACAO_DISCIPLINA'	83,57	84,23	4,48

Fonte: Elaboração própria.

Complementando os resultados da Tabela 14, o Quadro 6 mostra uma apresentação geral dos 5 primeiros atributos mais frequentes no *ranking* de seleção de atributos para todos os experimentos realizados. A primeira coluna corresponde ao experimento realizado e as duas últimas colunas mostram o nome dos atributos (com a explicação descrita no Quadro) e a frequência que corresponde a quantidade de vezes que o atributo aparece dentre os 5 primeiros no *ranking* de seleção de atributos.

Quadro 6 - Atributos mais relevantes e frequentes por experimento.

Experimento	Atributo	Frequência
1° EXPERIMENTO 'RETIDO'	historico_geral_reprovadas_quantidade	7
	forma_ingresso	6
	historico_geral_reprovadas_media	6
	historico_geral_media	6
	historico_geral_cr	5
2° EXPERIMENTO 'CATEGORIA_RETENCAO'	historico_geral_reprovadas_quantidade	8
	historico_geral_media	7
	forma_ingresso	6
	historico_geral_cr	6
	historico_geral_cursadas	6
3° EXPERIMENTO 'EVADIDO'	historico_geral_reprovadas_quantidade	7
	historico_recente_cr	7

	historico_geral_media	6
	historico_recente_media	5
	historico_geral_cr	5
4° EXPERIMENTO 'CATEGORIA_RETENCAO'	historico_geral_cr	7
	historico_recente_cr	6
	historico_geral_reprovadas_quantidade	6
	historico_geral_media	5
	historico_recente_media	4
5° EXPERIMENTO 'SITUACAO_DISCIPLINA'	historico_recente_cr	8
	historico_recente_media	8
	historico_recente_aprovadas_quantidade	8
	historico_recente_percentual_frequencia	7
	historico_recente_classe_rendimento	7

Fonte: Elaboração própria.

Os resultados do Quadro 6 evidenciam a importância da criação de novos atributos, pois todos os atributos referente ao histórico do desempenho do aluno foram criados a partir dos atributos existentes e contribuíram para o processo de previsão pois aparecem frequentemente como os primeiros no *ranking* na seleção de atributos.

4 CONSIDERAÇÕES FINAIS

Este capítulo apresenta na Seção 4.1 as contribuições, na Seção 4.2 as limitações encontradas no decorrer da pesquisa e as recomendações para trabalhos futuros na Seção 4.3.

4.1 CONTRIBUIÇÕES

Este trabalho analisou a retenção e a evasão no curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, utilizando a tarefa de classificação da Mineração de Dados. Para realizar a pesquisa, foram coletados dados dos estudantes de Sistemas de Informação, utilizando a metodologia baseada no processo do KDD como mostrado nas Seções 3.1 e 3.2.

Foram realizados dois experimentos com o objetivo de prever a retenção dos alunos, dois experimentos para prever a evasão dos alunos e um experimento sobre a aprovação em disciplinas do curso. Para isso, foram avaliados os algoritmos: *Naive Bayes*, *Sequential Minimal Optimization (SMO)*, *IBK*, *J48* e *Random Forest*. A partir disso, o algoritmo *Random Forest* apresentou as melhores médias de acurácia para todos os experimentos realizados.

Os objetivos especificados foram alcançados, que consistiam na construção de modelos que possibilitasse a previsão da situação de retenção e de evasão dos estudantes, este trabalho mostrou a importância da investigação dessa problemática utilizando meios computacionais, mostrando dessa forma resultados satisfatórios.

Além de construir modelos preditivos com uma boa taxa de acurácia, também foi aplicado um método InfoGain Attribute Eval de seleção dos atributos para alcançar uma melhoria na acurácia, que possibilitou que os resultados obtidos nos experimentos fossem aperfeiçoados.

