CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

EVOLVEDTREE: UM SISTEMA DE MINERAÇÃO DE DADOS EDUCACIONAIS BASEADO EM ÁRVORE DE DECISÃO E ALGORITMO GENÉTICO PARA CLASSIFICAR EVASÃO NO ENSINO SUPERIOR

Gustavo Alexandre Sousa Santos

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Diego Nunes Brandão
Coorientador(a): Luís D. T. Jardim Tarrataca

Rio de Janeiro,
14 de Julho de 2020

EVOLVEDTREE: UM SISTEMA DE MINERAÇÃO DE DADOS EDUCACIONAIS BASEADO EM ÁRVORE DE DECISÃO E ALGORITMO GENÉTICO PARA CLASSIFICAR EVASÃO NO ENSINO SUPERIOR

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

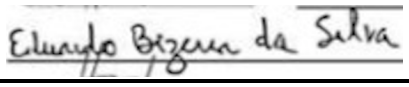
Gustavo Alexandre Sousa Santos

Banca Examinadora:

Presidente, Prof. Diego Nunes Brandão D.Sc. (CEFET/RJ) (Orientador(a))

Prof. Luís Domingues Tomé Jardim Tarrataca, Ph.D. (CEFET/RJ) (Coorientador(a))

Prof. Diego Barreto Haddad D.Sc. (CEFET/RJ)



Prof. Eduardo Bezerra da Silva, D.Sc. (CEFET/RJ)

Prof. Alexandre Plastino de Carvalho, D.Sc. (UFF)

Rio de Janeiro,

14 de Julho de 2020

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S237 Santos, Gustavo Alexandre Sousa
Evolvedtree: um sistema de mineração de dados educacionais baseado em árvore de decisão e algoritmo genético para classificar evasão no ensino superior / Gustavo Alexandre Sousa Santos – 2020.
97f : il. color. + anexos , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2020.
Bibliografia : f. 82-97
Orientador: Diego Nunes Brandão
Coorientador: Luís D. T. Jardim Tarrataca

1. Evasão escolar – Ensino Superior. 2. Big data – Aspectos sociais. 3. Mineração de dados (Computação). 4. Centro Federal de Educação Tecnológica Celso Suckow da Fonseca. I. Brandão, Laura Diego Nunes (Orient.). II. Tarrataca, Luís D. T. Jardim (Coorient.). III. Título.

CDD 371.2913

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

DEDICATÓRIA

Aos meus pais e ao meu irmão, juntos são a prova irrefutável do amor incondicional em minha vida; a vocês dedico todas as minhas vitórias. À minha noiva Georgia, companheira de todos os momentos. Obrigado pela compreensão e carinho.

AGRADECIMENTOS

Primeiramente, a Deus, por ter me dado saúde, força e perseverança para superar as dificuldades.

Aos meus orientadores, ao CEFET/RJ, ao seu corpo docente, à direção e à administração que promoveram e despertaram a capacidade e a confiança necessária para o desenvolvimento deste trabalho.

À UFF e à STI, *staff* e amigos, por toda colaboração no acesso aos dados utilizados nesta pesquisa, assim como no entendimento e compreensão dos processos de negócio da educação superior federal. Além disso, por propiciar e permitir um ambiente de P&D, capaz de estimular e propôr inovação tecnológica.

À minha família, meus pais Ezielia e Josadack, e ao meu irmão Alan, pelo apoio incondicional, incentivo e motivação durante toda a minha vida e, principalmente, nesse período.

À minha noiva, por toda compreensão, companheirismo e parceria no processo de escrita deste trabalho.

Aos meus amigos e colegas do PPCIC, pelas experiências compartilhadas, conversas e risos, assim como a todos os demais que, direta ou indiretamente, fizeram parte desta formação.

A todos vocês, meu muito obrigado!

RESUMO

EvolveDTree: Um sistema de Mineração de Dados Educacionais baseado em Árvore de Decisão e Algoritmo Genético para classificar Evasão no Ensino Superior

A educação é um dos alicerces para o desenvolvimento econômico e social de um país. Garantir que os investimentos em educação sejam feitos de forma eficiente é um grande desafio para toda a sociedade. Neste aspecto, um dos grandes problemas da educação pública de nível superior ocorre quando os estudantes se desassociam da instituição sem completar o curso no qual estavam matriculados, caracterizando o fenômeno de evasão. Assim, os recursos investidos na formação desses estudantes acabam sendo perdidos, representando um desperdício financeiro significativo. Neste contexto, o desenvolvimento de ferramentas que auxiliem no processo de minimização dos casos de evasão torna-se imprescindível. O presente trabalho propõe o desenvolvimento de um sistema que permite avaliar diferentes técnicas de mineração de dados para classificar a tendência de um aluno abandonar ou graduar no curso em que está matriculado. Por meio desse sistema, busca-se a identificação de características que indiquem a evasão antes que ela ocorra, permitindo que alguma ação possa ser tomada de maneira a minimizá-la. Para este objetivo, foi desenvolvido um Data Warehouse Educacional (EDW) que permite a integração dos dados educacionais de uma instituição de ensino superior. Os resultados obtidos demonstram que o EDW desenvolvido é robusto o suficiente para permitir que diversas análises sejam realizadas pela gestão acadêmica. Os modelos de classificação avaliados foram comparados por meio de diferentes métricas, destacando-se a estratégia baseada em árvores de decisão. Uma técnica de redução de dimensionalidade baseada em algoritmo genético também foi avaliada, permitindo uma diminuição do tempo de processamento da fase de treinamento em todos os modelos de classificação avaliados. Contudo, foi identificado um aumento no tempo total da abordagem proposta, quando avaliadas as fases de pré-processamento e treinamento, simultaneamente.

Palavras-chave: Ensino Superior. Evasão. Data Warehouse Educacional. Algoritmo Genético. Árvore de Decisão.

ABSTRACT

EvolveDTree: A system of the Educational Data Mining based on Decision Tree and Genetic Algorithm to classify Dropout in Higher Education

Education is one of the foundations for the economic and social development of a country. Ensuring that investments in education are made efficiently is a significant challenge for the whole of society. In this regard, one of the major problems of public higher education occurs when students disassociate themselves from the institution without completing the course in which they were enrolled, a phenomenon known as dropout. As a result, the resources invested in the training of those students end up being lost, representing a significant financial waste. The development of tools that assist in the process of minimizing dropout cases is therefore essential. The present work proposes the development of a system that allows the evaluation of different data mining techniques to classify a student's tendency to drop out or graduate from the course in which he is enrolled. The system seeks to identify characteristics that indicate dropout before it occurs, allowing some action to be taken to minimize it. For this purpose, an Educational Data Warehouse (EDW) was developed that enables the integration of educational data from a higher education institution. The results obtained demonstrate that the developed EDW is robust enough to allow several analyzes to be carried out by academic management. Different classification models were evaluated using different metrics. Of these, the strategy based on decision trees showed the most promise. A dimensionality reduction technique based on a genetic algorithm was also evaluated. This strategy allowed for a reduction in the processing time of the training phase in all the classification assessed models. However, an increase in the total time of the proposed approach was identified when the preprocessing and training phases were measured simultaneously.

Keywords: Higher Education. Dropout. Machine Learning. Genetic Algorithm. Decision tree.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AB	*AdaBoost*
AD	Árvore De Decisão
AG	Algoritmo Genético
AM	Aprendizado De Máquina
CART	*Classification and Regression Trees*
CHAID	*Chi-square Automatic Interaction Detection*
EDM	*Educational Data Mining*
EDW	*Educational Data Warehouse*
ETL	*Extract, Transform and Load*
ID3	*Iterative Dichotomiser 3*
IES	Instituição De Ensino Superior
INEP	Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira
KNN	*K-Nearest Neighbours*
MCC	Coeficiente De Correlação De Matthews
MD	Mineração De Dados
MEC	Ministério De Educação
MLP	*Multilayer Perceptron*
NB	*Naive Bayes*
RB	Redes Bayesianas
RF	*Random Forest*
RL	Regressão Logística
RN	Redes Neurais
SEMESP	Sindicato Das Entidades Mantenedoras De Ensino Superior De São Paulo
SVM	Máquina De Vetor De Suporte
UFF	Universidade Federal Fluminense
UNESCO	Organização Das Nações Unidas Para A Educação, A Ciência E A Cultura

SUMÁRIO

1- Introdução

Os investimentos corretos em educação são fundamentais para garantir o desenvolvimento econômico e social de um país. No contexto da educação de nível superior, um problema significativo ocorre quando os estudantes se desassociam das instituições sem concluir o curso no qual estavam matriculados, caracterizando o fenômeno de evasão.

No ano de 2018, o Censo da Educação Superior mostrou que foram ofertadas mais de 13,5 milhões de vagas em cursos de graduação, sendo mais de 3,6 milhões remanescentes de evasão. Entretanto, apenas 11,3% dessas vagas remanescentes foram preenchidas [BRASIL, 2019]. As vagas, originalmente ocupadas por alunos evadidos, tornam-se ociosas e demandam algum tipo de remanejamento que, possivelmente, exigirá mais recursos [Barros, 2015].

O governo brasileiro tem estimulado a educação superior por meio de diversos mecanismos que facilitem o acesso a uma Instituição de Ensino Superior (IES). Porém, a quantidade de investimento alocado em função da evasão situa-se em patamares superiores aos desejados. De acordo com a Organização para Cooperação e Desenvolvimento Econômico (OCDE), o custo médio anual por estudante de graduação na educação pública brasileira no ano de 2013 foi de US$ 13.539,90, implicando uma perda financeira significativa com a ocorrência da evasão [OECD, 2016]. Por esta razão, as IES têm tentado ativamente compreender as causas deste fenômeno [Baggi and Lopes, 2011].

Ao buscar soluções para o problema da evasão, algumas IES têm adotado soluções tecnológicas baseadas em sistemas de apoio à tomada de decisão. As áreas de conhecimento que abrangem estes sistemas no contexto educacional são: Mineração de Dados Educacionais [1] (EDM, do inglês *Educational Data Mining*) e *Data Warehousing* na Educação Superior [2] (HEDW, do inglês *Higher Education Data Warehousing*). A presente dissertação se enquadra dentro do escopo de tais áreas.

[1] http://educationaldatamining.org/
[2] https://hedw.org/

1.1- O Problema

A evasão é uma adversidade que atinge várias IES no mundo, tanto públicas como privadas [BRASIL, 2017, 2018; Nascimento and Verhine, 2017]. A quantidade de investimento despendida devido ao abandono representa um problema de alocação de recursos [Silva Filho et al., 2007]. A identificação da origem de tal problema tem sido objeto de estudo de pesquisadores da educação e de outras áreas de conhecimento, inclusive ligadas à tecnologia [DeBerard et al., 2004; Silva Filho et al., 2007; Baker, 2010; Rodriguez, 2011; Baker et al., 2011; Romero and Ventura, 2013].

As IES têm políticas e programas que visam auxiliar, estimular e fomentar o plano de ensino e o processo de aprendizagem na comunidade acadêmica [Seiffert and Hage, 2008; de Assis et al., 2013]. No entanto, mesmo diante dessas iniciativas de suporte à melhor formação dos alunos, os índices de desempenho ainda se revelam pouco satisfatórios, com turmas grandes no início do curso e muito reduzidas ao final, principalmente, devido à evasão e transferências de alunos [BRASIL, 2018]. Segundo o Mapa do Ensino Superior do Brasil [3] de 2015, produzido pelo Sindicato das Entidades Mantenedoras de Ensino Superior de São Paulo (SEMESP), no ano de 2013, aproximadamente 25% dos alunos de IES evadiram [Dalongaro et al., 2016].

Assim, torna-se compreensível questionar, dado o aluno de uma IES e seu custo médio, quais ações devem ser aplicadas para reduzir as chances de abandono deste estudante e quando elas devem ser executadas. As razões para um estudante decidir seguir ou não um curso em uma determinada instituição são diversas e variam desde os níveis pessoal, social e até institucional [Baggi and Lopes, 2011].

Como pode ser observado, a evasão é uma questão complexa, resultante de uma conjunção de fatores que pesam na decisão de o aluno permanecer ou não no curso. Alguns desses fatores são [Tigrinho, 2008; Baggi and Lopes, 2011]:

- Repetência;
- Orientação vocacional;
- Mudança de curso;

[3] https://www.semesp.org.br/pesquisas/mapa-do-ensino-superior/

- Pouco prestígio da formação profissional;
- Incompatibilidade com o horário de trabalho;
- Entre outros.

Em função disso, é razoável avaliar se, após a reprovação em uma ou mais disciplinas, os alunos estão mais propensos a abandonarem seus cursos. Adicionalmente, quais as informações sobre a carreira profissional pertinentes ao curso ingressado podem influenciar de alguma forma na permanência desse aluno? Um exemplo desse questionamento é o estudo da Pró-Reitoria de Ensino e Graduação desenvolvido pela PROEG [2016] na UFRR (Universidade Federal de Roraima), destacando que cursos como medicina e direito têm altas taxas de sucesso. Desse modo, torna-se questionável, também, analisar se uma expectativa de vida melhor e a realização profissional do aluno são fatores suficientes para motivá-lo a concluir o curso.

1.2- Motivação

De acordo com o que foi apresentado, a evasão gera prejuízos financeiros significativos, pois, ao evadirem, as vagas ocupadas anteriormente por esses alunos tornam-se ociosas e, dificilmente, serão preenchidas. Dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) mostram que apenas 36,3% das vagas ociosas oferecidas em 2017 foram efetivamente ocupadas [BRASIL, 2018]. Assim, a busca por alternativas que minimizem a evasão é de suma importância para a sociedade e inúmeros pesquisadores vêm abordando o assunto, conforme apresentado a seguir.

O trabalho de Burgos et al. [2017] utiliza Regressão Logística (RL) para analisar o desempenho dos alunos de vários cursos, combinando uma abordagem de predição e um plano de ação, a fim de reduzir o abandono nos anos de 2014 e 2015.

Ahuja and Kankane [2017] utilizaram um conjunto de algoritmos para prever a probabilidade de um aluno concluir o curso de graduação. Segundo os autores, os resultados dos testes efetuados destacam o desempenho da AD em prever a insatisfação dos alunos com base nos dados obtidos em uma pesquisa de satisfação sobre cursos de graduação.

No trabalho de Manhães et al. [2014], é apresentada uma arquitetura que visa prever a evasão, denominada *WAVE*. Nessa proposta, são analisados apenas dados do sistema de gestão acadêmica, desconsiderando informações sociodemográficas e qualquer integração de dados com outros sistemas.

Também no intuito de redução da evasão, a pesquisa de Oliveira Júnior et al. [2017] apresenta uma solução com o objetivo de identificar padrões mediante uma abordagem de engenharia de atributos para criação e seleção de atributos.

Tais estudos propuseram análises de dados educacionais em formas distintas para o problema da evasão no contexto de diferentes instituições. Dentre esses trabalhos, não foi percebido que tenha sido desenvolvido um sistema de informação que permitisse a utilização das análises produzidas por especialistas e demais interessados.

Sendo assim, esta pequisa propõe um processo analítico composto pelo desenvolvimento de uma base de dados integrada por várias áreas de negócio e um sistema de apoio à tomada de decisão organizado em análises descritiva e preditiva. Além disso, propõe como estudo de caso aplicar o sistema desenvolvido no contexto da Universidade Federal Fluminense (UFF). A escolha da UFF como estudo de caso deve-se a importância nacional da instituição, que segundo dados do Censo de 2017[4], foi a responsável por $4,2\%$ de todas as vagas ofertadas no ensino superior federal. Atualmente a UFF é a responsável pelo maior número de matrículas na graduação, com cerca de 47.254 alunos.

1.3- Objetivos

Esta dissertação desenvolve um sistema de apoio à tomada de decisão para instituições de ensino superior, focando principalmente em minimizar a evasão, mediante dois componentes: a) o desenvolvimento de um *Data Warehouse* Educacional e b) um sistema de apoio à Tomada de Decisão baseado em modelo de predição. Esse modelo é implementado com uma abordagem de redução de dimensionalidade, combinando uma avaliação de técnicas de classificação em Aprendizado de Máquina (AM) para prever alunos em risco de evasão. Dessa forma, este trabalho busca alcançar seus objetivos de acordo com as seguintes atividades:

[4]http://portal.inep.gov.br/censo-da-educacao-superior

- Criar uma base de dados analítica sobre o tema evasão;
- Produzir análises descritivas com a base de dados desenvolvida;
- Avaliar uma abordagem para redução de dimensionalidade utilizando AG;
- Produzir um modelo de predição, com base em estudo comparativo entre AD e outras técnicas de aprendizado de máquina, a fim de prever alunos em risco de evasão.

Com o intuito de justificar a escolha do algoritmo de AD – o método *Classification and Regression Trees* (CART), é válido mencionar algumas de suas vantagens perante outras técnicas de AD: a) construção de árvores binárias [Breiman et al., 1984], favorecendo as tarefas de classificação binária, principal característica do conjunto de dados utilizado nesta dissertação; b) capacidade de detecção de *outliers* [Singh and Gupta, 2014; Timofeev, 2004]; e c) praticidade para gerar árvores de classificação com dados que contenham valor categórico [Singh and Gupta, 2014; Sharma and Kumar, 2016; Breiman, 2017].

1.4- Estrutura do documento

Esta dissertação está estruturada em mais sete capítulos. O Capítulo 2 descreve os conceitos sobre o problema da evasão e algumas informações sobre o ensino superior brasileiro. O Capítulo 3 destaca um conjunto de trabalhos relacionados ao problema apresentado no escopo de abordagens computacionais, especificamente à área de Mineração de Dados Educacionais. Em seguida, o Capítulo 4 detalha em etapas as atividades de integração de dados para a criação da base de dados analítica, avaliada pelo sistema de predição descrito no Capítulo 5, sendo este composto de AD e AG. O Capítulo 6 enfatiza os cenários de avaliação e os resultados de treinamento e teste dos algoritmos utilizados. O Capítulo 7 elucida os resultados alcançados na etapa de validação pelo modelo algorítmico de melhor desempenho durante as fases de treinamento e teste. Por fim, o Capítulo 8 discute as considerações finais, os objetivos atingidos e os trabalhos futuros desta pesquisa.

2- A Evasão no Ensino Superior

A evasão é um tema que atinge diversas instituições de ensino, conforme já mencionado, compreender sua origem e como isso afeta o sistema educacional tem sido motivo de estudo para pesquisadores de diversas áreas [Delavari et al., 2005; Silva Filho et al., 2007; Baker and Yacef, 2009; Romero and Ventura, 2013]. Este capítulo está dividido em duas partes, a Seção 2.1 contém as definições e conceitos que serão necessários para uma compreensão do problema da evasão e a Seção 2.2 retrata brevemente um cenário de dados históricos sobre a evasão no ensino superior brasileiro na última década.

2.1- Definições e Conceito

O conceito de evasão no ensino superior adotado pelo Ministério de Educação (MEC) é "a saída definitiva de um aluno matriculado no referido curso sem concluí-lo ou a diferença entre alunos ingressantes e concluintes, após uma geração completa" [ANDIFES, 1996]. De acordo com o Comitê Especial de Estudos de Evasão do MEC, o conceito de evasão pode ser caracterizado também pelo tipo de abandono [ANDIFES, 1996]:

- **Abandono de curso** – o aluno desliga-se do curso em situações como transferência de curso, reprovação em disciplina ou exclusão por norma institucional;
- **Abandono da instituição** – o aluno desiste da instituição em que está matriculado;
- **Abandono do sistema** – o aluno abandona permanentemente ou temporariamente o ensino superior.

Além disso, a evasão pode também ser medida com base nos critérios de: IES, curso, área de conhecimento e, até mesmo, período de oferta de cursos [Silva Filho et al., 2007]. Em princípio, pode-se estudar a evasão no âmbito de uma IES ou de um sistema de ensino, o que representa um conjunto de instituições. Entretanto, a evasão no

ensino superior é um fenômeno complexo e não deve ser avaliada fora de uma análise histórica mais ampla, pois para um aluno a decisão de evadir é reflexo de uma realidade de períodos anteriores [Baggi and Lopes, 2011]. Neste trabalho, todo o processo de identificação e análise da evasão é baseado no conceito de abandono de curso.

2.2- Uma Visão sobre o Ensino Superior Brasileiro na última década

Nos estudos apresentados [Silva Filho et al., 2007; Dias et al., 2010; Amaral, 2013; Matta et al., 2017], é possível verificar que a resposta para a problemática da evasão é, frequentemente, uma simplificação para um dilema que envolve as questões de ordem acadêmica, as expectativas do aluno em relação à sua formação e a própria integração do estudante com a IES.

No intuito de elucidar um pouco mais a evasão, a Tabela 1 apresenta uma visão sobre a última década, baseada em dados do Censo da Educação Superior [1] quanto aos alunos ingressantes, matriculados e concluintes dos cursos de graduação das IES públicas.