Em resumo, com base nos experimentos realizados é possível chegar em três conclusões:

1. O curso possui um alto índice de retenção após o primeiro ano da graduação. O algoritmo *Random Forest* alcançou uma média de acurácia de 87,76% no 1º experimento com o atributo alvo 'RETIDO' e 73,29% no 2º experimento com o atributo alvo 'CATEGORIA_RETENCAO'.
2. A evasão ocorre predominantemente por jubramento nos 2 primeiros anos de curso. No 3º com o atributo alvo 'EVADIDO' e no 4º experimento com o atributo alvo 'CATEGORIA_EVASAO', o algoritmo *Random Forest* alcançou uma média de acurácia de 87,89% e 83,76%, respectivamente.
3. As aprovações têm o maior índice em todos os períodos, mas apresentam os menores valores nos 2 primeiros anos. Neste 5º experimento com o atributo alvo 'SITUACAO_DISCIPLINA', o algoritmo *Random Forest* obteve uma média de acurácia de 84,23%.

4.2 LIMITAÇÕES DO ESTUDO

Com relação às limitações do estudo, uma das principais dificuldades encontradas foi com relação à coleta dos dados, alguns dados de alunos, tais como informações a respeito do aluno se foi bolsista da instituição ou da relação e da quantidade de empréstimos realizados na biblioteca por período, não foram possíveis de se obter em um formato compatível com o escopo da pesquisa. Outros dados que poderiam agregar para a tarefa de predição mas que não foram possíveis de se obter tais como a nota do Exame Nacional do Ensino Médio (Enem) e dados socioeconômicos dos estudantes.

4.3 RECOMENDAÇÕES

Como recomendações para trabalhos futuros, têm-se a construção dos modelos de classificação em uma ferramenta, para que seja possível a utilização dos modelos preditivos com novas entradas de dados dos alunos. Dessa forma, seria viabilizada a análise dos alunos com possibilidade de retenção ou, até mesmo, de evasão do curso, auxiliando os tomadores de decisão em qual abordagem utilizar para prevenir e tentar minimizar a retenção ou a evasão dos estudantes.

Como exemplo das possíveis aplicações, o modelo do experimento de retenção binário poderia ser aplicado para identificar os alunos que possam se tornar retidos, enquanto o modelo da retenção em classes poderia contribuir para a previsão de quanto tempo um estudante retido concluiria a graduação. Da mesma forma, o modelo do experimento de evasão poderia ser usado para identificação de possíveis alunos que podem se evadir, ou seja, sair do curso sem a concluí-lo.

Outra recomendação seria um estudo com a aplicação da tarefa de regras de associação, porque permite a visualização e identificação de padrões relacionados aos casos de retenção e de evasão, dentre os dados dos alunos.

REFERÊNCIAS

ALGARNI, A. **Data Mining in Education**. *International Journal of Advanced Computer Science and Applications*, v. 7, n. 6, 2016.

ALVES, A. **Análise de novas abordagens para mineração de regras de classificação utilizando algoritmos genéticos**. Universidade Federal de Uberlândia, 20 fev. 2020.

ANDIFES, A.; ABRUEM, A.; SESU/MEC, S. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**: resumo do relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. Avaliação: Revista da Avaliação da Educação Superior, v. 1, n. 2, 1996.

BAKER, R.; ISOTANI, S.; CARVALHO, A. **Mineração de Dados Educacionais**: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03, 31 ago. 2011.

BERRY, M., LINOFF, G. **Mastering Data Mining**: the art and science of customer relationship management. John Wiley & Sons, 2000.

CAMILO, C, SILVA, J. **Mineração de Dados**: Conceitos, Tarefas, Métodos e Ferramentas. p. 29, 2009.

CHICCO, D.; JURMAN, G. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation**. *BMC Genomics*, v. 21, n. 1, p. 1–13, 2020.

CHEN, S. et al. **A novel selective naïve Bayes algorithm**. Knowledge-Based Systems, v. 192, p. 105361, 15 mar. 2020.

COSTA, E. et al. **Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações**. Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1–29, 4 fev. 2013.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, v. 17, 1996.

FARIA, M. **Detecção de Intrusões em Redes de Computadores com Base nos Algoritmos KNN, K-Means++ e J48**. Faculdade Campo Limpo Paulista, 2016.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 3. ed. Morgan Kaufmann, 2011.

HOED, R. **Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação**, 2016.

KAMSU-FOGUEM, B., RIGAL, F., MAUGET, F. **Mining association rules for the quality improvement of the production process**. Expert Systems with Applications, v. 40, p. 1034-1045, 2013.

LEITE, D. R. A.; MORAES, R. M. DE; LOPES, L. W. **Método de Aprendizagem de Máquina para Classificação da intensidade do desvio vocal utilizando Random Forest**. Journal of Health Informatics, v. 12, n. 0, 15 mar. 2021.

LIMA, F. S.; ZAGO, N. **Evasão no Ensino Superior: Desafios conceituais**. p. 7, 2017.

MANHÃES, L. **Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais**, 2015.

OLIVEIRA JÚNIOR, J. **Identificação de padrões para a análise da evasão em cursos 2015 de graduação usando mineração de dados educacionais**. Curitiba: Universidade Tecnológica Federal do Paraná, 2015.

OLIVEIRA JÚNIOR, G. **Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado**. Pernambuco: Universidade Federal de Pernambuco, 2010.

OLIVEIRA, RODRIGO; BARBOSA, JENNY. **Retenção universitária: fatores condicionantes e ações da gestão acadêmica no curso de administração da UFS**, 2016.

SALDANHA, V. **Retenção e Evasão do Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre**. Rio Branco: Universidade Federal do Acre, 2016.

SILVA, D. **Modelo para predição de risco de evasão na educação a distância utilizando técnicas de mineração de dados educacionais**. Niterói: Universidade Federal Fluminense, 2019.

SILVA FILHO, R. L. L. et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641–659, dez. 2007.

ROMERO, C., VENTURA, S. **Data mining in education**. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12–27, 2013.

PATRÍCIO, L. **Apuração do custo das Universidades Federais, e sua relação com os respectivos quantitativos de alunos**. MINISTÉRIO DA EDUCAÇÃO SECRETARIA EXECUTIVA, 2018.

PEREIRA, A. **Retenção discente nos cursos de graduação presencial da UFES**. Vitória: Universidade Federal do Espírito Santo, 2013.

PROPLAN-PRÓ-REITORIA DE PLANEJAMENTO. **Ufac em Números**. Rio Branco: PROPLAN, 2016.

PROPLAN-PRÓ-REITORIA DE PLANEJAMENTO. **Ufac em Números**. Rio Branco: PROPLAN, 2019.

THARWAT, A. Classification assessment methods. **Applied Computing and Informatics**, v. 17, n. 1, p. 168–192, 30 jul. 2020.

TRIBUNAL DE CONTAS DA UNIÃO; SECRETARIA DE EDUCAÇÃO SUPERIOR; SECRETARIA FEDERAL DE CONTROLE INTERNO. **Orientações para o cálculo dos indicadores de gestão**, 2004.

UNIVERSIDADE FEDERAL FLUMINENSE. **Retenção: perfil do aluno retido e suas percepções sobre as políticas existentes na UFF**. Disponível em: <<http://www.uff.br/?q=noticias/29-06-2015/pesquisa-inedita-analisa-causas-da-retenc-ao-de-alunos-da-uff>>. Acesso em: 1 jan. 2021.

VIEIRA, E. et al. **Avaliação da performance do algoritmo J48 para construção de modelos baseados em árvores de decisão**. *Revista Brasileira de Computação Aplicada*, v. 10, n. 2, p. 80–90, 17 jul. 2018.

XIONG, Z. et al. *Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation*. **Computational Materials Science**, v. 171, p. 109203, jan. 2020.

ZEVAREX, V. A. DE O.; SANTOS, C. H. DA S. **Estudos das implicações do teorema de Bayes na computação natural**. Revista Brasileira de Iniciação Científica, v. 7, n. 5, p. 58–79, 19 out. 2020.

WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação**. 2° Ed. Brasil: Elsevier, 2014.

WITTEN, I. H. et al. **Data Mining: Practical Machine Learning Tools and Techniques**. 4ª ed. Morgan Kaufmann, 2016.