Tabela 1 – Um panorama estatístico do Ensino Superior Público

Ano	Ingressos (A)	Matrículas (B)	Concluintes (C)	Índice C/A	Índice C/B
2008	352.615	1.273.965	187.758	0,53	0,14
2009	379.134	1.351.168	187.804	0,49	0,13
2010	435.710	1.461.696	178.407	0,40	0,12
2011	456.635	1.595.391	194.666	0,42	0,12
2012	499.370	1.715.752	202.394	0,40	0,11
2013	494.940	1.777.974	206.261	0,41	0,11
2014	504.627	1.821.629	225.714	0,44	0,12
2015	504.038	1.823.752	224.196	0,44	0,12
2016	505.002	1.867.477	231.752	0,45	0,12
2017	502.621	1.879.784	238.061	0,47	0,12
2018	580.936	1.902.972	259.302	0,44	0,13

Fonte: Resumos Técnicos do Censo da Educação Superior

[1] http://portal.inep.gov.br/censo-da-educacao-superior

Mediante aos dados apresentados nessa tabela, é possível perceber que mesmo com um aumento no número de alunos ingressantes e no número de alunos matriculados em mais de 50%, o número de concluintes não representa um aumento proporcional aos alunos ingressantes e matriculados. Inclusive, é possível evidenciar também que os percentuais de concluintes por ingressantes e concluintes por matriculados têm-se reduzido ao longo do tempo, respectivamente apresentados pelas colunas de Índice C/A e Índice C/B. Assim, observa-se que a evasão é um fenômeno grave que ocorre tanto nas instituições públicas quanto nas privadas em todo Brasil Tigrinho [2008].

A Figura 1 apresenta a distribuição de matrículas nas IES públicas e privadas, descrevendo um aumento de 44,6% no número de matrículas de 2008 até 2018. As IES privadas têm uma participação de 75,4% (6.373.274) no total de matrículas de graduação e as IES públicas representam 24,6% (2.077.481). No gráfico exibido, quando se compara de 2008 até 2018, observa-se um aumento no número de matrículas de 49,8% na rede privada e de 33,8% na rede pública. Este aumento torna evidente que medidas para o acompanhamento dos alunos são fundamentais para reduzir os números da evasão.

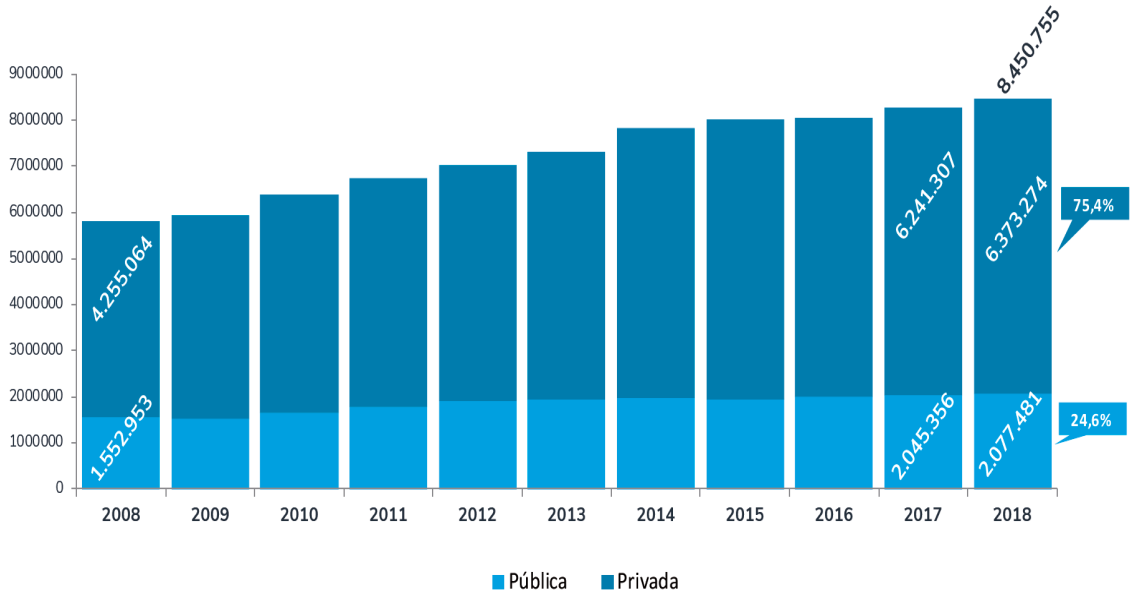


Figura 1 – Matrículas em cursos de graduação dentre os anos 2008 até 2018. Fonte: Notas Estatísticas do Censo da Educação Superior 2018

No contexto brasileiro, várias pesquisas têm sido desenvolvidas sobre o tema em diversas perspectivas, as quais destacam-se os trabalhos de Paredes [1994], Andriola [2003], Gaioso [2005] e Speller et al. [2012].

O estudo apresentado por Paredes [1994] refere-se às desistências nos cursos. Segundo Paredes, a desistência de cursos é subestimada, no que se refere ao rendimento dos cursos de cada instituição e, superestimada quando é vista como abandono definitivo da formação. O autor afirma que 64% dos alunos que desistiram do curso de origem, obtiveram a titulação em outra instituição, uma vez que o sistema permite a mobilidade dos alunos entre as IES.

O trabalho de Andriola [2003] corrobora com a análise apresentada por Paredes, destacando que a mudança de curso nas universidades brasileiras é algo alarmante. Isto não só sinaliza os equívocos na identificação de vocação, mas também representa um ônus para a sociedade por causa da ocupação indevida de vagas, acarretando desperdício financeiro para sociedade no caso das IES públicas.

Na pesquisa desenvolvida por Gaioso [2005], foram entrevistados alunos e gestores de várias IES nas regiões Sudeste, Centro-Oeste, Sul e Nordeste do Brasil. Nessa pesquisa, o autor menciona inconsistências entre os números divulgados oficialmente sobre evasão e os números apresentados pelos gestores durante as entrevistas. Segundo Gaioso, muitos entrevistados não puderam mencionar números sobre a evasão, pois algumas instituições a enxergam como “tabu”, o que destaca a necessidade de mais estudos para qualificar melhor esse tipo de resultado.

A Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) publicou em 2012 um trabalho sobre os desafios e perspectivas da educação superior brasileira [Speller et al., 2012]. Segundo a pesquisa, é necessário dedicar mais atenção aos temas “evasão” e “vagas ociosas”, pois são tópicos bastante significativos para o desenvolvimento da educação superior no país. Desse modo, os fatores que têm sido apresentados como motivadores à ociosidade das vagas e à evasão, tais como a insuficiência de recursos financeiros e a recente diversificação e qualidade do sistema, ainda são pouco compreendidos.

No próximo capítulo apresentaremos mais detalhadamente os trabalhos relacionados ao tema de pesquisa desta dissertação com enfoque em abordagens computacionais.

3- Trabalhos Relacionados

Este capítulo apresenta uma visão concisa sobre os trabalhos científicos relacionados à questão da evasão, focando, principalmente, naqueles que apresentam contribuições utilizando técnicas computacionais. Sendo assim, o capítulo divide-se em duas partes: Seção 3.1, que traz uma exposição à área de conhecimento da *Educational Data Mining* (EDM), e Seção 3.2, que aborda as pesquisas mais recentes desenvolvidas no Brasil e no mundo, com enfoque no problema da evasão em relação à EDM.

3.1- *Educational Data Mining*

Nos últimos anos estudos mostram que as instituições de ensino adquiriram uma grande quantidade de dados sobre os alunos [Chen et al., 2012; Picciano, 2012; Inmon and Linstedt, 2014; Sin and Muthu, 2015; Luna et al., 2016]. Tais dados foram produzidos pelos sistemas de informação que suportam gerencialmente os processos, procedimentos e tarefas do dia a dia das IES. Com base nesses dados, é possível iniciar atividades para promover descoberta de conhecimento sobre os diversos problemas da instituição, em particular, sobre a evasão.

Para suportar as atividades de descoberta de conhecimento em dados educacionais, existe a *Educational Data Mining* (EDM), uma área da ciência da computação que tem se revelado promissora na verificação desses dados. As pesquisas em EDM são feitas principalmente para analisar a aprendizagem, compreender e prever a evasão e, consequentemente, estimular a permanência de alunos na instituição [Baker and Yacef, 2009; Romero and Ventura, 2010; Costa et al., 2013].

Sobre EDM, Baker and Yacef [2009] destacam o conceito da IEDMS [1]: “uma disciplina emergente, dedicada ao desenvolvimento e aplicação de métodos para explorar dados educacionais e compreender melhor os alunos e seus padrões”.

De acordo com Baker [2010], os trabalhos em EDM podem ser categorizados

[1] http://educationaldatamining.org

da seguinte forma: predição, agrupamento, construção de regras, avaliação de regras de associação e descoberta de conhecimento através de modelos. As três primeiras categorias são predominantes e tradicionais na pesquisa de mineração de dados. As duas últimas consistem, respectivamente, em visualização de dados com análise estatística e modelos descritivos. Neste contexto, a pesquisa proposta nesta dissertação pode ser categorizada entre duas: (1) predição e (2) descoberta de conhecimento através de modelos.

Um estudo apresentado por Santos et al. [2018] destaca um crescimento no desenvolvimento de pesquisas sobre EDM nos últimos anos. Para este trabalho, realizou-se uma análise na base de dados *Scopus*, usando a expressão de busca "educacional data mining", no intuito de observar o quanto a EDM tem se destacado nos últimos anos. Na Figura 2, é possível observar um crescimento acentuado a partir do ano de 2015. No ano seguinte, esse crescimento foi superado em dobro, reforçando a notoriedade que a área tem alcançado recentemente.

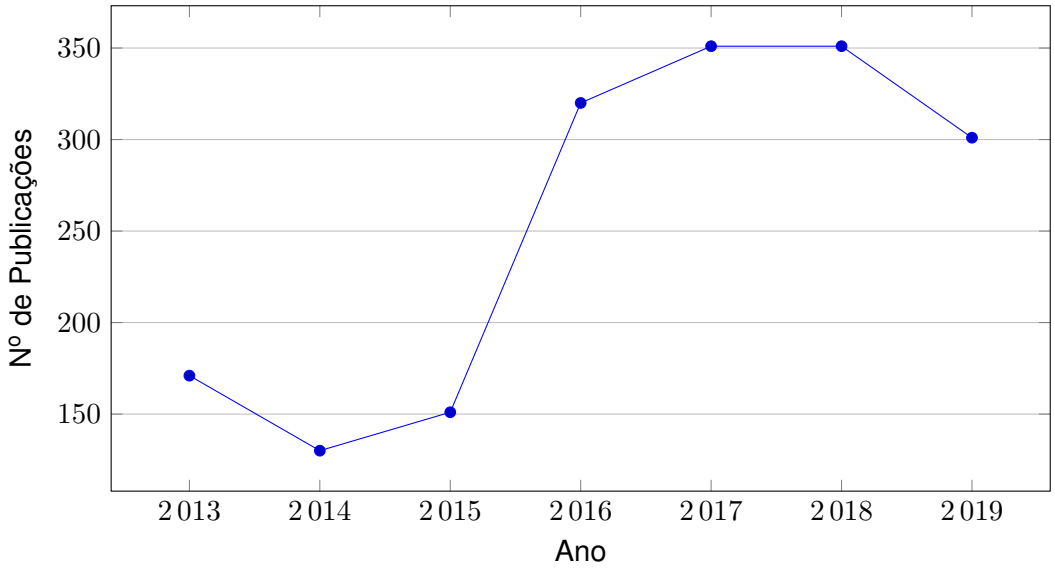


Figura 2 – Nº de publicações em EDM na base Scopus nos anos de 2013 a 2019

A Figura 3 identifica os autores que mais publicaram trabalhos em EDM. Em seguida, na Figura 4, são apresentados os tipos de publicações baseadas no catálogo *Scopus*. A Figura 5 mostra um quantitativo do desenvolvimento das pesquisas em EDM, de acordo com o país de origem entre 2013 e Nov/2019.

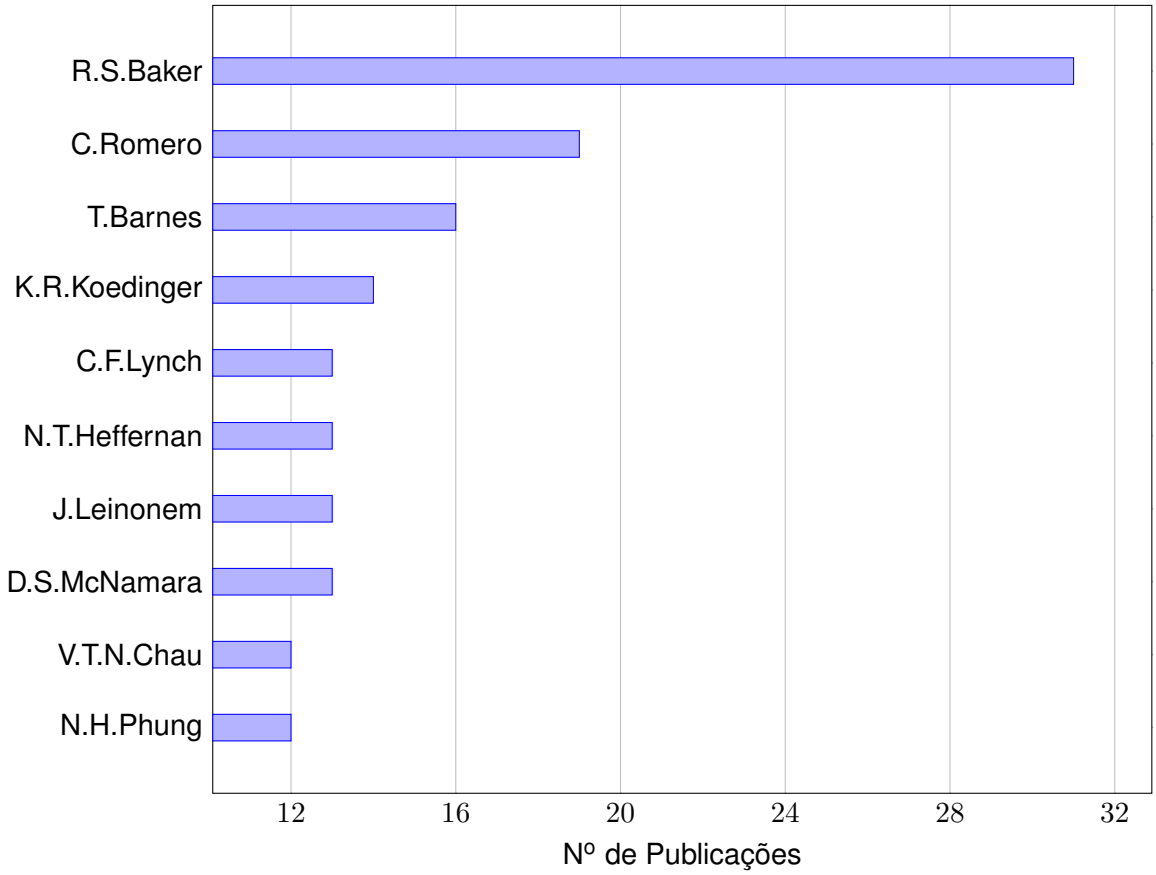


Figura 3 – Número de publicações por autor em EDM na base Scopus nos anos de 2013 a 2019

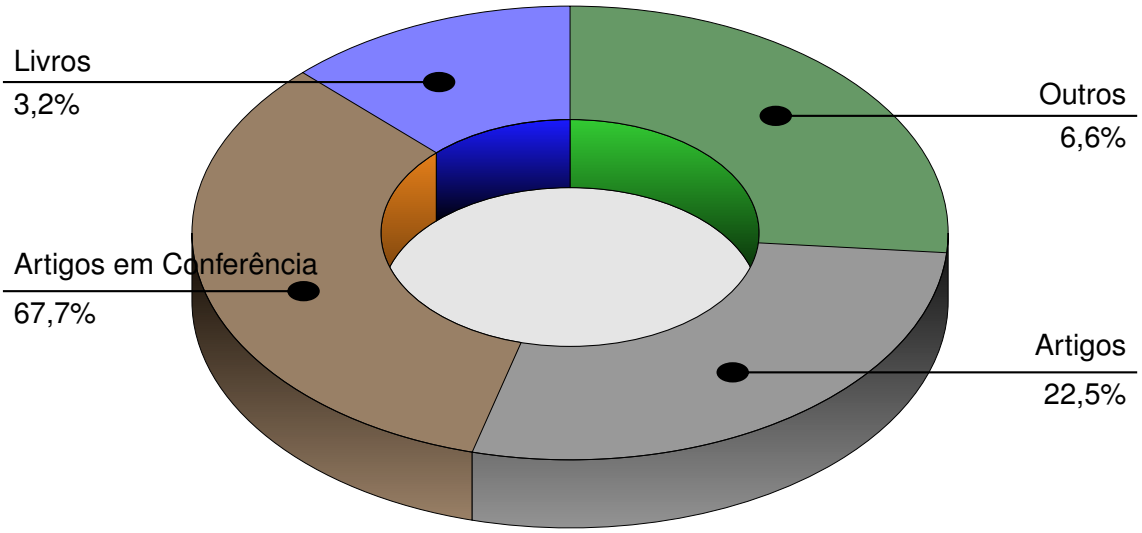


Figura 4 – Percentual por tipo de publicação em EDM na base Scopus nos anos de 2013 a 2019

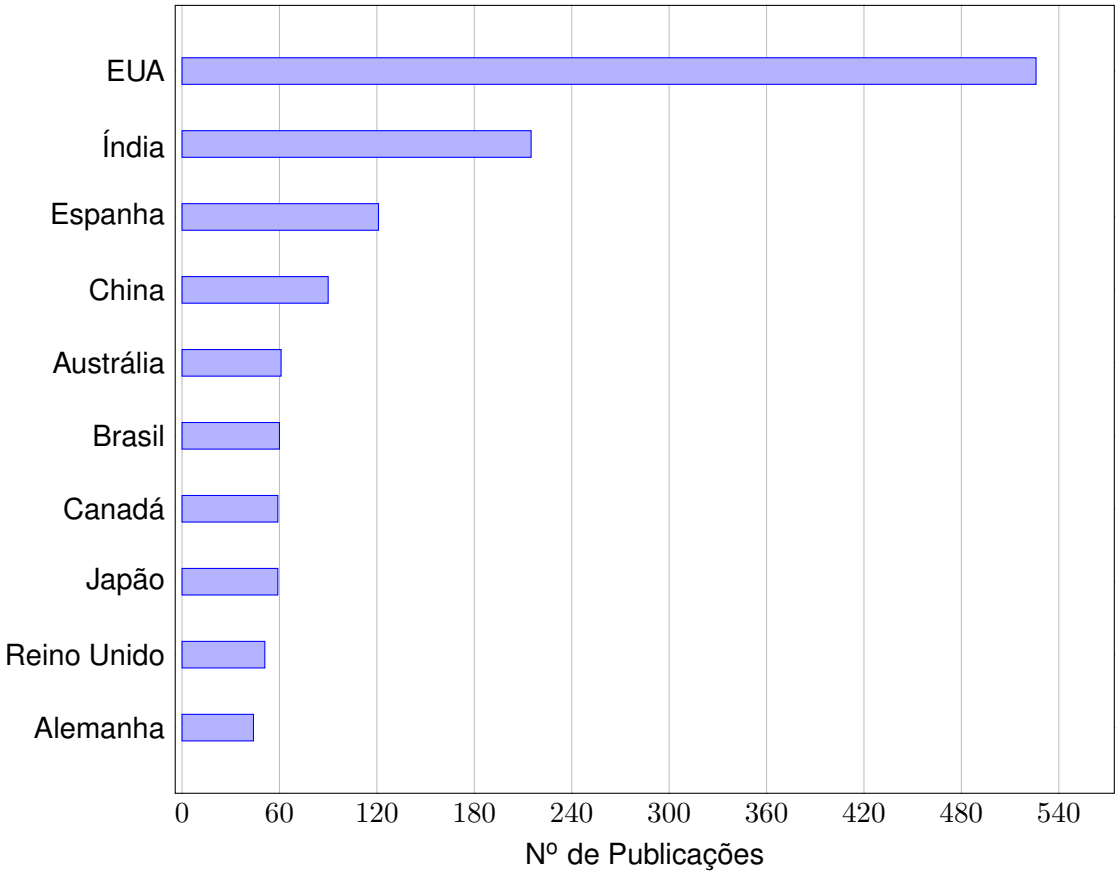


Figura 5 – Número de publicações por país em EDM na base Scopus nos anos de 2013 a 2019

Ainda visando enfatizar a produção científica desta área de pesquisa, a Figura 5 destaca os países através de um ranqueamento conforme a quantidade de publicações em EDM. Com base neste *ranking*, os três países melhor posicionados são, respectivamente: EUA, Índia e Espanha. O Brasil aparece em sexto lugar com um número de publicações muito próximo ao de países como Austrália (5º lugar), Canadá (7º lugar) e Japão (8º lugar), destacando-se à frente de países como Reino Unido (9º lugar) e Alemanha (10º lugar).

Por fim, os trabalhos de Maschio et al. [2018] e Costa et al. [2013] apresentam um panorama atual sobre as pesquisas desenvolvidas, os caminhos percorridos e a percorrer no Brasil, explorando a área de EDM à resolução dos desafios presentes na educação brasileira.

3.2- Revisão Sistemática

Está seção descreve os trabalhos relacionados a área de EDM com foco na predição de evasão, detalhando o processo aplicado e os procedimentos envolvidos para a escolha de apresentá-los aqui. A metodologia utilizada para esta revisão sistemática foi originada dos trabalhos apresentados por Kitchenham et al. [2009] e Dutt et al. [2017]. A questão de pesquisa desta revisão da literatura é "**identificar e conhecer as pesquisas em EDM na predição de evasão em alunos do ensino superior presencial do Brasil e do mundo**".

Segundo Baker and Yacef [2009], os estudos de predição em EDM têm como objetivo desenvolver um modelo que possa inferir um único aspecto dos dados a partir de alguma combinação de outros elementos dos dados. Predição é um método que precisa identificar a variável de saída para um conjunto de dados, em uma característica (rótulo) capaz de representar informações confiáveis sobre a variável alvo. De acordo com Baker [2010], geralmente os tipos mais comuns de abordagens de predição em EDM são classificação, regressão e estimativa de densidade.

Para realizar o levantamento bibliográfico, a primeira expressão de busca foi *"Educational Data Mining AND Dropout"*, resultando em 30900 artigos no *Google Scholar*, 966 artigos no *Scopus* e 68 artigos no *IEEEXplore*. A partir desta primeira expressão de busca, outras expressões foram testadas e validadas, a fim de obter o grupo mais apropriado de publicações que representasse a melhor similaridade com o objeto de pesquisa deste trabalho: *"Construir um Modelo de Predição capaz de Identificar Alunos do Ensino Superior Presencial com Risco de Evasão"*.

Dito isto, expressões como "MOOC" (do inglês, *Massive Open Online Courses*[2]) foram necessárias para evitar que trabalhos relacionados à análise da aprendizagem em cursos online fossem trazidos, pois, apesar da relevância atual para o tema evasão, compreende-se que o aluno pertencente a esta modalidade de ensino possui características e comportamentos que diferem do aluno presencial. Consequentemente, expressões como "Online" e "Online Course" foram também adicionadas à expressão de busca.

[2]https://www.mooc.org

A expressão “High School” foi também inserida para evitar que estudos relacionados ao ensino médio fossem selecionados. Da mesma forma, foi necessário incluir a expressão “Learning Analytics” para evitar que os trabalhos relacionados a essa subárea de pesquisa da EDM fossem acrescentados nesta revisão da literatura. Sendo assim, a expressão de busca foi aprimorada em etapas, de acordo com a Tabela 2, e executada de maneira gradativa, na base *Scopus*, conforme a seguir:

1. *(“Educational Data Mining” AND “Dropout” AND “Predict”) AND PUBYEAR > 2012 AND PUBYEAR < 2020*;
2. *(“Educational Data Mining” AND “Dropout” AND “Predict”) AND NOT TITLE-ABS-KEY("MOOC” OR "Online”) AND PUBYEAR > 2012 AND PUBYEAR < 2020*;
3. *(“Educational Data Mining” AND “Dropout” AND “Predict”) AND NOT TITLE-ABS-KEY(“MOOC” OR “Online” OR “Online Course” OR “High School”) AND PUBYEAR > 2012 AND PUBYEAR < 2020*;
4. *(“Educational Data Mining” AND “Dropout” AND “Predict”) AND NOT TITLE-ABS-KEY(“MOOC” OR “Online” OR “Online Course” OR “High School” OR “Learning” OR “Learning Analytics”) AND PUBYEAR > 2012 AND PUBYEAR < 2020*.

A Tabela 2, baseada nas expressões de busca realizadas, apresenta o número de artigos retornados em cada expressão executada nos referidos catálogos. Esta pesquisa visa obter boas orientações e direcionamentos para desenvolver novos estudos em EDM. Desse modo, publicações sobre o estado da arte, previsão de abandono, análise de evasão, desafios e perspectivas em EDM são muito bem recebidas.

Tabela 2 – Resultados das Expressões de Busca

Expressão de Busca	**Catálogo de Periódico**		
	Google Scholar	***Scopus***	***IEEEXplore***
1)	2180	439	25
2)	1860	276	12
3)	1460	261	4
4)	615	85	0

A partir da Expressão de Busca 3, foram realizados testes para validar o processo de pesquisa e, principalmente, tratar termos correlatos e sinônimos. Dessa maneira, foi obtido um resultado mais preciso na expressão de busca 4, cujo resultado está apresentado na Tabela 3.

Neste estudo, os critérios de inclusão e exclusão dos artigos retornados na Expressão de Busca 4 foram aplicados por meio do título, resumo e palavras-chave.

A finalidade desse procedimento é remover os trabalhos que, apesar de selecionados, não estão aderentes ao objeto de pesquisa e, sucessivamente, incluir trabalhos não selecionados que foram adicionados. Os critérios de inclusão são descritos a seguir:

- Publicações encontradas na *International Conference on Educational Data Mining* e no periódico *Journal of Educational Data Mining*;
- Publicações de autoria ou coautoria na área de EDM com os seguintes autores: Ryan Shaun Baker [3] (R.S.Baker) e Cristobal Veloso Morales [4] (C. Romero);
- Publicações desenvolvidas no Brasil e no mundo no contexto da EDM com foco na 'predição de evasão';
- Publicações desenvolvidas no Brasil e no mundo no contexto da EDM com foco na revisão de literatura e estudos comparativos.

Agora, serão descritos os critérios de exclusão:

- Publicações desenvolvidas no Brasil e no mundo no contexto de Educação Online: *Massive Open Online Courses* (MOOC) ou Ambiente Virtual de Aprendizagem (*i.e. Moodle*);
- Publicações desenvolvidas fora do contexto do Ensino Superior (Ensino Básico, Ensino Médio ou Ensino Técnico);
- Publicações desenvolvidas no contexto da EDM com foco em Análise de Aprendizagem (*Learning Analytics*);
- Publicações que, mesmo retornadas da EBF, não estiverem escritas em inglês ou português.

[3] http://www.upenn.edu/learninganalytics/ryanbaker/
[4] http://www.uco.es/users/in1romoc/

Sendo assim, uma amostra dos artigos selecionados é apresentada na Tabela 3, mas a listagem completa está disponibilizada aqui [5]. Respectivamente, as publicações foram organizadas nesta tabela, de acordo com autoria, ano de publicação, país de origem, objetivo de trabalho e número de citações.

Tabela 3 – Publicações selecionadas sobre EDM

Autor	Ano	País	Objetivo	Cit.
Romero and Ventura	2013	Espanha	Revisão da Literatura	367
Costa et al.	2013	Brasil	Revisão da Literatura	30
Papamitsiou and Economides	2014	Greece	Revisão da Literatura	205
Doshi and Chaturvedi	2014	Índia	Seleção de Atributos	0
Tekin	2014	Turquia	Predição de Evasão	20
Manhães et al.	2014	Brasil	Predição de Evasão	19
Manhães et al.	2014	Brasil	Predição de Evasão	0
Sin and Muthu	2015	Índia	Big Data	62
Thakar and Mehta	2015	Índia	Predição de Evasão	30
Guarín et al.	2015	Colombia	Predição de Desempenho	27
Aziz et al.	2015	Malásia	Predição de Desempenho	6
Barbosa Manhães et al.	2015	Brasil	Predição de Desempenho	5
Ogwoka et al.	2015	Quênia	Predição de Desempenho	3
Al-Barrak and Al-Razgan	2015	Arabia	Predição de Desempenho	3
Nakhkob and Khademi	2016	Irã	Predição de Evasão	3
Kohli and Birla	2016	Índia	Predição de Evasão	3
Gonzalez et al.	2016	México	Predição de Evasão	2
Patel and Dharwa	2016	Índia	Predição de Evasão	1
Cunha et al.	2016	Brasil	Análise de Padrões	0
Burgos et al.	2017	Espanha	Predição de Evasão	8
Oliveira Júnior et al.	2017	Brasil	Seleção de Atributos	0
Ahuja and Kankane	2017	Índia	Predição de Evasão	0
Sultana et al.	2017	Paquistão	Predição de Desempenho	0
Chaturvedi	2017	Índia	Revisão da Literatura	0
Couto and Santana	2017	Brasil	Predição de Evasão	0
Santos et al.	2018	Brasil	Revisão da Literatura	0
Sarra et al.	2018	Itália	Predição de Evasão	0
Miguéis et al.	2018	Portugal	Predição de Desempenho	0
Zaffar et al.	2018	Brasil	Seleção de Atributos	0
Mason et al.	2018	EUA	Predição de Evasão	0
Backenkohler et al.	2018	Alemanha	Predição de Evasão	0
Alban and Mauricio	2019	Equador	Predição de Evasão	1
Tasnim et al.	2019	Bangladesh	Predição de Evasão	1
Santoso et al.	2019	Indonésia	Predição de Desempenho	1
Ramentol et al.	2019	Suécia	Predição de Evasão	1
Khan et al.	2019	Índia	Big Data	0

[5] bit.ly/selecao-artigos

O abandono dos estudantes é um elemento extremamente importante para o gerenciamento de matrículas, pois afeta não apenas a classificação das universidades, mas também a reputação da IES, o que compromete apoio financeiro [Delen, 2010].

No contexto internacional, destacaram-se os trabalhos de Romero and Ventura [2013], Yukselturk et al. [2014], Tekin [2014], Sin and Muthu [2015], Thakar and Mehta [2015] e Guarín et al. [2015]. Além destes, os trabalhos recentes de Ahuja and Kankane [2017], Sultana et al. [2017], Sarra et al. [2018], Alban and Mauricio [2019] e Tasnim et al. [2019] também devem ser enfatizados pois analisam dados de estudantes de graduação com o objetivo de identificar alunos com maiores chances de abandono.

A abordagem de Sultana et al. [2017] analisou dados de estudantes de engenharia elétrica, fornecidos por meio de diferentes questionários. O trabalho fez uso de diferentes métodos, como AD, RL, *Naive Bayes* (NB) e Redes Neurais (RN), para explorar características cognitivas e não cognitivas dos estudantes, a fim de prever os resultados da evasão. Os autores descrevem que aspectos cognitivos melhoram a precisão preditiva nos métodos da árvore de decisão, mas não em outros métodos.

De maneira complementar, Burgos et al. [2017] também utiliza a RL para a análise de desempenho dos alunos de vários cursos. O objetivo do trabalho é a proposta de um método para detecção de abandono que permita produzir, em tempo hábil, um plano de ação capaz de evitá-lo. Os resultados mostraram que a previsão combinada com o plano de ação ajudou a reduzir o abandono durante os anos letivos de 2014 e 2015, em comparação com os outros anos em que essa abordagem não havia sido implementada.

Outro trabalho relevante é apresentado por Chaturvedi [2017], na forma de uma compilação de vários estudos na área. Embora o trabalho de Chaturvedi não tenha discutido a análise dos dados, ele apresenta algumas ferramentas que podem ser usadas para analisar os conjuntos de dados e resume os algoritmos usados na área descrevendo suas características básicas.

Sarra et al. [2018] usou um método de regressão bayesiano para analisar as respostas ao questionário de mais de 500 estudantes envolvendo questões destinadas a obter informações sobre competências, motivações e resiliência acadêmicas. Os autores verificaram os fatores relevantes para identificar os grupos com maior chance de abandono, incluindo alguns destes fatores na forma de indicadores motivacionais, indicadores de dificuldade e indicadores de satisfação durante a vida acadêmica.

Ahuja and Kankane [2017] usaram vários algoritmos como NB, RL, *K-Nearest Neighbours* (KNN), *Random Forest* (RF), AD para prever a probabilidade de conclusão do curso de graduação. Os dados utilizados incluíram notas dos alunos e dados sociodemográficos. Os autores compararam os métodos e, com base nos experimentos, identificaram que os algoritmos RF e AD foram os melhores para classificar e prever os alunos a graduar. Porém, de acordo com os resultados dos testes efetuados na seleção de atributos, o algoritmo com melhor acurácia para prever a insatisfação dos alunos foi a Árvore de Decisão (AD), pois favorece a interpretação dos fatores que levaram ao resultado final. Os resultados mostram que a abordagem é viável, com uma precisão de até 97,87% nas experiências realizadas.

No Brasil, o enfoque direcionou-se às pesquisas de Costa et al. [2013], Manhães et al. [2014], Manhães et al. [2014], Barbosa Manhães et al. [2015], Cunha et al. [2016], Oliveira Júnior et al. [2017], Couto and Santana [2017]. e Henley [2018].

O trabalho de Manhães et al. [2014] visa auxiliar os gerentes da instituição. Os autores identificaram atributos que ajudam a detectar estudantes com desempenho insatisfatório ou risco de abandono. Os dados de estudantes de graduação de uma grande instituição educacional foram analisados usando os seguintes algoritmos de classificação: AD, Máquina de Vetor de Suporte (SVM), NB e *Multilayer Perceptron* (MLP). O modelo MLP foi utilizado para apresentar uma abordagem quantitativa. Os autores estimaram que o uso dos métodos de EDM ajudam a identificar os alunos com maiores chances de evasão e oferecem à instituição uma forma de análise adicional para o problema da alta taxa de evasão.

Em um trabalho subsequente, Manhães et al. [2014] apresentam uma arquitetura denominada *WAVE* que utiliza técnicas de EDM para prever e identificar estudantes que estão em risco de abandono. A arquitetura usa apenas dados do aluno armazenados no sistema de gerenciamento acadêmico, não exigindo dados sociais ou econômicos.

Na pesquisa de Cunha et al. [2016] foram analisados dados de cursos em diferentes níveis de ensino para detectar quais atributos mais influenciaram no abandono e reprovação, a fim de traçar um perfil dos alunos evadidos e alunos com baixo desempenho. Para isso, aplicou-se o método AD junto à ferramenta *Analysis Services*. Com base nos resultados obtidos, os autores levantaram algumas ações preventivas para que a gestão da instituição possa tomar decisões a fim de minimizar as taxas de abandono e reprovação.

Couto and Santana [2017] apresentam um documento que visa criar subsídios para auxiliar os gerentes de instituições de ensino superior a identificar estudantes propensos a desistência ou retenção em seus cursos. Para isso, os autores utilizaram algoritmos de classificação aplicados aos dados de graduação. Os métodos RF e Redes Bayesianas (RB) foram os mais satisfatórios para analisar os dados e, assim, auxiliar os gestores.

Com uma abordagem de estudo de caso semelhante ao desta pesquisa, o trabalho de Henley [2018] apresentou uma arquitetura de sistema para analisar a evasão: PRELUDE [6]. Nesse trabalho, Henley propõe um modelo de predição para evasão usando aprendizado logico-relacional através de *Prolog* [7] para gerar regras de inferência que descrevem um comportamento de evasão.

Já para reduzir o número de alunos evadidos, a pesquisa de Oliveira Júnior et al. [2017] apresenta uma solução que visa identificar padrões que ajudariam os gestores na tomada de decisões. Esse trabalho propõe um modelo preditivo de evasão por meio da criação e seleção de atributos de bancos de dados educacionais da Universidade Federal do Paraná.

Baseado nos trabalhos identificados, vislumbrou-se a possibilidade de utilizar a seleção automática de atributos por meio de algoritmos genéticos, associada ao modelo preditivo de árvores de decisão para o problema de evasão, aplicados no contexto da Universidade Federal Fluminense, sendo este o foco do presente trabalho.

[6] Do inglês, *Prediction using machine Learning for Undergraduate courses*.
[7] É uma linguagem de programação lógica expressa na forma de regras [Clocksin and Mellish, 2012].

4- EDW – Um Data Warehouse Educacional

Ao buscar soluções para o problema da evasão, algumas IFES têm adotado soluções tecnológicas baseadas em sistemas de apoio à tomada de decisão. A área de conhecimento que abrange esses sistemas é conhecida como inteligência de negócios ou *Business Intelligence* (BI) [Negash, 2004; Davenport, 2012; Inmon et al., 1997; Kimball and Ross, 2011].

Sistemas de BI usam funcionalidades de banco de dados, processos de negócio, modelagem e visualização de dados para auxiliar no entendimento e tomada de decisão [Chen et al., 2012]. Tais funcionalidades permitem aplicações sofisticadas no desenvolvimento de análises de dados e projeções analíticas, que podem auxiliar no aprimoramento de ações de tomada de decisão e em melhorias nos processos de negócio [Shim et al., 2002].

Dentre as funcionalidades de um sistema de BI, o *Data Warehouse* (DW) assume um papel de destaque [Olszak and Ziemba, 2007]. O DW consiste em uma coleção de bancos de dados que visa manter dados íntegros, consistentes e centralizados sobre as áreas estratégicas e de negócios, a fim de aprimorar o processo decisório das atividades de gestão [Inmon et al., 1997]. Diversas organizações estão aprimorando seus sistemas de tomada de decisão utilizando DW, conforme pode ser visto nos seguintes trabalhos: Rudra and Yeo [1999]; Olszak and Ziemba [2007]; Ranjan [2009]; Kimball and Ross [2011]; Davenport [2012] e Inmon and Linstedt [2014].

A fim de contribuir de forma semelhante aos trabalhos mencionados, este capítulo apresenta o EDW – uma base de dados analítica capaz de auxiliar a tomada de decisão na gestão acadêmica da Universidade Federal Fluminense (UFF). Dessa forma, a Seção 4.1 descreve o processo analítico para identificar os sistemas de informação da instituição. Na Seção 4.2, são apresentadas as fontes de dados selecionadas para compor uma base de dados analítica sobre evasão, as quais são unificadas através de um processo de ETL descrito na Seção 4.3. Em seguida, a Seção 4.4 detalha o modelo de dados analítico na forma do EDW e a Seção 4.5 evidencia algumas das análises produzidas nesta pesquisa. Ao final do capítulo, a Seção 4.6 destaca algumas considerações e observações sobre o EDW.

4.1- Processo analítico e domínio da informação

O planejamento de recursos acadêmicos é um procedimento administrativo complexo baseado na análise extensiva das demandas relacionadas à atividade educacional. Isso envolve corpo administrativo e corpo docente, recursos didáticos, cursos oferecidos, estrutura de curso, currículos, processo de matrícula, vagas ociosas, retenção, evasão, dentre outras necessidades.

Com a utilização de tecnologias de informação mais recentes, o ambiente operacional das universidades tem sido transformado por meio de sistemas computacionais que auxiliam os processos de negócio. Sendo assim, os dados pertencentes às atividades identificadas como relevantes para uma gestão mais eficiente passam a ser armazenados computacionalmente [Delavari et al., 2005].

A Figura 6 apresenta, de forma generalizada, o fluxo da informação para o processo analítico que envolve o conjunto de dados necessário para construir o EDW em uma IES. Neste processo, são representadas as seguintes entidades: IFES, Aluno, Sistemas de Informação, Bases de Dados e o *Data Warehouse* Educacional (EDW).

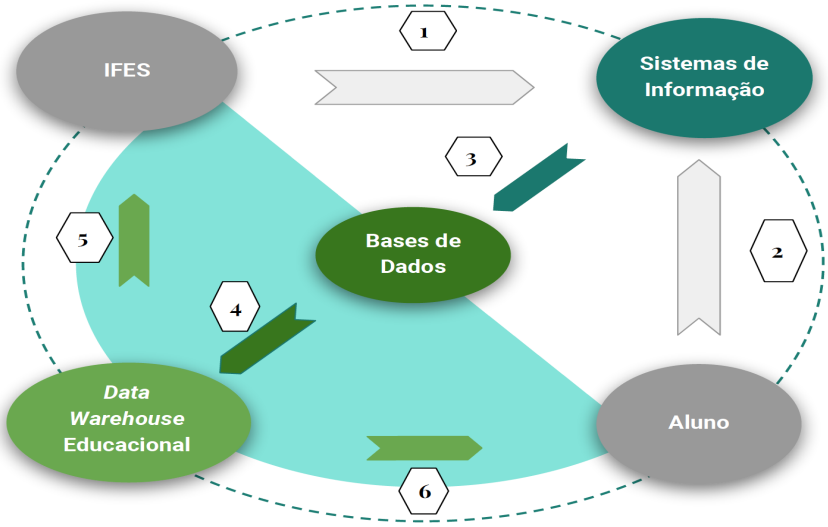


Figura 6 – Fluxo de Dados para o *Data Warehouse* Educacional [Santos et al., 2019]

.

As etapas 1 e 2 da Figura 6 representam as interações dos alunos e dos funcionários da instituição por meio das atividades acadêmica e estudantil, assim como as atividades administrativas e gerenciais nos diversos sistemas de informação (SI) da

IFES. Na etapa 3, as informações inseridas nos SI durante as etapas anteriores, são persistidas, atualmente, em bases de dados pertencentes a esses sistemas. Em seguida, a etapa 4 apresenta o processo de ETL composto pelas seguintes atividades: extração das informações das fontes de dados; transformação dos dados baseado nas correções necessárias; e carregamento desses dados no EDW. Por fim, as etapas 5 e 6 compreendem a obtenção de conhecimento por meio da base de dados analítica produzida pelo EDW. Na etapa 5, são disponibilizadas informações analíticas de apoio à tomada de decisão; e, na etapa 6, dados com base no perfil do aluno, do curso e de suas interações armazenadas.

Após detalhado o processo analítico do EDW, é preciso compreender o contexto das atividades mediante o domínio da informação, o qual será utilizado como o principal referencial das análises produzidas. O domínio da informação do EDW é pertencente ao problema da evasão e será atendido com base nos requisitos de negócio apresentados na Tabela 4, a seguir.

Tabela 4 – Requisitos do EDW

Requisito	Descrição
1	Analisar a evasão com relação ao desempenho acadêmico
2	Analisar a evasão com base nos cursos de graduação
3	Analisar a evasão considerando a localidade de curso
4	Analisar a evasão considerando as bolsas de estudos

Cada um dos requisitos da Tabela 4 corresponde a uma perspectiva específica sobre o problema da evasão. Diante das descrições propostas para compor o EDW, os principais sistemas de informação escolhidos para abarcar o domínio da informação são descritos na seção seguinte.

4.2- Descrição do Ambiente Transacional

Os sistemas acadêmicos da UFF estão distribuídos em conjuntos distintos de dados, pertencentes às áreas e unidades organizacionais da instituição, na forma de pró-reitorias, superintendências, departamentos e coordenações. Devido a essa estrutura

funcional, os sistemas são geridos, em grande maioria, diretamente pela sua unidade organizacional, responsável pela regra de negócio daquele referido domínio de informação. Os principais sistemas são o idUFF, PIBIC, SISBOL e o MONITORIA.

O idUFF é o sistema de identificação única da UFF com o objetivo de centralizar os dados das pessoas que têm ou tiveram vínculo com a universidade e, principalmente, as informações acadêmicas pertencentes aos cursos de graduação geridos pela pró-reitoria de graduação [Fluminense, 2014].

O sistema PIBIC foi desenvolvido para administrar o processo de submissão dos projetos de pesquisa, a seleção e concessão de bolsa de Iniciação Científica, assim como a avaliação destes projetos pela pró-reitoria de pesquisa.

O sistema SISBOL auxilia no gerenciamento das bolsas de assistência estudantil da pró-reitoria de assistência ao estudante. Esse sistema otimiza a gestão de informações em torno dos programas de assistência estudantil, relativos aos editais de participação, processo seletivo e concessão de bolsas.

O sistema MONITORIA visa auxiliar o processo de submissão de alunos-monitores dos cursos de graduação nas disciplinas ofertadas para tal proposta, tendo como foco contribuir no auxílio da aprendizagem dos alunos.

O EDW proposto teve que integrar as fontes de dados desses sistemas em um ambiente analítico de banco de dados utilizando várias técnicas, denominadas de modelagem multidimensional, extração, transformação, carregamento e visualização de dados.

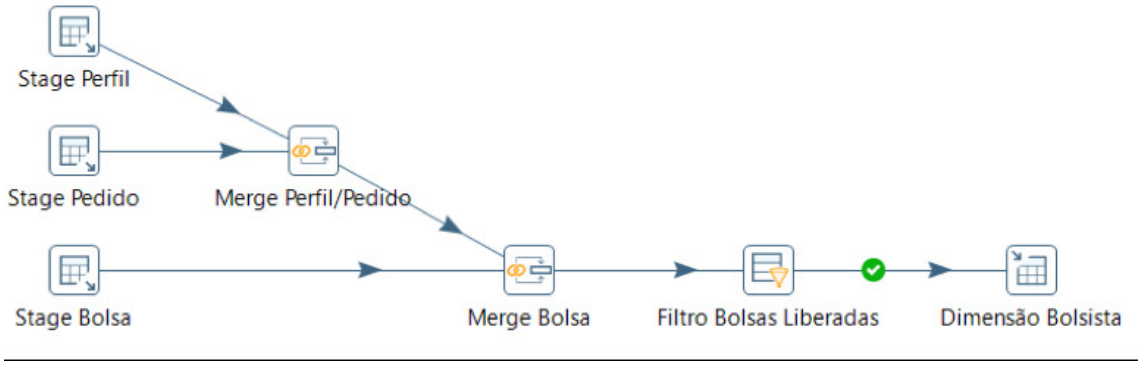
4.3- Processo de ETL

O processo de ETL ocorre em três etapas: extração, transformação e carga [Kimball and Caserta, 2004]. Neste trabalho, a primeira refere-se à extração dos dados dos sistemas acadêmicos, consistindo na geração dos arquivos das tabelas de cada fonte de dados no formato *.csv*. A segunda detalha a transformação desses dados, a fim de conceder ajustes nos valores dos atributos e respectivas padronizações, como por exemplo, unidades de tempo. E a terceira corresponde ao carregamento desses dados, conforme o ambiente analítico.

É importante ressaltar que, durante as etapas de exportação dos dados, é executado um procedimento que os carrega em uma base de dados de preparação (*staging area*). Essa base de dados é utilizada como etapa intermediária para auxiliar na limpeza e transformação dos dados necessários, que serão armazenados no EDW.

A Figura 7 apresenta o processo de ETL detalhando dois procedimentos de extração dentre os vários desempenhados na implementação do EDW. No procedimento (a), os dados são extraídos das tabelas de origem do sistema de bolsas: Perfil, Pedido e Bolsa. Nesse ETL, os dados dessas tabelas são unificados e selecionados para serem inseridos na Dimensão Bolsista. No procedimento (b), é feito um dos processos de carga da tabela Fato Evasão, unificando os dados de Bolsista e de Aluno para filtrar os alunos classificados como evadidos. Ao fim desse procedimento, todas as ocorrências de alunos bolsistas evadidos são inseridas na tabela Fato Evasão.

a)



b)

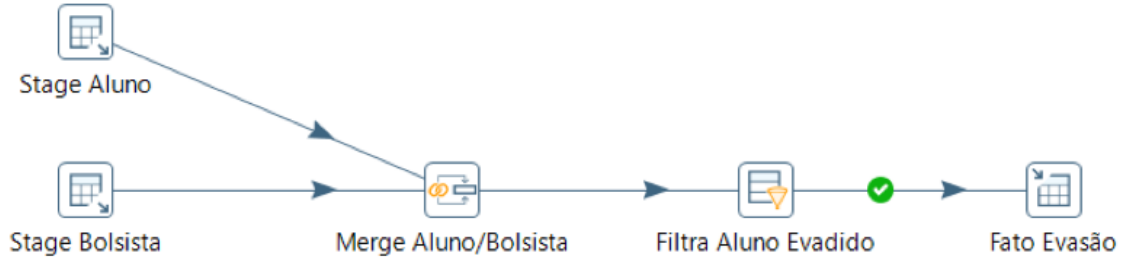


Figura 7 – Etapas do processo de ETL: a) Carga da tabela Dimensão Bolsista e b) Carga da tabela Fato Evasão

4.4- Modelagem do EDW

Modelagem dimensional é uma técnica de projeto de banco de dados para organizar dimensões e fatos em um modelo analítico [Kimball and Ross, 2011; Inmon and Linstedt, 2014]. A Figura 8 apresenta a modelagem multidimensional desenvolvida neste trabalho.

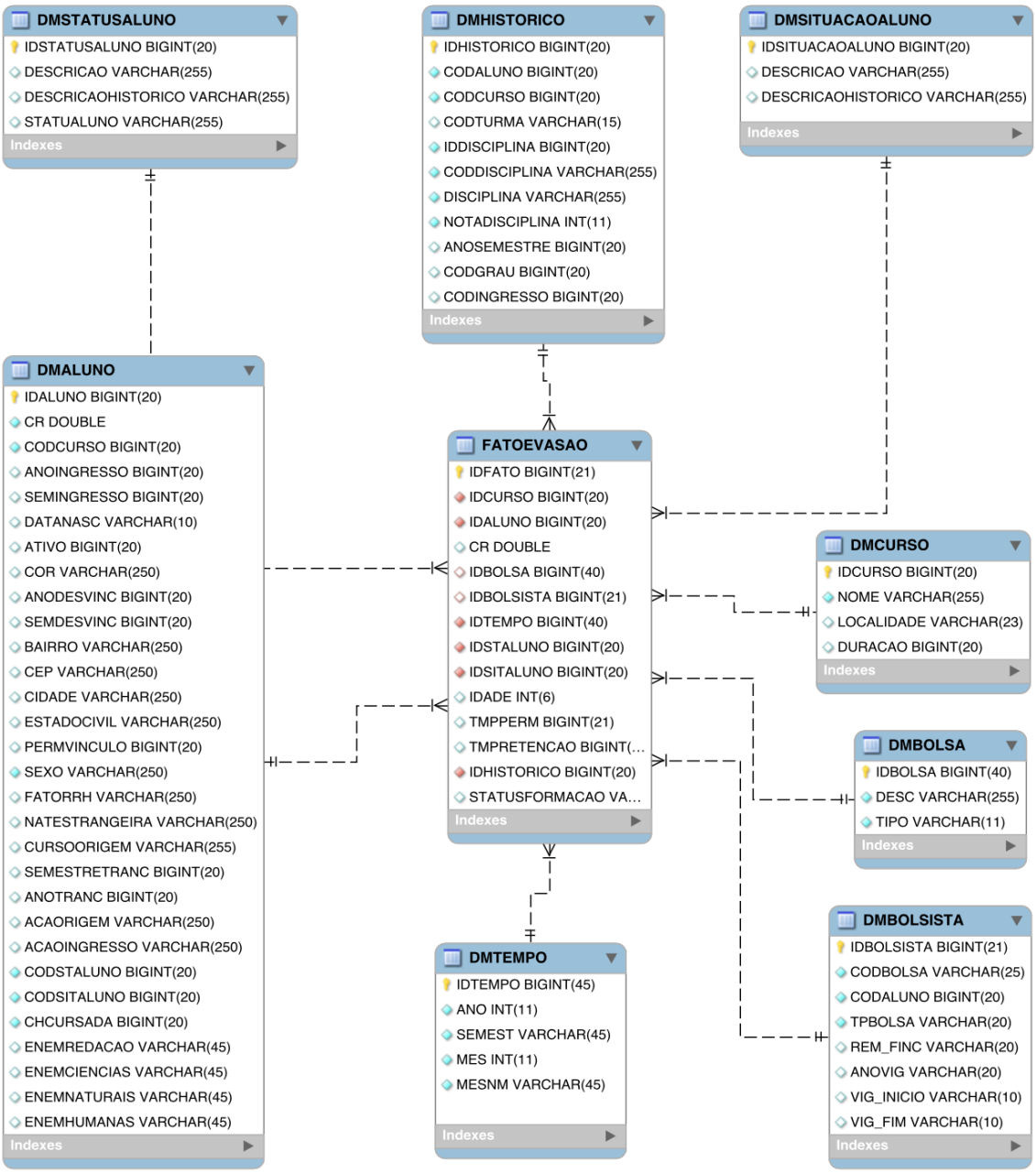


Figura 8 – Modelo do EDW para Evasão

Para atender aos requisitos de negócio, foram criadas oito tabelas de dimensões e uma tabela fato, representando as seguintes entidades: Aluno, Bolsa, Bolsista, Curso, Histórico, Status do aluno, Situação do aluno, Tempo e assunto Evasão. Cada uma dessas entidades serão descritas a seguir.

A dimensão Aluno representa as informações do indivíduo, destacando as notas de entrada, dados socioeconômicos, curso matriculado, carga horária cursada, coeficiente de desempenho acumulado, entre outras. A partir disso, é possível explorar as características do aluno correlacionando com as outras dimensões. Sendo assim, há também a dimensão Histórico contendo o desempenho acadêmico dos alunos a partir das avaliações das disciplinas cursadas.

Há também as dimensões de Status e Situação do aluno que, respectivamente, descrevem se ele está ativo ou não, mediante a situação de: matriculado, trancado ou cancelado, conforme suas especifidades (i.e., desempenho, frequência e etc.).

As dimensões Curso, Bolsa e Bolsista complementam o EDW detalhando as informações pertinentes a cada uma das entidades. A dimensão Curso traz consigo duração e localidade, semelhante à dimensão Bolsa, que descreve tipo de fomento e o projeto/programa de referida bolsa. Por intermédio da dimensão Bolsa, os alunos vinculados tornam-se bolsistas, podendo também ser caracterizados por código de bolsa, período de vigência e remuneração.

De forma centralizada, tem-se a tabela fato Evasão que agrega cada uma das dimensões reveladas, tornando essas dimensões uma projeção específica para o problema evasão. Essa projeção pode ser também agrupada com uma ou mais dimensões, o que enriquece o contexto informacional. Além disso, é possível avaliar essas dimensões sobre uma perspectiva de dados históricos, proporcionada pela caracterização da dimensão Tempo.

A partir dessa modelagem, tornou-se viável atender as análises solicitadas pelos requisitos, pois as fontes de dados que estavam dissociadas foram unificadas. Por exemplo, através do EDW, pode-se fazer consultas que listem todos os alunos bolsistas graduados em todos os cursos ou especificamente em um curso.

Outra maneira de avaliar a viabilidade do EDW pode ser mencionada na forma de questões de negócio. Para elucidar a capacidade analítica do EDW, seguem algumas questões que antes não poderiam ser respondidas, mas que atualmente são factíveis. Quais alunos bolsistas evadiram no ano de 2017? Quais alunos bolsistas graduaram com

CR acima de 9 na localidade de Niterói? Qual o programa de bolsas que obteve maior número de alunos graduados?

Além das análises exemplificadas acima, muitas outras são possíveis, pois o conjunto informacional atendido pelo EDW é bastante vasto. A tabela Aluno, por exemplo, possui informações que favorecem análises sobre etnia, ações afirmativas, notas de entrada do ENEM (Exame Nacional do Ensino Médio), bairro, cidade de origem, entre outras. Com o EDW desenvolvido, pode-se também desenvolver novas análises além das já mencionadas. Por exemplo, abordagens com análises mais robustas que possibilitem prever alunos em risco de evasão e, principalmente, identificar padrões de evasão.

É importante destacar que as análises desenvolvidas tornaram-se factíveis devido ao EDW. Anteriormente, elas não eram possíveis; pois, além das bases de dados serem desagregadas, a UFF não tinha nenhuma solução analítica com alcance institucional. Com o intuito de demonstrar o potencial desta abordagem, a próxima seção apresenta as análises que atendem aos requisitos elencados, através de gráficos e relatórios.

4.5- Uma perspectiva sobre a Evasão através do EDW

Esta seção apresenta as análises produzidas através dos dados disponibilizados pelo EDW que contem cerca de 80 mil registros dos alunos dos 106 cursos de graduação oferecidos pela UFF, no período de 2005 até 2018 . O conjunto de dados produzido durante o processamento analítico, passa a ser utilizado para a criação de relatórios, geração de gráficos e *dashboards*. Essas análises trazem informações que auxiliam no processo de tomada de decisão sobre o problema da evasão, a fim de atender aos requisitos de negócios elencados na Tabela 4 da Seção 4.1.

Mediante o **Requisito #1**, uma perspectiva sobre o desempenho acadêmico dos alunos foi fornecida por meio de um histograma apresentado na Figura 9. Nesta figura, é possível perceber que há uma frequência maior da amostra de alunos com baixo desempenho nas regiões de CR ’0’ a ’3’, sendo parte significativa deste grupo constituída por alunos evadidos e com maior ocorrência de evasão no primeiro ano de ingresso.

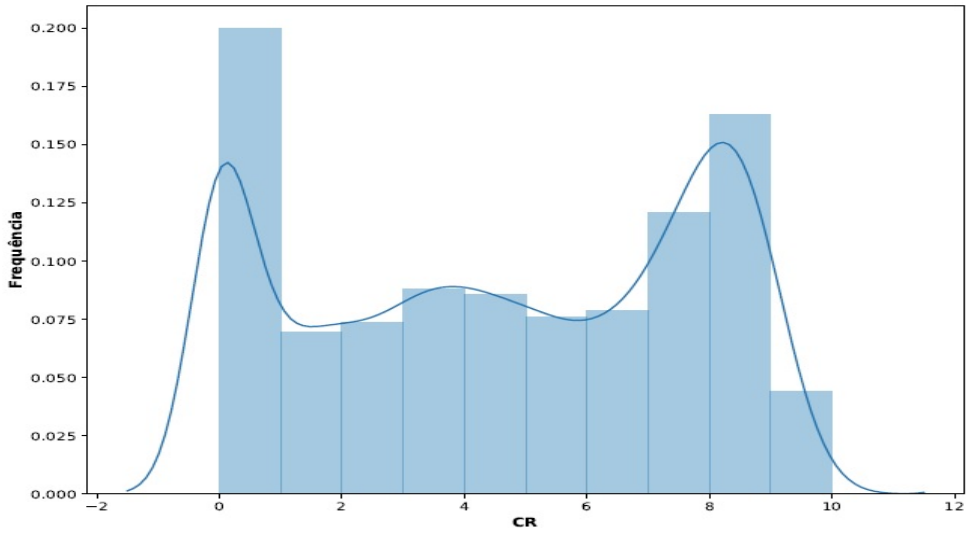


Figura 9 – Histograma do CR dos alunos graduados e evadidos

Para atender ao **Requisito #2**, é exibido um ranqueamento dos cursos nos quais ocorreu maior número de alunos evadidos, dentre os anos de 2005 a 2015, conforme a Tabela 5.

Tabela 5 – Ranking de Evasão por Curso

Curso	Nº de Alunos Evadidos
Matemática	1417
Letras	1288
Ciências Econômicas	1114
Física	1029
Farmácia	1025
Geografia	934
História	914
Administração Pública	876
Serviço social	820
Ciências Contábeis	706
Ciências da Computação	701
Direito	698
Administração	689
Ciências sociais	655

Nesta tabela, é possível identificar que em cursos como matemática, letras, economia e física a evasão ainda apresenta-se com a mesma necessidade de atenção de anos anteriores. Por isso, soluções como o EDW, que proporcionam uma visão holística sobre o contexto informacional da evasão, podem auxiliar em decisões melhor direcionadas, possibilitando ações mais assertivas diante desse problema.

Tratando do **Requisito #3**, a Figura 10 apresenta um gráfico de informação georeferenciada sobre a localidade de campi da UFF, relacionado com o percentual da faixa de coeficiente de rendimento do aluno matriculado no curso da referida localidade. O intuito dessa informação é destacar as faixas de CR por cores, representando essas faixas com grupos de valores.

Diante dessa distribuição, pode-se estender a análise feita, comparando também o perfil dos alunos formados ou evadidos, exibido pela Tabela 7, de maneira correlacionada com o quantitativo de alunos bolsistas e não bolsistas e a localidade do curso.

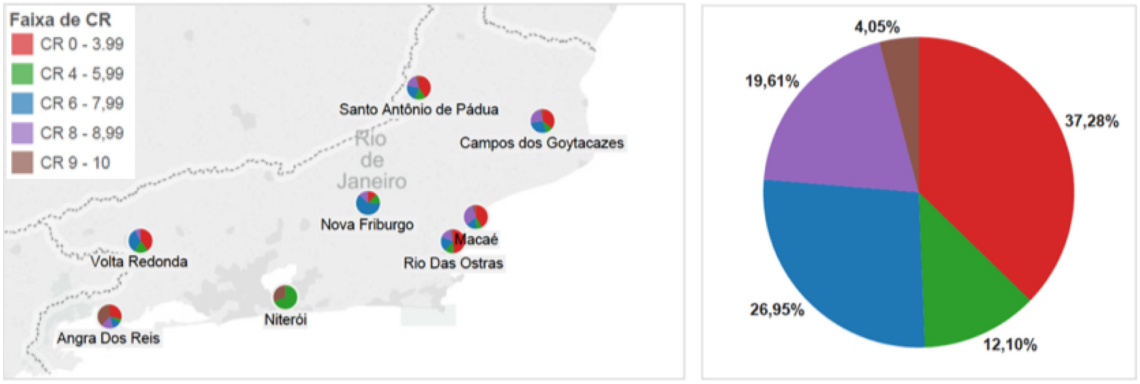


Figura 10 – Distribuição dos Alunos por Rendimento e Localidade.

Por último, mas não menos importante, foi efetuado um detalhamento que atende ao **Requisito #4**, onde é apresentada a quantidade de alunos que participaram de algum programa de fomento da UFF, conforme a Tabela 6. Nesta tabela, há uma sumarização do "Nº de Alunos" vinculado ao "Tipo de Bolsa", considerando o "Ano" de credenciamento do aluno. Ao analisar a Tabela 6, identifica-se a quantidade de alunos que possuem apoio de fomento pela instituição, conforme o tipo de bolsa oferecida nos referidos anos. É importante destacar que dentre os anos de 2013, 2014 e 2015, as bolsas de assistência têm recebido uma porção maior do orçamento de fomento na UFF. Esse fato é decorrente da expansão do Plano Nacional de Assistência Estudantil (PNAES) no orçamento da instituição, reajustado em 2013.

Tabela 6 – Quantidade de Bolsistas por Tipo de Bolsa e Ano

Ano	Tipo de Bolsa	Nº de Alunos
	Assistência	141
2011	Iniciação Científica	2286
	Monitoria	1124
	Assistência	1137
2012	Iniciação Científica	1856
	Monitoria	1133
	Assistência	2357
2013	Iniciação Científica	1718
	Monitoria	1270
	Assistência	2662
2014	Iniciação Científica	1443
	Monitoria	1308
	Assistência	2004
2015	Iniciação Científica	1346
	Monitoria	1174

A Tabela 7 também apresenta um relatório que compara a distribuição dos alunos formados e evadidos entre os anos de 2005 a 2015, sendo estes vinculados aos programas de bolsas ou não. As colunas referentes aos *Tipo I* e *Tipo II* são, respectivamente, a representação dos alunos bolsistas e não bolsistas. A distribuição apresentada é formada por um agrupamento de alunos evadidos que participaram ou não de programas de fomento.

Tabela 7 – Perfil dos Alunos Formados e Evadidos por Curso

Curso	Formados		Evadidos	
	Tipo I	***Tipo II***	***Tipo I***	***Tipo II***
Direito	1989	152	694	4
Administração	1519	41	689	0
Ciências Contábeis	1345	15	703	3
Serviço Social	1298	195	803	17
Medicina	1278	355	117	2
Matemática	1122	73	1380	37
Ciências Econômicas	1061	55	1110	4
Pedagogia	1017	135	598	11
Letras	1014	163	1256	32
Comunicação Social	999	77	475	3

4.6- Discussão

Neste capítulo foi detalhado um banco de dados analítico para fornecer uma solução capaz de auxiliar a gestão acadêmica da UFF na identificação de padrões que promovem a evasão. A identificação desses padrões pode permitir um maior êxito na redução dos impactos da evasão e, consequentemente, uma melhor gestão de recursos financeiros e humanos.

Os relatórios implementados apresentaram uma utilização prática sobre o problema de evasão. No entanto, apesar da utilidade dos relatórios apresentados, é necessário algum entendimento sobre o tema da evasão e uma visão analítica sobre a gestão acadêmica de cursos de graduação para avaliar os resultados já implementados e promover ações de melhoria.

Durante a fase de coleta dos dados, ficou aparente a necessidade de complementar as informações da base de dados. De acordo com observações iniciais sobre o EDW, algumas novas informações podem ser consideradas primordiais para aprimoramentos no processo de gestão acadêmica. Essa novas informações devem ser relacionadas ao histórico pregresso do aluno, como atributos que descrevam as instituições de ensino básico e ensino médio dos alunos de graduação: nome da instituição, sistema de ensino (municipal, estadual ou privado), localidade da instituição (bairro, cidade e estado), ano de ingresso e ano de conclusão.

Por fim, foram percebidas novas etapas de desenvolvimento dessa pesquisa, permitindo aplicar também técnicas de mineração de dados, reconhecimento de padrões e modelos de predição no tema evasão, demonstrando a versatilidade e robustez do sistema proposto.Tais técnicas são exploradas nos capítulos seguintes.

5- *EvolveDTree*: Um Sistema de Predição de Evasão

Tarefas de predição compreendem algumas das atividades mais difundidas em Aprendizado de Máquina(AM) com aplicações consolidadas em várias áreas [Baldi et al., 2000]. Dentre as técnicas utilizadas em tarefas dessa natureza, modelos baseados em árvores têm contribuído de maneira relevante para AM [Breiman et al., 1984; Breiman, 2001; Cieslak and Chawla, 2008; Chen and Guestrin, 2016]. Sob essa perspectiva, este capítulo apresenta abordagem que denominamos de *EvolveDTree* (do inglês *"**Evolve**d **D**ecision **Tree**"*), composto por um modelo que utiliza a técnica de Árvore de Decisão (AD) associada a um Algoritmo Genético (AG).

A AD foi escolhida porque permite uma representação visual do problema, o que favorece a compreensão do resultado e facilita a vida de gestores, ao contrário do que aconteceria com outras técnicas, como por exemplo, o SVM. Contudo, AD é muito suscetiva ao problema de *overfitting*, pois tende a estimar de forma enviesada os dados de treinamento, refletindo em resultados ruins quando aplicados em novos dados [Timofeev, 2004; Horning, 2013]. Com o intuito de mitigar esse problema, o AG é utilizado para realizar a seleção de atributos e melhorar a capacidade de generalização por meio de redução de dimensionalidade [Stein et al., 2005; Farissi et al., 2020].

A abordagem desenvolvida combinando as duas técnicas (AG e AD) visa uma execução eficiente no conjunto de dados avaliado, o qual pode ser composto por até uma centena de atributos. Nesse sentido, a Seção 5.1 apresenta de forma geral o funcionamento do *EvolveDTree*. Já a Seção 5.2 descreve as etapas de AG e a Seção 5.3 especifica AD para a tarefa de predição. Por fim, a Seção 5.4 discute sobre as pesquisas que aplicaram de forma unificada AG e AD em diferentes problemas.

5.1- Visão Geral sobre o *EvolveDTree*

O conjunto de dados utilizado nesta pesquisa, fornecido pelo EDW, contém informações sobre o aluno com enfoques acadêmico e sociodemográfico. Esses dados são caracterizados por 28 atributos, conforme pode ser visto no Apêndice B, representando 12.969 alunos, ingressos nos anos de 2012 até 2014. Considera-se que tais alunos podem ter evadido ou graduado até o ano de 2018.

A partir da obtenção dos dados, a próxima fase consiste em realizar o seu pré-processamento. Esta fase implica em transformar, converter e particionar os dados. As informações relativas a Cor, Curso, Ação Afirmativa e Turno foram ajustadas de forma que cada um dos seus valores categóricos fossem transformados em atributos, baseando-se na técnica *one-hot encoding* [Cerda et al., 2018]. Em seguida, os dados foram divididos em dez subconjuntos pela técnica de validação cruzada estratificada [Purushotham and Tripathy, 2011], a fim de melhorar o processo de AM [Diamantidis et al., 2000]. Foi decidido usar os alunos ingressantes em 2012 e 2013 como sendo o conjunto de treinamento e, os de 2014, o conjunto de validação.

Após a etapa de pré-processamento, é realizada uma redução de dimensionalidade utilizando um AG. Por fim, os atributos selecionados pelo AG são submetidos à AD, que recebe como entrada dez partições de dados de forma iterativa, sendo cada uma delas avaliadas segundo um conjunto de métricas escolhidas.

Com o objetivo de entender mais detalhadamente o funcionamento do AG e da AD desenvolvidos neste trabalho, as próximas seções apresentam algumas definições e como tais técnicas foram adaptadas para o problema da evasão.

5.2- Algoritmo Genético

O AG desenvolvido nesta pesquisa teve como base o trabalho de Farissi et al. [2020]. Em geral, um AG tem quatro componentes conhecidos [Goldberg and Holland, 1988]: $a)$ Uma população de indivíduos, em que cada um representa uma amostra na forma de cromossomo, composto por genes; $b)$ Operadores genéticos, como cruzamento

e mutação, que exploram novas combinações de genes e, ao mesmo tempo, mantém um percentual das informações contidas no cromossomos atuais; $c)$ Uma função de seleção, que decide como escolher bons indivíduos da população atual para criar a próxima geração; e $d)$ Uma função de avaliação pela qual é possível avaliar se um indivíduo está apto a compor a solução final.

A compreensão de cada um destes componentes é fundamental para o entendimento do funcionamento do *EvolveDTree*. Assim, as próximas seções descrevem de forma mais detalhada cada um deles.

5.2.1 População inicial

A atividade de gerar uma população para AG é o processo no qual são criados os primeiros indivíduos capazes de representar a atividade definida. Nessa população é preferível que os indivíduos sejam os mais diferentes possíveis, distribuídos aleatoriamente no grupo das possíveis soluções [Zhu et al., 2006]. Na Figura 11 é apresentado um conjunto de atributos, em que cada letra representa um atributo do conjunto de dados inicial.

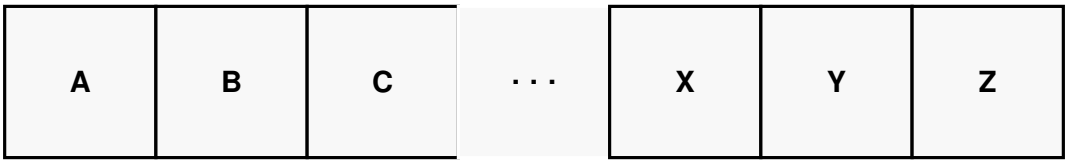


Figura 11 – Representação dos atributos no conjunto de dados

Para efetuar a discretização dos atributos sob uma nomenclatura de AG, cada atributo passa a ser identificado como gene e estes formarão um cromossomo. Discretizar um atributo é um procedimento utilizado neste trabalho para substituir a posição do atributo por um valor numérico [Liu et al., 2002].

Cada cromossomo será formado por genes de 0s e 1s, onde 0 significa que o atributo foi descartado e 1 que o atributo foi incluído. A Figura 12 representa um gene.

Durante o processo de seleção de atributos, cada um dos atributos é escolhido aleatoriamente para compor o cromossomo, conforme a Figura 13. Esta figura tam-

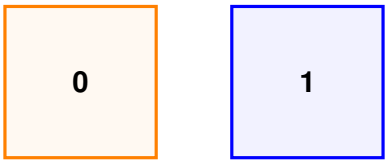


Figura 12 – Representação dos genes do AG

bém representa a conversão dos atributos na forma de genes, bem como a etapa de discretização dos atributos para o AG.

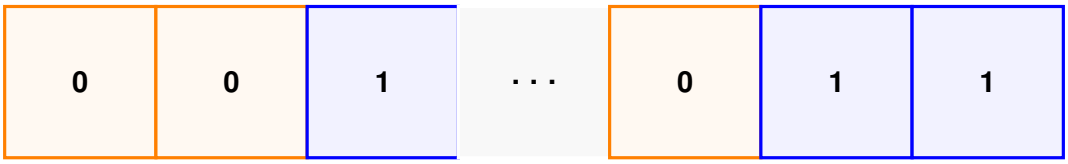


Figura 13 – Representação do cromossomo para o AG

Uma das práticas mais comuns na representação de uma população inicial para AG consiste no método baseado em representação binária, neste os indivíduos têm seus genes preenchidos por uma combinação de 0s e 1s. Ao final dessa etapa, espera-se que a população inicial de indivíduos represente soluções do problema em questão [Zhou et al., 2011].

Nesta dissertação a representação utiliza uma sequência binária de 0s e 1s com um tamanho fixo n de posições, menor ou igual ao número de características de indivíduos do tipo *Aluno*.

5.2.2 Operadores genéticos

Este componente é responsável por evoluir a população através de gerações, estendendo-se iterativamente até chegar a um resultado satisfatório. Esse procedimento torna-se necessário para que a população obtenha variabilidade e, ao mesmo tempo, mantenha uma proporção das características dos indivíduos ascendentes [Forrest, 1993]. São operadores genéticos as funções de cruzamento e de mutação:

1. **Cruzamento**: a operação de cruzamento (do inglês, *crossover*) efetua a recombinação dos genes das gerações ascendentes, no intuito de garantir que as gerações

descendentes herdem um percentual de genes ancestrais. O cruzamento é o operador genético prioritário e, por isso, deve ser aplicado com uma probabilidade maior que a taxa de mutação. Existem alguns tipos de cruzamento que são comumente utilizados, conforme Back et al. [2018]. São eles:

- **Ponto único** (do inglês, *'One-Point crossover'*): é o procedimento de definir um determinado "ponto de corte" no cromossomo para efetuar o cruzamento e os genes serem permutados entre os indivíduos;
- **Uniforme** (do inglês, *'Uniform crossover'*): é o procedimento que determina, através de um parâmetro global, a probabilidade de cada gene ser permutado;
- **Multipontos** (do inglês, *'Punctuated crossover'*): é uma generalização do procedimento de cruzamento do único ponto, o qual é aplicado em mais de um "ponto de corte" no cromossomo.

2. **Mutação**: A operação de mutação é necessária para manter a diversidade genética da população, pois altera arbitrariamente um ou mais componentes do conjunto de genes escolhido, fornecendo meios para inserção de novos indivíduos na população. Com isto, a mutação assegura que a probabilidade de se chegar a qualquer parte do conjunto de solução nunca será zero [Vasconcelos et al., 2001]. Segundo Wright [1991], a técnica de mutação deve ser aplicada com uma taxa de mutação pequena.

A abordagem de cruzamento de 'Ponto único' foi primeiramente proposta por De Jong [1975]. Neste trabalho optou-se por essa abordagem devido a simplicidade de implementação.

5.2.3 Seleção de indivíduos

Os indivíduos são selecionados de uma população para gerarem descendentes. Dessa forma, é imprescindível efetuar uma escolha adequada, a fim de que tais descendentes sejam mais aptos que as gerações anteriores. Há diversos métodos de seleção de indivíduos apresentados pela literatura. Um aparato dessas várias estratégias e técnicas para selecionar indivíduos em AG pode ser visto em Sivaraj and Ravichandran [2011].

Neste trabalho a estratégia de seleção desenvolvida é baseada no método de torneio [Goldberg et al., 1990]. Essa estratégia é bastante difundida, pois oferece a vantagem de não precisar que uma comparação entre todos os indivíduos da população seja feita [O'Neill et al., 2010]. Além do mais, o método de torneio é preferível quando comparado a outros métodos, principalmente, devido a sua capacidade de paralelização e minimização de viés estatístico [Fleming and Purshouse, 2002].

5.2.4 Função de aptidão

Função de aptidão, também conhecida como função de avaliação (em inglês, *fitness function*), é um tipo particular de função objetivo usada para identificar quão próximo um determinado indivíduo solução está apto a atingir os objetivos definidos. A função de aptidão é normalmente usada em AG para guiar as simulações, representando uma evolução do conjunto de indivíduos retornados da função de seleção, os quais devem conter resultados cada vez mais próximos da solução [Hruschka et al., 2009].

Após cada ciclo de simulação, descartam-se os piores indivíduos e criam-se novos. Cada indivíduo precisa atingir um limiar, para indicar o seu alcance perante à especificação geral, avaliado pela função de aptidão em cada geração [Zhou et al., 2011].

5.3- Árvore de Decisão

Árvore de Decisão (AD) é um método de aprendizagem supervisionada muito usado em tarefas de classificação e regressão, com o objetivo de criar um modelo capaz de prever o valor de uma variável alvo, baseado nas regras de decisão inferidas a partir dos dados [Breiman, 2017].

Uma AD tem a estrutura semelhante a de um fluxograma, no qual cada nó interno representa um teste em um atributo, cada ramificação significa um resultado do teste e o rótulo de cada classe é apresentado como um nó folha. Ou seja, dado um conjunto de atributos $Dados$, os valores desses atributos são testados na árvore de decisão e, como

resultado, um caminho é traçado desde o nó raiz até os nós folhas com a previsão de classe referente aquele conjunto. Um exemplo de AD gerada para um conjunto de dados pode ser visto na Figura 14.

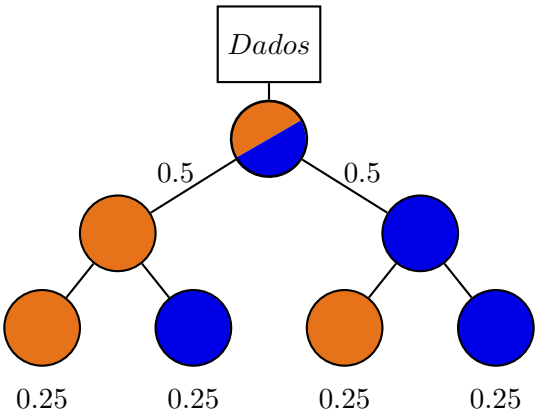


Figura 14 – Representação de uma Árvore de Decisão

O método de AD usa um modelo preditivo hierárquico que mapeia observações sobre um determinado item, objetivando inferir conclusões sobre o valor alvo por meio das inferências obtidas. É uma das abordagens de modelagem preditiva mais usadas em Estatística, Mineração de Dados (MD) e AM. Os modelos de AD em que a variável de destino possui um conjunto finito de valores são chamados de 'Árvore de Classificação". Nessa estrutura de AD para classificação, as folhas representam rótulos de classe e, as ramificações, conjunções de atributos que levam a esses rótulos de classes [Sharma and Kumar, 2016].

A classificação de AD significa uma forma de particionamento recursivo da amostra da instância. Neste tipo de AD cada nó interno divide a amostra em dois ou mais subespaços, de acordo com uma determinada função discreta aplicada aos valores dos atributos de entrada, onde, para cada folha, é atribuída uma classe que representa o valor alvo mais apropriado [Rokach and Maimon, 2005].

Aplicações com AD têm sido muito utilizadas por diversas razões e algumas delas são [Peng et al., 2009]:

- Boa generalização para instâncias ainda não avaliadas e que correspondem ao modelo treinado;
- Método eficiente computacionalmente e proporcional ao número de instâncias de treinamento observadas;

- Resultado interpretável de maneira compreensível ao humano, tornando o processo de classificação mais evidente.

O algoritmo de AD tem sido usado com sucesso no processo de extração de conhecimento. Dado que a tarefa frequente realizada nesse contexto é o uso de métodos indutivos para identificar padrões, a fim de expressar a classificação deste objeto de acordo com as regras geradas pela AD [Fayyad et al., 1996]. As implementações de AD mais conhecidas, conforme Singh and Gupta [2014] são: *Iterative Dichotomiser 3 (ID3), C4.5, CART, Chi-square Automatic Interaction Detection (CHAID), QUEST, GUIDE, CRUISE e CTREE*.

5.3.1 Critérios de Separação

Os algoritmos de AD requerem, fundamentalmente, critérios para dividir um nó até formar uma árvore [Breiman, 1996; Drummond and Holte, 2000]. Na maioria dos casos, as funções de divisão são univariadas. Neste contexto, o termo "univariado" significa que um nó interno é dividido conforme o valor de um único atributo, em que o algoritmo utilizado procura o melhor atributo para ser particionado [Rokach and Maimon, 2005]. O principal objetivo dessa atividade de separação é determinar o particionamento mais adequado dos dados. Os critérios de separação mais usados no particionamento de AD são[Tan et al., 2018]:

- **Entropia**: consiste em uma medida que visa identificar a importância dos atributos, avaliando a influência de um atributo em relação à tarefa de predição [Zhang et al., 2016; Boonchuay et al., 2017];
- **Índice de Gini**: consiste em uma medida de divergência para as distribuições de probabilidade dos valores do atributo alvo [Breiman, 2017; Rutkowski et al., 2014; Timofeev, 2004];
- **Erro de classificação**: é uma medida que mostra a fração de instâncias pertencentes à classe i em um determinado nó t [Rutkowski et al., 2015];

- **Ganho de informação**: é um critério baseado na diferença entre a entropia do nó pai e a entropia dos seus nós filhos [Quinlan, 1986];
- **Taxa de ganho**: é um critério que normaliza o ganho de informação, atendendo em casos de atributos que possuem valores numéricos distintos e não apresentam bom desempenho nos critérios de entropia e índice de Gini [Quinlan, 1993; Rokach and Maimon, 2005];
- **Critério de *Twoing***: é baseado em um particionamento binário, definido por $p(i|t)$ que mostra a fração de instâncias pertencentes à classe i de um determinado nó t [Zambon et al., 2006].

5.3.2 O Algoritmo CART

O algoritmo CART (do inglês, *Classification and Regression Tree*), proposto por Breiman et al. [1984], inicia-se com um único nó raiz L_0. Durante o processo de aprendizado, para cada nó L_q criado, um subconjunto particular S_q do conjunto de dados de treinamento S é processado, fazendo com que o nó raiz S_0 seja inicializado com S. Se todos os elementos do conjunto S_q pertencerem à mesma classe, o nó será marcado como folha e a divisão não será feita. Caso contrário, conforme o critério de separação, o atributo divisor é escolhido entre os atributos disponíveis a partir do nó selecionado. Para cada atributo disponível a^i, um conjunto de atributos A^i é particionado em dois subconjuntos disjuntos A^i_L e A^i_R onde ($A^i = A^i_L \cup A^i_R$).

A escolha do subconjunto A^i_L determina automaticamente o subconjunto complementar A^i_R e, portanto, a partição é representada apenas por A^i_L. O conjunto de todas as partições possíveis de A^i é denotado pelo seu conjunto de valores V_i, onde os subconjuntos A^i_L (esquerdo) e A^i_R (direito) dividem-se do conjunto de dados S_q [Rutkowski et al., 2014]. Dessa forma, as equações (1) e (2) representam, respectivamente, os dois subconjuntos originados:

$$L_q(A^i_L) = \{s_j \in S_q | V^i_j \in A^i_L\}, \quad (1)$$

$$R_q(A_L^i) = \{s_j \in S_q | V_j^i \in A_R^i\}. \quad (2)$$

Os conjuntos $L_q(A_L^i)$ e $R_q(A_L^i)$ dependem do atributo escolhido e da partição fracionada de seus valores. A partição de todos os elementos de dados S_q no nó considerado L_q, de uma classe k, é denotada por $p_{kL,q}$. Note que $p_{kR,q}$, $k = 1, ..., K$ são não dependentes do atributo escolhido a^i e da partição A_L^i. Desse modo, para qualquer subconjunto S_q do conjunto de dados de treinamento, denota-se o "Índice de Gini" através de (3):

$$Gini(S_q) = 1 - \sum_{k=1}^{K} (pk, q)^2. \quad (3)$$

Por meio da Equação (3), o "índice de Gini" atinge seu ponto de mínimo, onde $Gini(S_q) = 0$, quando todas as instâncias se representam em uma única classe e o ponto de máximo quando as instâncias são igualmente distribuídas entre todas as classes.

Os principais algoritmos para construção de AD diferem-se em duas características: tipo de árvore (binária ou não binária) e tipo de medida de impureza (*e.g.* entropia, índice de gini, taxa de ganho) [Quinlan, 2014]. Tais características do CART demonstram que ele se adequa melhor ao problema aqui abordado [Singh and Gupta, 2014; Sharma and Kumar, 2016; Breiman, 2017; Timofeev, 2004]. Para compreender em mais detalhes o processo de construção do CART consulte o trabalho de Timofeev [2004].

5.4- Discussão

A ideia de combinar AG e AD não é uma novidade na área de ciência de dados, inúmeros trabalhos realizaram tal combinação em diferentes contextos [Carvalho, 2005; Basgalupp et al., 2009; Blomberg et al., 2014; Silva, 2015]. Recentemente, encontramos o trabalho [Farissi et al., 2020] que reforça a relevância da abordagem desenvolvida nesta

dissertação.

Farissi et al. [2020] propõem um método baseado na seleção de atributos de AG com uma técnica de classificação para prever o desempenho acadêmico de um aluno. Segundo os autores, AG é usado em conjunto com RF para melhorar o resultado de predição e reduzir a dimensionalidade nos dados de classes desbalanceadas. Apesar das semelhanças, o presente trabalho além de utilizar uma abordagem baseada em AD, apresenta o desenvolvimento de um EDW que integra diferentes sistemas da universidade. Além disso, a base de dados utilizada nesta dissertação contem cerca de 80 mil registros de alunos, enquanto os autores daquele trabalho utilizam uma base de dados do repositório Kaggle utilizando contendo cerca de 480 registros.

Ainda sobre a redução de dimensionalidade, é imprescindível compreender que essa questão pode ser crucial para o problema. Observando-a sobre o enfoque da área de otimização, nota-se que o espaço de busca do problema cresce exponencialmente em relação ao número de atributos, o que poderia tornar inviável o desenvolvimento de um modelo [Huang, 2003; Aggarwal, 2001].

Este trabalho considerou apenas algumas dezenas de atributos a serem avaliados. No entanto, o conjunto de dados fornecido pelo EDW poderá alcançar a ordem de milhares de atributos, bastando que sejam adicionadas as informações de notas de desempenho e disciplinas de cada um dos 12969 alunos avaliados no contexto da evasão. Neste caso, não existiriam limitações para a utilização do *EvolveDTree*, não tendo sido realizado até o momento devido a limitações de tempo e escopo da dissertação.

Algoritmicamente, o *EvolveDTree* funciona conforme descrito a seguir. No primeiro momento, o AG recebe os dados particionados, a partir do conjunto total de atributos, onde cada indivíduo é representado por uma combinação aleatória desses atributos, compondo a população inicial de 100 indivíduos. Em seguida, estes indivíduos são submetidos à função de aptidão, composta por uma AD que retorna o *F-score* de cada indivíduo. Esse procedimento é executado usando os n atributos repetidamente. Cada iteração k produz uma população de indivíduos, com um desempenho e uma taxa de erro $e(k)$ que, durante essa etapa de seleção de atributos, descarta os indivíduos com desempenho menor que $0,75$. Ao final desse processo, os melhores indivíduos selecionados pelo AG são obtidos, representando o melhor subconjunto de atributos a serem considerados para o desenvolvimento do modelo preditivo. A segunda fase do *EvolveDTree* consiste no componente de predição. Para isso é empregado o método CART [Breiman et al., 1984],

uma das técnicas mais utilizadas de AD [Singh and Gupta, 2014]. Este método consegue realizar a tarefa de definir quais estudantes tendem a evadir ou concluir o curso.

As implementações das técnicas empregadas são provenientes da biblioteca *Sklearn* [Pedregosa et al., 2011]. Todas as implementações, desde as etapas de Análise Exploratória, Seleção de Atributos com AG e o procedimento de validação cruzada estratificado (com $k = 10$) foram também desenvolvidas em linguagem Python [1], compondo o sistema de predição do *EvolveDTree*.

[1] https://www.python.org/

6- Cenários Avaliados

Neste capítulo são detalhados os cenários avaliados nas simulações experimentais para selecionar um modelo de predição dentre oito algoritmos de classificação. Esta fase experimental é analisada sob a perspectiva das métricas de avaliação. Para apresentar essa etapa, a Seção 6.1 descreve os cenários experimentados, enquanto o detalhamento destes é realizado na Seção 6.2.

6.1- Descrição dos Cenários

Os cenários avaliados nesta pesquisa compreendem:

1. **Cenário Convencional**: realiza o treinamento com todos os atributos aplicados aos classificadores e o de melhor desempenho é selecionado, conforme o resultado das métricas observadas;
2. **Cenário Evolutivo**: efetua pré-processamento dos atributos selecionando-os por meio do Algoritmo Genético (AG) e posteriormente enviando somente os atributos selecionados aos classificadores. Ao final, estes classificadores serão avaliados com base no desempenho;
3. **Cenário Guloso**: apresentou uma implementação comparativa à redução de dimensionalidade, usando um processo iterativo em uma abordagem gulosa para encontrar o melhor subconjunto de atributos.

6.2- Apresentação dos Cenários de Avaliação

Durante esta seção, cada cenário será detalhado objetivando explicar, de forma reprodutível, o processo envolvido: recursos computacionais do ambiente, tratamento

da amostra e particionamento dos dados. Em todos os cenários, foi utilizada validação cruzada estratificada em dez partes (do inglês, *ten-fold stratified cross-validation*).

Todo código produzido nessa dissertação foi desenvolvido no ambiente *Google Cloud Platform* [1]. O *Google Cloud* é uma plataforma disponível por meio de computação em nuvem, fornecendo serviços de infraestrutura computacional. Nesta plataforma do Google ®, utilizou-se como ambiente de simulação, o *Google Colab* [2], ambiente para execução de código Python [3] em formato Jupyter [4]. Em relação às características do ambiente computacional utilizado para as simulações, o quadro a seguir, descreve os recursos utilizados.

Processador GPU: 1 Tesla P100-PCIE-16GB NVidia CUDA, 16GB GDDR5

Processador CPU: 1 *Single Core Hyper Threaded* (2 threads) Processador Xeon @$2,3$Ghz (Sem Turbo Boost) 45MB Cache

Memória RAM: $12,6$ GB

Espaço de armazenamento: 320 GB

Todos os cenários apresentarão os resultados obtidos durante as execuções efetuadas. Cada cenário será composto por tabelas e gráficos que exibem os resultados obtidos nas fases de treinamento e teste. Para a etapa de treinamento, foram disponibilizadas 8441 amostras, as quais representam o conjunto de alunos ingressantes dos anos de 2012 e 2013.

6.2.1 Cenário convencional

Conforme pode ser visto na Tabela 8, os resultados de treinamento destacam o desempenho dos modelos utilizando a amostra de treinamento com todos os atributos. Todas as técnicas tiveram desempenho acima de $0,75$ nas métricas avaliadas.

[1] https://cloud.google.com
[2] https://colab.research.google.com
[3] https://www.python.org/
[4] https://jupyter.org/

Tabela 8 – Resultados da Etapa de Treino no Cenário convencional

Algoritmo	F-Score	Mcc	Prec	Roc	Acc	Kappa
ArvoreDecisao	0,9875	0,9674	0,9876	0,9631	0,9858	0,9671
AdaBoost	0,9812	0,9727	0,9893	0,9703	0,9882	0,9726
SVM	0,9756	0,9645	0,9857	0,9630	0,9846	0,9644
RegLogistica	0,9740	0,9623	0,9869	0,9527	0,9834	0,9618
KNN	0,9740	0,9623	0,9869	0,9527	0,9834	0,9618
RandomForest	0,9736	0,9616	0,9827	0,9663	0,9834	0,9615
RedeNeural	0,9632	0,9473	0,9691	0,9803	0,9775	0,9470
NaiveBayes	0,8383	0,7620	0,8971	0,7724	0,8898	0,7556

Nesta fase de treinamento, o algoritmo que alcançou o melhor desempenho de *F-Score* foi AD com $0,9875$, seguido pelo AB com $0,9812$. Além das métricas de avaliação dos classificadores, foi observado também o tempo de execução durante a fase de treinamento, sendo NB ($0,01$ segundos), KNN ($0,10$ segundos) e AD ($0,11$ segundos) os mais rápidos e os com maiores tempo de execução foram RN ($9,78$ segundos) e SVM ($88,71$ segundos).

Tabela 9 – Relatório da Etapa de Teste no Cenário convencional usando AD

Classe	Precision	Recall	F-Score	Support
Evadido	1,00	0,99	0,99	581
Graduado	0,98	0,99	0,98	263

Tabela 10 – Matriz de Confusão do Cenário convencional usando AD

		Valor Predito	
		Evadido	**Graduado**
Valor Real	**Evadido**	*575*	*6*
	Graduado	*2*	*261*

A Tabela 9 apresenta os resultados da fase de teste para o modelo de predição da AD, onde foi obtido desempenho de *F-Score* acima de $0,98$. Esse desempenho pode ser melhor observado na Tabela 10, dado que, na atividade de prever os alunos evadidos, o modelo errou 6 dentre 581 evadidos e, para alunos graduados, o modelo classificou corretamente 261.

6.2.2 Cenário evolutivo

Neste cenário, o treinamento foi aplicado aos modelos com base nos dados referentes ao subconjunto de atributos selecionados pelo AG. A configuração do AG é definida pelos parâmetros propostos por Rainville Fortin et al. [2012], os quais estão descritos na Tabela 11.

Tabela 11 – Descrição dos Parâmetros do AG

Parâmetros:	**pop** – 100 – quantidade de indivíduos **cxpb** – $0,5$ – taxa de cruzamento **mutpb** – $0,2$ – taxa de mutação **ngen** – 10 – número de gerações **train** – $XTrain$ – dados de treino **test** – $Xtest$ – dados de teste
Retorno:	melhores indivíduos de cada geração

É importante destacar que nos atributos analisados nesta pesquisa, o conjunto inicial é representado por 27 características. Porém, após aplicada uma transformação nos dados categóricos (*One-Hot Encoding*), esse conjunto de atributos chegou um novo conjunto contendo 125 atributos. Especificamente neste cenário, o AG foi aplicados nesses dois conjuntos, sendo apresentados aqui, a listagem dos atributos que foram selecionados do conjunto inicial:

- ‘ACAOAFIRMATIVA’;
- ‘ENEMNATURAIS’;
- ‘ENEMREDACAO’;
- ‘CURSO’;
- ‘CODTURNOATUAL’;
- ‘TURNOATUAL’;
- ‘CR’;
- ‘SEMESTREINGRESSO’;

- ‘IDADE’;
- ‘COR’;
- ‘MOBILIDADE’;
- ‘CHCURSADA’;
- ‘ESTADOCIVIL’;
- ‘SEXO’.

E os atributos descartados pelo AG foram: ‘IDALUNO’, ’ENEMLINGUAGEM’, ’ENEMHUMANAS’, ’ENEMMATEMATICA’, ’CODTURNOINGRESSO’, ’ANOINGRESSO’, ’ANODESVINCULACAO’, ’SEMESTREDESVINCULACAO’, ’BAIRRO’, ’CEP’, ’CIDADE’, ’TRANCAMENTOS’, ’TEMPOPERMANENCIA’. Para o o segundo conjunto contendo 125 atributos foram selecionados 64 atributos (veja o Apêndice C). Após reduzida a dimensionalidade do conjunto de dados, os classificadores recebem esses dados como entrada e inicia-se a fase de treinamento. A avaliação desses classificadores e suas métricas são apresentados na Tabela 12.

Tabela 12 – Resultados da Etapa de Treino no Cenário evolutivo

Algoritmo	F-Score	Mcc	Roc	Prec	Acc	Kappa
EvolveDTree	0,9981	0,9972	0,9991	0,9962	0,9988	0,9972
AdaBoost	0,9943	0,9917	0,9964	0,9924	0,9964	0,9917
KNN	0,9943	0,9917	0,9964	0,9924	0,9964	0,9917
RegLogistica	0,9924	0,9890	0,9955	0,9886	0,9953	0,9890
SVM	0,9924	0,9890	0,9966	0,9850	0,9953	0,9890
RandomForest	0,9924	0,9889	0,9945	0,9924	0,9953	0,9889
RedeNeural	0,9667	0,9519	0,9835	0,9388	0,9787	0,9510
NaiveBayes	0,8799	0,8289	0,9369	0,7903	0,9159	0,8164

Diante dos resultados apresentados pela Tabela 12, os melhores desempenhos de *F-Score* foram obtidos pelos algoritmos *EvolveDTree*, AB e KNN, respectivamente, $0,9981$, $0,9943$ e $0,9943$. As técnicas do RL, SVM e RF também apresentaram desempenho acima de $0,99$ nos algoritmos avaliados. O desempenho de RN e NB foram, respectivamente, $0,9885$, $0,9866$ e $0,9794$. Neste cenário, observou-se também o tempo de execução durante o treinamento, sendo NB ($0,01$ segundos) e *EvolveDTree* ($0,05$ segundos) os mais rápidos e os de maior tempo foram RN ($3,53$ segundos) e SVM ($26,93$ segundos).

Conforme pode ser visto na Tabela 13, o melhor algoritmo na fase de treinamento e também na fase de teste foi AD, que obteve aproximadamente $1,0$ em todas as métricas: *Precision, Recall* e *F-Score*.

Tabela 13 – Relatório da Etapa de Teste no Cenário evolutivo AD

Classe	Precision	Recall	F-Score	Support
Evadido	1,00	1,00	1,00	582
Graduado	1,00	1,00	1,00	262

Tabela 14 – Matriz de Confusão do Cenário evolutivo AD

		Valor Predito	
		Evadido	**Graduado**
Valor Real	**Evadido**	*581*	*1*
	Graduado	*0*	*262*

Na Fase de Treinamento, os resultados obtidos pela AD e por AB foram os mais elevados, conforme apresentados na Tabela 13. Durante os testes, esses algoritmos obtiveram desempenho de *F-Score* acima de $0,99$ na classificação das amostras de alunos. Esse desempenho pode ser evidenciado na Tabela 14, dado que a atividade de classificar os alunos evadidos, o modelo fez corretamente 581 amostras de evadidos, errando apenas uma. No caso dos alunos graduados, o modelo identificou corretamente todos 262 alunos.

6.2.3 Cenário com Seleção de Atributos por Método Guloso

Neste cenário o treinamento foi aplicado aos modelos com base nos dados referentes ao subconjunto de atributos selecionados pelo Método Guloso (MG); abordagem conhecida em problemas de otimização [Kannan et al., 2018] e que também tem sido aplicada na redução de dimensionalidade e seleção de atributos [Huang, 2003; Farahat et al., 2013]. O MG descrito neste cenário, efetua um processo iterativo de remoção de atributos, conforme a ordem de entrada dos dados de treinamento, removendo até o penúltimo atributo.

Neste procedimento cada um dos atributos é removido conforme a sua ordem de entrada e o subconjunto restante é submetido aos algoritmos de classificação até que seja identificado um subconjunto ótimo de atributos com o melhor desempenho de *F-Score*. Cada subconjunto avaliado é particionado através do *ten-fold cross validation* realizado a cada novo atributo removido, repetindo-se $N - 1$ vezes, até o último atributo. Desse modo, atributos selecionados pelo MG foram:

- ‘ENEMLINGUAGEM’;
- ‘ENEMCIENCIAS’;
- ‘ENEMREDACAO’;
- ‘CODTURNOINGRESSO’;
- ‘CR’;
- ‘SEMESTREINGRESSO’;
- ‘SEMESTREDESVINCULACAO’;
- ‘IDADE’;
- ‘CHCURSADA’;
- ‘CEP’;
- ‘MOBILIDADE’;
- ‘ESTADOCIVIL’;
- ‘TEMPOPERMANENCIA’;
- ‘TRANCAMENTO’;
- ‘SEXO’.

E os atributos descartados pelo método foram: ‘IDALUNO’, ‘ACAOAFIRMATIVA’, ‘ENEMHUMANAS’, ‘ENEMMATEMATICA’, ‘CURSO’, ‘CODTURNOATUAL’, ‘TURNOATUAL’, ‘ANOINGRESSO’, ‘ANODESVINCULACAO’, ‘BAIRRO’, ‘CIDADE’ e ‘COR’.

Segundo a Tabela 15, os melhores desempenhos foram obtidos pelos algoritmos AB, AD e SVM, respectivamente, $0,9920$, $0,9912$ e $0,9904$. Para este cenário, o melhor algoritmo selecionado na fase de treinamento foi o AB.

Tabela 15 – Resultados da Etapa de Treino com Seleção de Atributos Gulosa

Algoritmo	F-Score	Mcc	Prec	Roc	Acc	Kappa
AdaBoost	0,9920	0,9895	0,9958	0,9888	0,9961	0,9895
ArvoreDecisao	0,9912	0,9885	0,9961	0,9857	0,9958	0,9884
SVM	0,9904	0,9874	0,9959	0,9842	0,9954	0,9874
KNN	0,9880	0,9843	0,9945	0,9810	0,9942	0,9842
RegLogistica	0,9865	0,9822	0,9940	0,9779	0,9934	0,9821
RedeNeural	0,9865	0,9823	0,9946	0,9764	0,9934	0,9822
NaiveBayes	0,9497	0,9348	0,9832	0,9042	0,9745	0,9327
RandomForest	0,8600	0,8350	0,8786	0,9916	0,9406	0,8231

Tabela 16 – Relatório da Etapa de Teste com Seleção de Atributos Gulosa usando AB

Classe	Precision	Recall	F-Score	Support
Evadido	0,98	0,98	0,98	582
Graduado	1,00	0,99	0,99	262

Tabela 17 – Matriz de Confusão do Cenário com Seleção de Atributos Gulosa usando AB

		Valor Predito	
		Evadido	**Graduado**
Valor Real	**Evadido**	*574*	*8*
	Graduado	*3*	*259*

Na Fase de Teste, os resultados obtidos por AB foram bons, conforme apresentados na Tabela 16. Durante os testes, o modelo AB obteve desempenho acima de $0,98$ para prever os alunos evadidos e graduados. Esse desempenho pode ser melhor observado na Tabela 17, dado que, na tarefa de predição, o modelo errou em 8 amostras, dentre 574 evadidos e em 3 amostras nos 263 graduados.

6.3- Discussão

Após avaliar as simulações dos três cenários apresentados, o melhor modelo selecionado foi produzido pelo Cenário evolutivo, resultando na abordagem AD com AG, denominada de *EvolveDTree*. Abordagens envolvendo métodos como o PCA e o NSGA-II para a redução de dimensionalidade deverão ser avaliadas em trabalhos futuros.

Os resultados obtidos pelo *EvolveDTree* são discutidos no Capítulo 7, junto com uma análise dos dados e observações estatísticas referente ao problema de evasão.

7- Discussão dos Resultados

O principal objetivo deste trabalho foi o desenvolvimento de uma abordagem que permitisse a gestão acadêmica integrar diversos outros sistemas com enfoque em classificar estudantes com risco de evasão, conforme descrito detalhadamente no Capítulo 4. Para o processo de classificação, o sistema de predição *EvolveDTree* foi desenvolvido. Este capítulo descreve os principais resultados obtidos pelo *EvolveDTree* no processo de classificação para o problema de evasão. A Seção 7.1 descreve uma análise com base no conjunto de dados oriundos do EDW desenvolvido, de maneira a alicerçar hipóteses sobre o problema de evasão na instituição analisada. A Seção 7.2 apresenta os resultados produzidos durante o processo de seleção de atributos, destacando as informações mais significativas para representar um aluno. Por fim, a Seção 7.3 discute o modelo de predição desenvolvido e a Seção 7.4 apresenta um estudo de caso com o *EvolveDTree* para prever a evasão avaliando os alunos de 2014 da Universidade Federal Fluminense.

7.1- A Análise dos Dados

Perante o objetivo de análise de dados, serão apresentadas algumas observações (*insights*) realizadas com este estudo. Como primeira observação, evidenciou-se que o conjunto de dados possui uma predominância de classe, o que promove um desbalanceamento dos dados. A amostra contém 76% de estudantes na classe “evadido” e 24% na classe “graduado”, conforme a Tabela 18. Este desequilíbrio na amostra tende a dificultar o processo de aprendizagem do modelo, podendo refletir em um viés estatístico e reduzindo a capacidade de generalização do modelo [Padmaja et al., 2007; Bhowan et al., 2011, 2012].

Tabela 18 – Mediana dos atributos baseada nos perfis “Evadido”e “Graduado”

Classe	Qtd.	Idade	SemestreFinal	CR	CargaHor	TempoPermanencia
Evadido (‘0’)	9852	25	1º	3,4	240	3 anos
Graduado (‘1’)	3117	24	2º	8,3	3199	5 anos

A Tabela 18 também apresenta detalhes que quantificam a amostra (“Qtd.”), exibindo para cada classe a mediana dos valores dos atributos: “Idade”, “SemestreFinal” (último semestre cursado), “CR” (Coeficiente de Rendimento), “CargaHor” (Carga horária cursada) e “TempoPermanencia”. Com base nos resultados, é possível observar que um padrão de evasão destaca-se com alunos abandonando em maior frequência no início do ano letivo com idade próxima a 25 anos, CR mediano de $3,4$ e carga horária cursada aproximada de 240 horas de currículo.

A partir das perspectivas anteriores, avaliou-se, também, a correlação entre os atributos gerados pelo conjunto de dados observado, conforme a Tabela 19. Essa tabela de correlações destaca que as notas do ENEM são os atributos com maior relação ao ‘CR’ e ao ‘StatusFormacao’.

Tabela 19 – Correlação dos atributos “CR” e “StatusFormacao”

Atributo	CR	StatusFormacao
EnemLinguagem	$0,141993$	$0,099262$
EnemHumanas	$0,094311$	$0,018340$
EnemCiencias	$0,103915$	$0,040897$
EnemMatematica	$0,067701$	$0,036235$
EnemRedacao	$0,139197$	$0,095648$
IdTurno	$0,076495$	$-0,014528$
IdTurnoAtual	$0,008767$	$-0,057659$
CR	$1,000000$	$0,633797$
AnoIngresso	$-0,09314$	$-0,208442$
SemestreIngresso	$-0,10859$	$-0,101707$
Idade	$-0,14578$	$-0,067144$
CargaHorCursada	$0,702870$	$0,901757$
Trancamento	$0,037985$	$-0,008887$
TempoPermanencia	$0,539021$	$0,480889$
StatusFormacao	$0,633797$	$1,000000$

Outra observação está relacionada com as distribuições dos alunos e sua etnia, conforme mostrado na Figura 15, destacando-se o valor mediano de CR para cada raça/cor. O gráfico desta figura ilustra que, para a maioria das etnias, o valor mediano do CR está entre 4 e 6. A exceção neste comportamento foi a etnia amarela, representada por 31 amostras de alunos. Todavia, ressaltamos que qualquer outra análise é mais complexa devido a maioria dos alunos não declararem sua raça/cor.

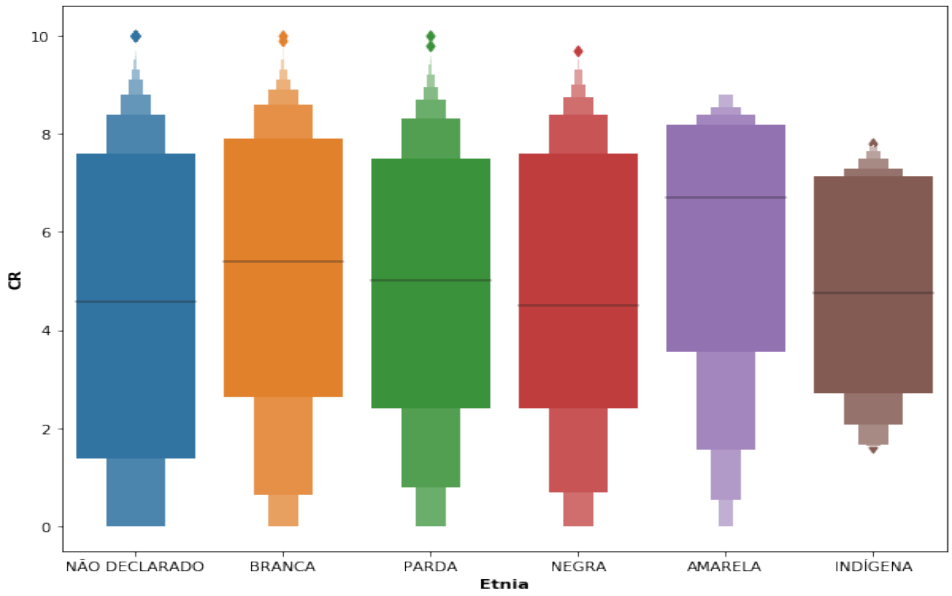


Figura 15 – Gráfico de *Boxplot* do CR em cada raça/cor da amostra

Baseada nesta análise de dados exploratória, foi possível concluir que a maior parte dos alunos evadidos (81%) tem idade entre 21 e 30 anos. Dentre todos os evadidos, 134 abandonaram os estudos com uma média de carga horária cursada de 2500 horas e 2192 tiveram CR mediano de $7,4$.

Pode-se observar também que a maior ocorrência de alunos evadidos aconteceu nos 1^{o} e 4^{o} anos dos cursos e que os homens evadem mais que as mulheres. Nesta amostra, também evidenciou-se que a maioria dos alunos evadidos foram matriculados no turno integral, ingressaram no ano de 2013, abandonaram o curso no 1^{o} semestre do ano letivo e tinham idade mediana de 23 anos.

7.2- O Aluno e seus Atributos – Uma Identificação Informativa

A Tabela 20 apresenta o comportamento do AG para a seleção de atributos durante a fase de treinamento. Nesta tabela a avaliação do AG é representada pelas seguintes colunas: “Número da geração” (**Geração**), “Quantidade de indivíduos” (**Indivíduos**), “ Média Ponderada de *F-Score*” (**Média**), “Desvio padrão de *F-Score*” (**Erro**), “Valor Mínimo de *F-Score*” (**Mínimo**) e “Valor Máximo de *F-Score*” (**Máximo**). Observe que o AG atingiu o desempenho de *F-Score* de $0,994536$ com apenas 6 gerações, mantendo este valor

estável até a 10^a geração.

Tabela 20 – Avaliação das gerações de indivíduos através de AG para selecionar o melhor conjunto de atributos

Geração	Indivíduos	Média	Erro	Mínimo	Máximo
0	100	0,851333	0,184753	0,378788	0,991796
1	61	0,966764	0,065023	0,586081	0,991796
2	67	0,985620	0,023870	0,778598	0,991811
3	67	0,988532	0,010675	0,883721	0,992714
4	69	0,984829	0,026180	0,803371	0,993631
5	72	0,990325	0,001918	0,82617	0,993631
6	53	0,990957	0,001746	0,985348	0,994536
7	66	0,990350	0,009432	0,898148	0,994536
8	70	0,990786	0,010488	0,888067	0,994536
9	60	0,990588	0,018996	0,802188	0,994536
10	60	0,990721	0,019774	0,794521	0,994536

Segue, abaixo, o subconjunto ótimo obtido como resultado do AG na seleção de atributos do conjunto inicial:

- ‘ACAOAFIRMATIVA’;
- ‘ENEMNATURAIS’;
- ‘ENEMREDACAO’;
- ‘CURSO’;
- ‘CODTURNOATUAL’;
- ‘TURNOATUAL’;
- ‘CR’;
- ‘SEMESTREINGRESSO’;
- ‘IDADE’;
- ‘COR’;
- ‘MOBILIDADE’;
- ‘CHCURSADA’;
- ‘ESTADOCIVIL’;

- ‘SEXO’.

Foram descartados pelo AG os seguintes atributos: ‘IDALUNO’, ’ENEMLINGUAGEM’, ’ENEMHUMANAS’, ’ENEMMATEMATICA’, ’CODTURNOINGRESSO’, ’ANOINGRESSO’, ’ANODESVINCULACAO’, ’SEMESTREDESVINCULACAO’, ’BAIRRO’, ’CEP’, ’CIDADE’, ’TRANCAMENTOS’, ’TEMPOPERMANENCIA’. Sobretudo, optou-se em avaliar o impacto de transformar os atributos de natureza categórica em atributos de natureza inteira. Para tanto os atributos: ACAOAFIRMATIVA, TURNOATUAL, COR e CURSO, foram selecionados e foram transformados por meio da técnica *One-Hot Encoding* [Cerda et al., 2018]. Dessa forma, o conjunto de dados passou a totalizar 125 atributos, dentre os quais o AG selecionou 64 com desempenho máximo de *F-Score* igual a $0,9945$. Assim, a abordagem usando AG alcançou melhores resultados para a fase preditiva, quando comparada com os outros dois cenários (convencional e guloso). O conjunto final dos atributos transformados está no Apêndice C.

A Figura 16 apresenta o desempenho obtido na avaliação do conjunto de teste para a geração de indivíduos nesta base de dados transformada. Neste gráfico é destacada a região dos melhores indivíduos gerados, a partir do *F-Score* de $0,85$, representando, aproximadamente, 80% da população. Desse modo, pode-se concluir que o processo de evolução do AG obteve sucesso na atividade de geração de indivíduos melhores para o conjunto solução, com desempenho de *F-Score* entre $0,90$ e $0,99$.

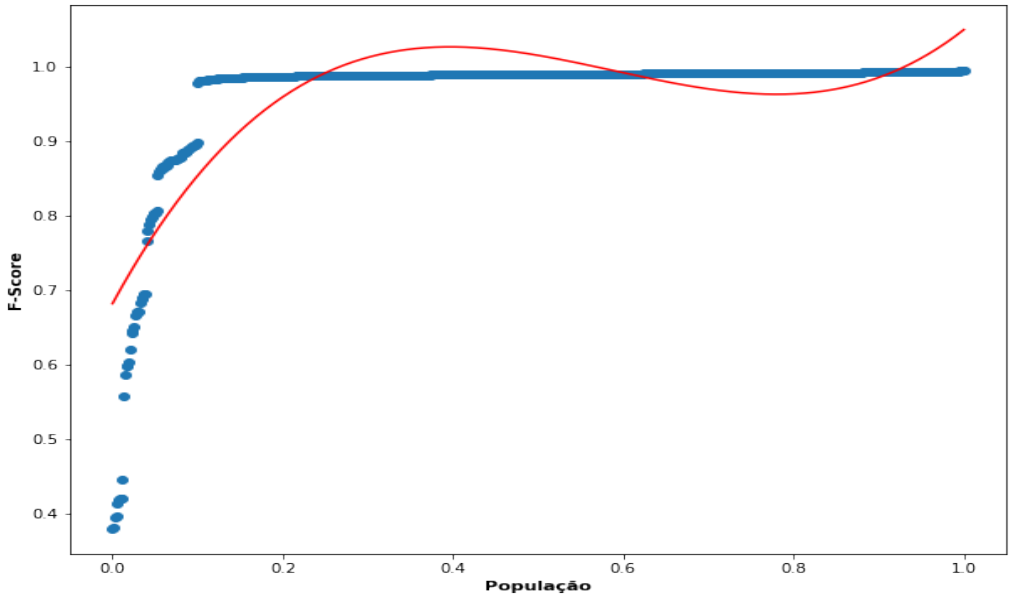


Figura 16 – Curva da *F-Score* no conjunto de validação

Para avaliar a relevância dos atributos na tarefa de predição, foi utilizado o método SHAP (do inglês *SHapley Additive exPlanations*). De acordo com [Lundberg and Lee, 2017], os valores SHAP visam: (i) explicar a predição de uma amostra analisando a influência de cada atributo nesta previsão, com base no valor de Shapley [Lipovetsky and Conklin, 2001]; e (ii) representar uma medida de importância de atributos.

A Figura 17 apresenta um gráfico da importância de cada atributo para o *EvolveD-Tree*. Ela ilustra que, para o modelo desenvolvido, a CHCURSADA e o CR são os dois atributos mais importantes na tarefa de predição.

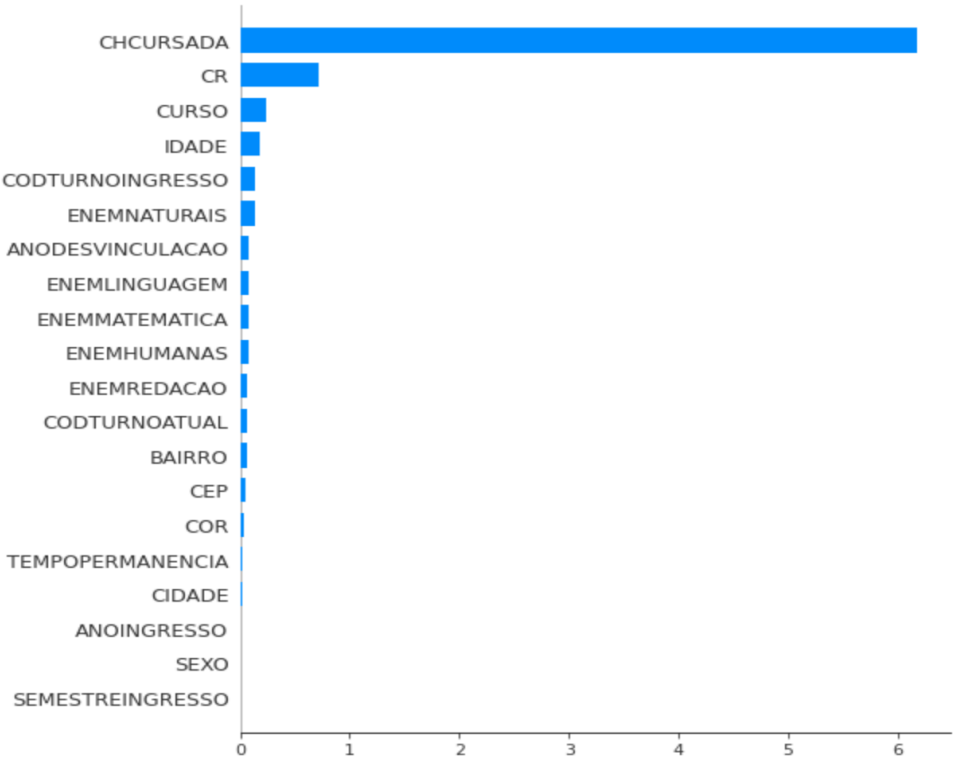


Figura 17 – Gráfico de importância dos atributos

A Figura 18 destaca o nível de impacto dos atributos no resultado de predição. Neste gráfico, também é possível observar as regiões de influência de cada atributo que mais impactam no modelo. Os de cor azul representam baixo impacto do atributo no alvo de predição; e, os de cor vermelha, alto impacto.

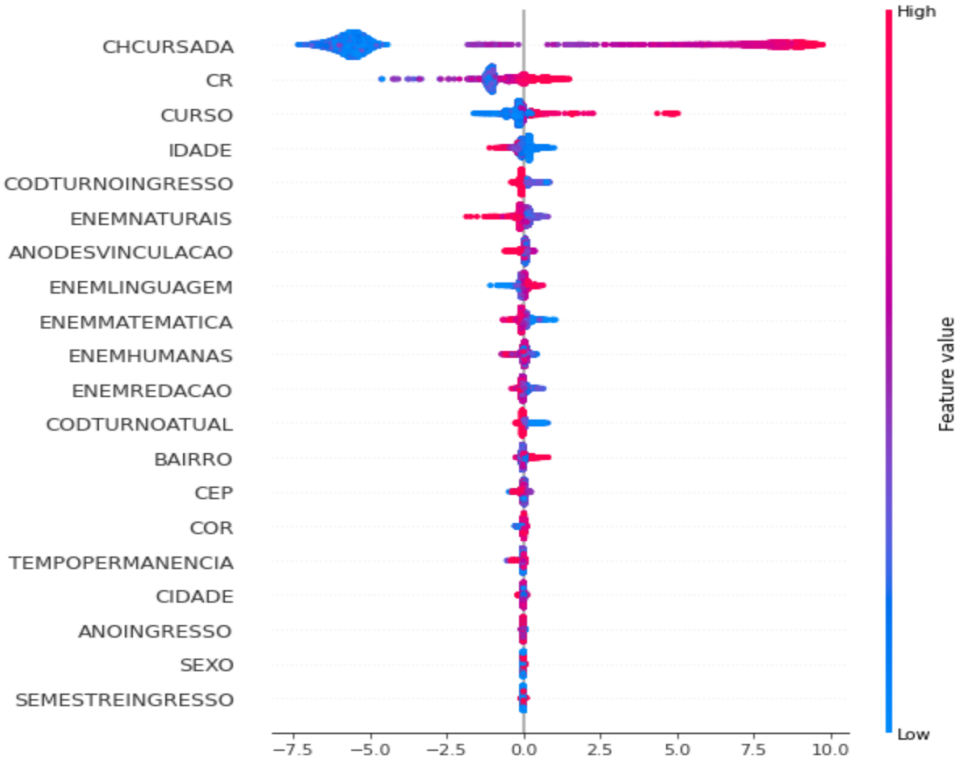


Figura 18 – Gráfico de impacto dos atributos no *EvolveDTree*

Diante disso, é possível concluir que, dos 20 atributos listados nos gráficos das Figuras 17 e 18, o AG identificou 9 destes, nos quais encontram-se os quatro mais importantes (CHCURSADA, CR, CURSO e IDADE).

Com o objetivo de validar o nível de importância desses atributos, o Método Guloso (MG), avaliado no Capítulo 6, apresentou uma diminuição do *F-Score* de $0,99$ para $0,87$ durante os experimentos sem o atributo CHCURSADA. Além disso, observou-se que, ao avaliar os algoritmos de classificação no conjunto de atributos, excluindo CHCURSADA e CR, o *F-Score* reduziu ainda mais $(0,68)$. Isso reforça a importância dos atributos, evidenciada nos gráficos das Figuras 17 e 18.

Por outro lado, é válido enfatizar que os atributos CHCURSADA e CR juntos representam um valor informacional agregado; pois, à medida que o valor de carga horária cursada aumenta, a probabilidade de o aluno graduar eleva. Isto será melhor detalhado na Seção 7.4, na qual é descrita a avaliação do *EvolveDTree* na predição dos alunos evadidos até o ano de 2014, comparando os resultados ao utilizar (ou não) o atributo CHCURSADA.

7.3- Um Modelo de Predição para Evasão

Nesta seção são comparados os resultados de execução do *EvolveDTree* com os resultados produzidos por outras técnicas de classificação testadas.

A AD gerada pelo *EvolveDTree* é exibida na Figura 19. De acordo com os atributos apresentados pelos nós de decisão da AD, é possível concluir que os estudantes com uma CHCURSADA abaixo de 2324 são altamente propensos a evadir do curso. Essa figura ilustra o motivo deste trabalho optar pela AD perante as demais técnicas: a simplicidade de interpretação visual do modelo.

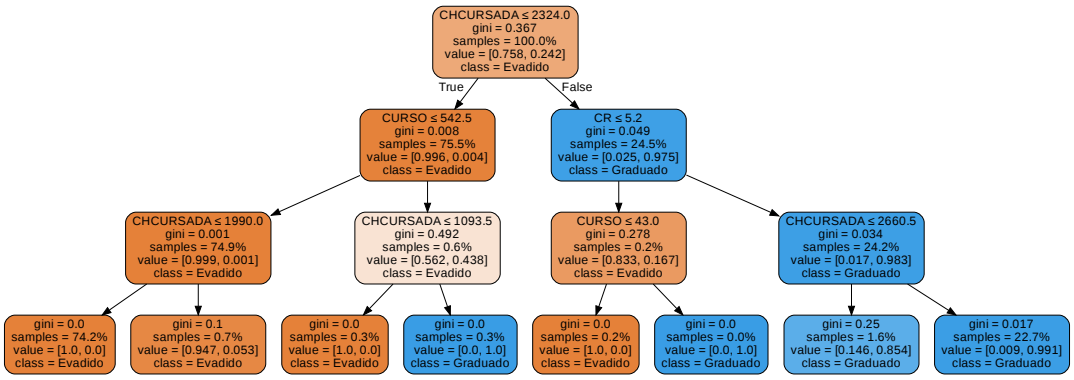


Figura 19 – Apresentação da Árvore de Decisão gerada pelo *EvolveDTree* com profundidade igual a 3

A interpretação visual permite outras observações, como por exemplo, o nó raiz particionar os dados em duas subárvores, de maneira semelhante à proporção das classes. Pode-se verificar, também, que o atributo CHCURSADA tem uma influência muito forte no particionamento da árvore de decisão, assim como o CR, pois, à medida com que as disciplinas do curso são concluídas, as chances do aluno graduar aumentam. A versão completa da saída gerada pelo *EvolveDTree* pode ser vista no Apêndice A.

A avaliação do modelo é realizada comparando o *EvolveDTree* com outras técnicas de AM, conforme as usadas em [Kotsiantis et al., 2007]. Para tanto, foram usadas o mesmo conjunto de métricas apresentadas em [Narudin et al., 2016].Os resultados

produzidos nessa etapa de avaliação são apresentados na Tabela 21. Estes valores se referem, respectivamente, a cada uma das seguintes métricas: “**F-Score**”, “Precisão” **(Prec)**, “Acurácia” **(Acc)**,“Coeficiente de Correlação de Matthews” **(Mcc)**, “Coeficiente Kappa” **(Kappa)** e “Curva ROC” **(Roc)**.

Tabela 21 – Resultados dos classificadores no conjunto de teste

Algoritmo	F-Score	Prec	Acc	Mcc	Kappa	Roc
AdaBoost	0,9789	0,9763	0,9912	0,9720	0,9954	0,9762
EvolveDTree	0,9770	0,9742	0,9927	0,9644	0,9949	0,9741
SVM	0,9750	0,9720	0,9916	0,9625	0,9945	0,9719
RegLogistica	0,9687	0,9649	0,9846	0,9639	0,9932	0,9649
KNN	0,9372	0,9298	0,9586	0,9519	0,9865	0,9296
RedeNeural	0,9339	0,9262	0,9556	0,9516	0,9859	0,9260
RandomForest	0,9338	0,9268	0,9486	0,9694	0,9861	0,9260
NaiveBayes	0,7023	0,6841	0,9223	0,5630	0,9139	0,6555

A Tabela 22 apresenta os resultados produzidos pelo modelo de predição durante a fase de teste, usando as métricas de “Precisão” (**Precisão**), “Sensibilidade” (**Sens.**), “F-Score” (**F-Score**), “Suporte ” (**Suporte**), “Verdadeiro Positivo” (**TP**) e “Falso Positivo” (**FP**). Os resultados detalham os valores das métricas para cada uma das classes *FinalStatus*: ‘**Evadido**’ e ‘**Graduado**’.

Tabela 22 – Avaliação do *EvolveDTree* no conjunto de teste

Classe	Precisão	Sens.	F-Score	Suporte	TP	FP
Evadido	1,00	1,00	1,00	4035	4017	18
Graduado	0,96	0,99	0,98	493	488	5

Ao diminuir o número de atributos, o número de possíveis configurações de árvores diferentes também é reduzido. Isso resulta em um modelo mais generalizado, minimizando a possibilidade de *overfitting*.

7.4- Prevendo a evasão dos alunos de 2014 com *EvolveDTree*

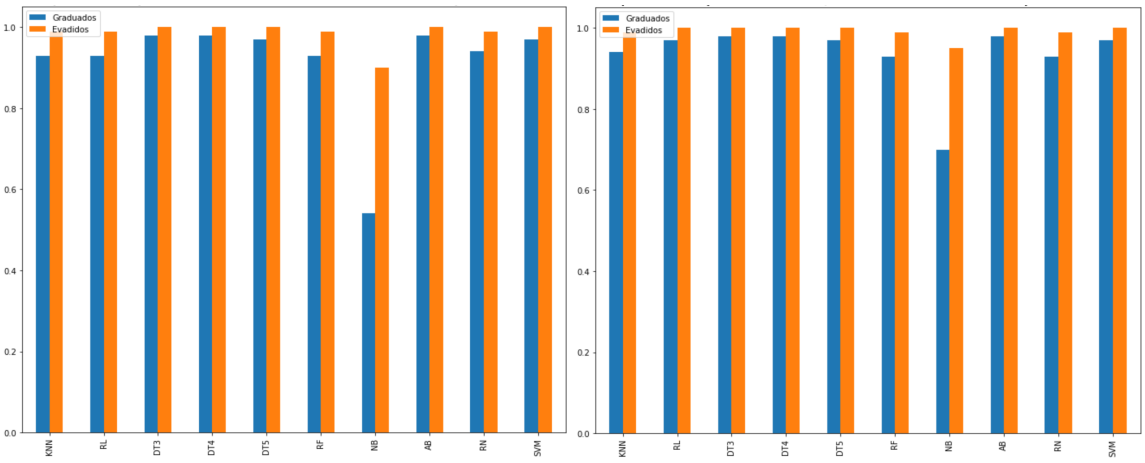
Esta seção destaca o estudo de caso no qual o *EvolveDTree* foi usado para prever os alunos ingressantes do ano de 2014, evadidos e graduados. Para isto, o *EvolveDTree* foi treinado com os alunos ingressos dos anos de 2012 e 2013. Esse conjunto de treinamento

conteve 8441 alunos; e, o conjunto de validação, 4528.

Essa avaliação visa verificar a capacidade de generalização do *EvolveDTree* em uma situação real. O critério observado nessa fase foi o valor de *F-Score* para a tarefa de predição dos alunos conforme sua classe (evadido ou graduado).

Durante os experimentos realizados buscou-se analisar os classificadores comparando a utilização (e não utilização) do AG e do atributo CHCURSADA. A Figura 20 apresenta esta avaliação com o atributo CHCURSADA. Já a Figura 21 apresenta os resultados sem o atributo CHCURSADA. Em ambos os casos, avaliamos também diferentes tipos de AD limitando as profundidades delas aos valores de 3(DT3), 4(DT4) e 5(DT5). Este teste objetivava avaliar a ocorrência de um possível *overfitting*.

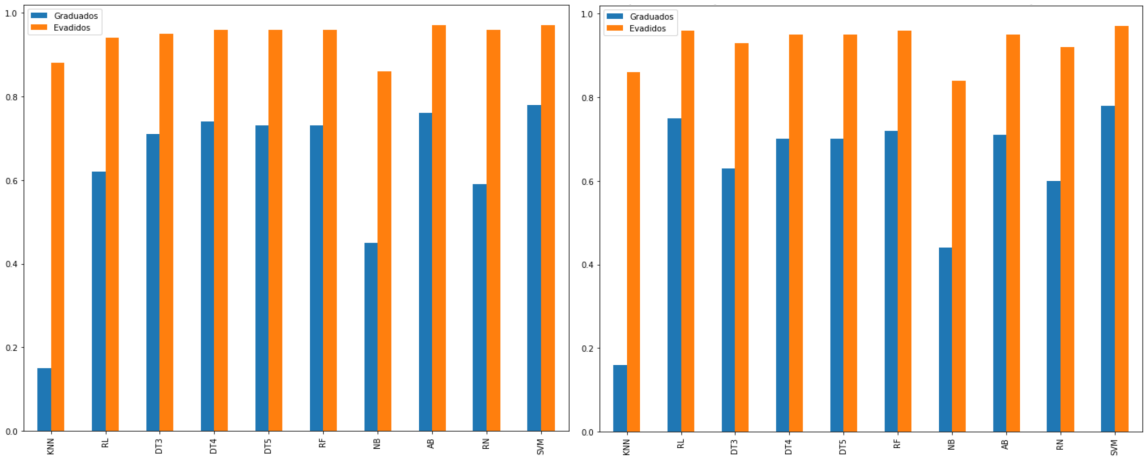
Quando observados os resultados das Figuras 20a e 20b percebe-se uma pequena melhora no valor do *F-Score* dos algoritmos RL e NB, quando aplicada a redução de dimensionalidade com AG. Todavia, para as AD a diferença de resultado foi quase insignificativa.



(a) Sem AG (b) Com AG

Figura 20 – Desempenho do *F-Score* dos classificadores avaliados para os alunos ingressantes no ano de 2014 com atributo CHCURSADA.

Na Figura 21 percebe-se uma melhora no valor de *F-Score* em alguns dos algoritmos (RL, NB, RN e SVM) quando aplicada a redução de dimensionalidade com AG nos experimentos. Por outro lado, as técnicas AD (DT3 – com profundidade 3) e AB apresentaram uma redução do F-Score. Os algoritmos que tiveram melhor resultado foram SVM e RL usando AG.



(a) Sem AG (b) Com AG

Figura 21 – Desempenho do *F-Score* dos classificadores avaliados para os alunos ingressantes no ano de 2014 sem atributo CHCURSADA.

Além do desempenho apresentado pelos algoritmos de classificação, observou-se, também, o tempo de execução na fase de treinamento. No experimento (a), foi utilizado o AG para redução de dimensionalidade, selecionando 63 entre 127 atributos, cujo tempo de execução levou em média 200 segundos, sendo o tempo total o tempo de treinamento de cada algoritmo somado ao tempo de execução do AG. No experimento (b), não foi feita redução de dimensionalidade, logo todos os atributos foram avaliados. Desse modo, o tempo total do experimento (a) foi de, aproximadamente, 539 segundos; e, o do experimento (b), 872 segundos.

Pode-se evidenciar que, detalhando o tempo de treinamento obtido, técnicas como RN e SVM tiveram maior proveito com a redução de dimensionalidade. Nos experimentos (a) e (b) da Figura 20, respectivamente, o tempo computacional desses algoritmos foi otimizado em, aproximadamente, 70% para RN e 300% para SVM. Já nos experimentos (a) e (b) da Figura 21, o comportamento se manteve, mas em proporções menores, 50% para RN e 200% para SVM.

Assim, como resultado final deste estudo de caso, o *EvolveDTree* foi o modelo escolhido e o atributo CHCURSADA seria removido do conjunto de dados. O descarte do atributo CHCURSADA aconteceu mediante o entendimento de que esse atributo representa um maior valor informacional. Essa compreensão é devida à informação quantitativa de créditos concluídos que esse atributo induz. Sob este ponto de vista, implicitamente, pode-se inferir um percentual de disciplinas aprovadas e de tempo de permanência de curso.

Para a escolha do *EvolveDTree*, o critério de facilidade na interpretação visual do resultado foi também relevante, uma vez que o especialista poderá ter maior autonomia na tomada de decisão. Além disso, na tarefa de prever os 4035 alunos evadidos de 2014, o *EvolveDTree* acertou 3675 e, na predição dos 493 alunos graduados, o número de acertos foi 464.

8- Considerações Finais

Esta pesquisa apresentou o desenvolvimento de um Data Warehouse Educacional (EDW) para a tarefa de mineração de dados educacionais, focando especificamente no problema de evasão no ensino superior. Para tanto, também foram desenvolvidos modelos de classificação baseados em diferentes técnicas de mineração de dados. Uma técnica de redução de dimensionalidade usando algoritmos genéticos também foi avaliada. O presente capítulo destaca os objetivos atingidos na Seção 8.1, os resultados alcançados na Seção 8.2 e os trabalhos futuros na Seção 8.3.

8.1- Objetivos Atingidos

Neste trabalho foi **efetuada uma análise dos dados** sobre uma amostra de um conjunto de dados de alunos da Universidade Federal Fluminense, no intuito de explorar os dados e validar possíveis caminhos para minimizar a evasão. Pode-se perceber, com esta análise, o desbalanceamento dos dados promovido pela diferença da proporção de alunos evadidos em relação aos graduados. Verificou-se, também, que há uma baixa qualidade de informação racial dos alunos, onde mais de 70% estão como "não declarados". Por outro lado, a amostra apresentou um quantitativo de mulheres e homens distribuídos de forma equilibrada (6514 mulheres para 6455 homens), o que é positivo para produção de novas análises. Por exemplo, nesta dissertação, observou-se que as mulheres graduaram mais que os homens e tiveram um melhor rendimento acadêmico. No entanto, também é importante, ter essa observação em nível de curso e sob outras perspectivas.

Após os resultados produzidos na análise de dados, percebeu-se a necessidade de identificar, nos 125 atributos da amostra, quais destes seriam relevantes para novas análises. Para isto, fez-se necessário **identificar os atributos mais representativos do aluno**, utilizando a técnica de redução de dimensionalidade. Um AG foi desenvolvido para isso, obtendo desempenho de $99,32$% com base na métrica de *F-Score*. Dentre

os atributos da amostra, 64 foram selecionados pelo AG, promovendo uma redução em $50,39$% do volume de dados.

Em seguida, o conjunto de dados selecionado foi analisado em cenários comparativos por oito algoritmos de classificação. Esta etapa resultou no *EvolveDTree*, um modelo de predição para **classificar alunos evadidos**, composto por um AG e uma AD. Na fase de testes, o *EvolveDTree* obteve o melhor desempenho dentre os demais algoritmos, classificando corretamente todos os 789 alunos evadidos e 249 nos 252 graduados.

Por fim, o *EvolveDTree* foi submetido à etapa de validação, que consistiu de uma observação sobre o conjunto de dados dos alunos ingressantes em 2014, evadidos e graduados, a fim de avaliar a capacidade de generalização do modelo para **prever alunos em risco de evasão**. Nesse conjunto de dados, ao analisar todos os atributos, o desempenho do *EvolveDTree* foi de $0,91$ de *F-Score* e de $0,73$ ao descartar o atributo CHCURSADA.

Como resultado dessa etapa de validação, é possível concluir que o *EvolveDTree* obteve resultados satisfatórios, destacando a capacidade de generalização do modelo. Pode-se observar, ainda, uma avaliação comparativa com e sem o atributo mais relevante (CHCURSADA) dentre o conjunto de dados, conforme identificado na Seção 7.2.

8.2- Resultados Alcançados

O *EvolveDTree* apresentou bom desempenho durante as fases de treinamento, teste e validação. Isto pode favorecer a IES desenvolver ações mais efetivas, no intuito de que os alunos sob o risco de evasão sejam observados de modo específico e, consequentemente, passem a ter um melhor direcionamento em sua vida acadêmica. Espera-se que com o desenvolvimento destas ações, a instituição possa reduzir o número de alunos evadidos e consequentemente ocorra uma redução nas despesas associadas ao problema da evasão.

Além do resultado apresentado por este modelo de predição, esta solução ainda pode ser estendida a outras instituições no Brasil e no mundo, desde que sejam feitos alguns ajustes. Para contribuir com isso, estão sendo discutidas ações para liberar o

conjunto de dados analisado neste trabalho no Portal de Dados Abertos da UFF[1]. O código de implementação da solução apresentada será disponibilizado em repositório público para todos os interessados, sob a condição de referenciá-lo quando utilizado. Além das contribuições já citadas, esta dissertação gerou os seguintes artigos:

- ERSI 2018 [2] – *"A Brief Review about Educational Data Mining applied to Predict Student's Dropout"*;
- SBBD 2019 [3] – "Data Warehouse Educacional: Uma visão sobre a Evasão no Ensino Superior";
- IWSSIP 2020 [4] – *"EvolveDTree: Analyzing Student Dropout in Universities"*.

Um novo artigo também está em processo de avaliação em uma conferência com foco em gestão educacional. Por fim, há mais um artigo em desenvolvimento, o qual deverá ser submetido a revista *Information Systems*.

8.3- Trabalhos Futuros

A partir dos resultados produzidos nesta pesquisa, foi possível obter alguns *insights* para o problema de evasão e, ao mesmo tempo, um reconhecimento de padrões nos alunos da UFF que abandonam ou graduam. Este tipo de suporte à tomada de decisão visa auxiliar a gestão acadêmica - atividade relevante e imprescindível.

Por outro lado, acredita-se que uma extensão deste trabalho, com o objetivo de também ajudar o aluno, possa ser muito útil. Para tanto, uma funcionalidade que seria capaz de direcionar na decisão dos alunos durante o processo de escolha de disciplinas se mostra interessante. Uma ferramenta de sistemas de recomendação seria importante, com ela seria possível indicar as disciplinas que melhor se adéquem ao aluno, tanto no quesito horário, quanto em um reforço na sua formação. Para atender essa funcionalidade, serão necessários alguns novos requisitos. São eles:

[1] http://dados.uff.br
[2] http://ersi2018.cefetfriburgo.com/ersi2018
[3] http://sbbd.org.br/2019/
[4] http://iwssip2020.ic.uff.br/

- Estender o conjunto de dados utilizado neste trabalho, buscando outras informações, principalmente disciplinas cursadas e o desempenho em cada uma;
- Desenvolver um conjunto de regras de associação capaz de identificar as relações e padrões frequentes de aprovação e reprovação dos alunos nas disciplinas dos cursos;
- Implementar uma interface apta a fornecer adequadamente as análises produzidas pelo sistema, de acordo com as práticas e políticas permitidas através da lei de proteção de dados pessoais.

Por conseguinte, para cumprir esta funcionalidade de recomendação, é necessário adicionar novas informações ao conjunto de dados, o que promoverá um aumento no conjunto de atributos e, possivelmente, uma maior necessidade de técnicas de engenharia de atributos (*feature engeneering*), principalmente para a criação de atributos na forma de métricas e indicadores.

É válido lembrar que o nível de precisão que buscamos conceder através desta solução pode estimular outras sub-atividades, como por exemplo, obter as informações das interações dos alunos nas atividades oferecidas em uma plataforma de conteúdo.

Sobretudo, em um cenário com esse tipo de informação, o conjunto de dados tende a se aproximar de um ambiente de *Big Data*. Isto promove mais um nível de complexidade, implicando em ajustes mais robustos, como por exemplo, aplicação de técnicas de aprendizado profundo (*Deep Learning*) e de tecnologias de armazenamento, as quais podem ser atendidas por banco de dados do tipo não relacional.

Levando-se em consideração os resultados alcançados, esta dissertação contribui para uma melhoria no processo de tomada de decisão da UFF, possivelmente auxiliando na redução da evasão dos alunos de cursos de graduação presencial.

Referências Bibliográficas

C. C. Aggarwal. On the effects of dimensionality reduction on high dimensional similarity search. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 256–266, 2001. 56

R. Ahuja and Y. Kankane. Predicting the probability of student's degree completion by using different data mining techniques. In *Image Information Processing (ICIIP), 2017 Fourth International Conference on*, pages 1–4. IEEE, 2017. 16, 30, 31, 32

M. Al-Barrak and M. Al-Razgan. Predicting students' performance through classification: A case study. *Journal of Theoretical & Applied Information Technology*, 75(2), 2015. 30

M. Alban and D. Mauricio. Neural networks to predict dropout at the universities. *International Journal of Machine Learning and Computing*, 9(2):149–153, 2019. 30, 31

J. B. d. Amaral. Evasão discente no ensino superior: estudo de caso no instituto federal de educação, ciência e tecnologia do ceará (campus sobral). Mestrado em políticas públicas e gestão da educação superior, Universidade Federal do Ceará, Fortaleza, CE, 2013. 20

ANDIFES. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. Technical report, Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras, 1996. 19

W. B. Andriola. Evasão discente na universidade federal do ceará (ufc): proposta para identificar suas causas e implantar um serviço de orientação e informação (soi). *Ensaio. Avaliação de Políticas Públicas em Educação, Rio de Janeiro,(11)*, 40:332–347, 2003. 21, 22

A. Aziz, N. Ismail, F. Ahmad, and H. Hassan. A framework for student's academic performance analysis using naïve bayes classifier. *Jurnal Teknologi (Sciences & Engineering)*, 75(3):13–19, 2015. 30

T. Back, D. B. Fogel, and Z. Michalewicz. *Evolutionary computation 1: Basic algorithms and operators*. CRC press, 2018. 50

M. Backenkohler, F. Scherzinger, A. Singla, and V. Wolf. Data-driven approach towards a personalized curriculum. *International Educational Data Mining Society*, 2018. 30

C. Baggi and D. Lopes. Dropout rates and institutional evaluation in higher education: a bibliographical discussion. (in portuguese). *Avaliação: Revista da Avaliação da Educação Superior*, 16(2), 2011. 14, 15, 20

R. Baker. Data mining for education. *International encyclopedia of education*, 7(3): 112–118, 2010. 15, 23, 27

R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2009. 19, 23, 27

R. Baker, S. Isotani, and A. Carvalho. Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03, 2011. 15

P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000. 46, 106

L. M. Barbosa Manhães, S. M. S. da Cruz, and G. Zimbrão. Towards automatic prediction of student performance in stem undergraduate degree programs. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 247–253. ACM, 2015. 30, 32

A. d. S. X. Barros. Expansão da educação superior no brasil: limites e possibilidades. *Educação e Sociedade*, 36(131):361–390, 2015. 14

M. P. Basgalupp, R. C. Barros, A. C. de Carvalho, A. A. Freitas, and D. D. Ruiz. Legal-tree: a lexicographic multi-objective genetic algorithm for decision tree induction. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1085–1090. ACM, 2009. 55

U. Bhowan, M. Johnston, and M. Zhang. Developing new fitness functions in genetic programming for classification with unbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):406–421, 2011. 66

U. Bhowan, M. Johnston, M. Zhang, and X. Yao. Evolving diverse ensembles using genetic programming for classification with unbalanced data. *IEEE Transactions on Evolutionary Computation*, 17(3):368–386, 2012. 66

L. C. Blomberg et al. Um algoritmo evolutivo para indução de árvores de regressão robusto a valores ausentes. *Pontifícia Universidade Católica do Rio Grande do Sul*, 2014. 55

K. Boonchuay, K. Sinapiromsaran, and C. Lursinsap. Decision tree induction based on minority entropy for the class imbalance problem. *Pattern Analysis and Applications*, 20 (3):769–782, 2017. 53

S. Boughorbel, F. Jarray, and M. El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6): e0177678, 2017. URL `https://journals.plos.org/plosone/Article/file?id=10.1371/journal.pone.0177678&type=printable`. 106

BRASIL. Instituto nacional de estudos e pesquisas educacionais anísio teixeira (inep). *Censo da Educação Superior 2018: notas estatísticas. Brasília, 2017*, 2017. Acessado em 6 de Dezembro de 2019. 15

BRASIL. Instituto nacional de estudos e pesquisas educacionais anísio teixeira (inep). *Censo da Educação Superior 2018: notas estatísticas. Brasília, 2018*, 2018. Acessado em 10 de Dezembro de 2019. 15, 16

BRASIL. Instituto nacional de estudos e pesquisas educacionais anísio teixeira (inep). *Censo da Educação Superior 2018: notas estatísticas. Brasília, 2019*, 2019. Acessado em 8 de Dezembro de 2019. 14

L. Breiman. Some properties of splitting criteria. *Machine Learning*, 24(1):41–47, 1996. 53

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 46

L. Breiman. *Classification and regression trees*. Routledge, 2017. 18, 51, 53, 55

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees (monterey, california: Wadsworth), 1984. 18, 46, 54, 56

C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 2017. 16, 30, 31

D. R. Carvalho. Árvore de decisão/algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados. *Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. Doctor Thesis. 162pp*, 2005. 55

P. Cerda, G. Varoquaux, and B. Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10):1477–1494, 2018. 47, 70

M. Chaturvedi. Data mining and it's application in edm domain. In *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on*, pages 829–834. IEEE, 2017. 30, 31

H. Chen, R. H. Chiang, and V. C. Storey. Business intelligence and analytics: from big data to big impact. *MIS quarterly*, pages 1165–1188, 2012. 23, 34

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 46

D. Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35, Dec 2017. ISSN 1756-0381. URL `https://doi.org/10.1186/s13040-017-0155-3`. 106

D. A. Cieslak and N. V. Chawla. Learning decision trees for unbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 241–256. Springer, 2008. 46

W. F. Clocksin and C. S. Mellish. *Programming in Prolog: Using the ISO standard*. Springer Science & Business Media, 2012. 33

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. URL `https://doi.org/10.1177/001316446002000104`. 107

J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968. 107

E. Costa, R. S. Baker, L. Amorim, J. Magalhães, and T. Marinho. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29, 2013. 23, 26, 30, 32

D. Couto and A. Santana. Educational data mining applied to the identification of variables associated with evasion and retention(in portuguese). In *CEUR Workshop Proceedings*, volume 1877, pages 333–344, 2017. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028023454&partnerID=40&md5=97e647372af4ec57654bbbba151a9b07`. cited By 0. 30, 32, 33

J. A. Cunha, E. Moura, and C. Analide. Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. In *New Advances in Information Systems and Technologies*, pages 189–198. Springer, 2016. 30, 32

R. Dalongaro, C. Ramos, and R. Azzolin. Estudo sobre a empregabilidade dos cursos de graduação no brasil. *URI (Org.), Anais do Encontro Missioneiro de Estudos Interdisciplinares em Cultura-EMiCult, São Luiz Gonzaga, RS, Brasil*, 2, 2016. 15

T. H. Davenport. Business intelligence and organizational decisions. In *Organizational Applications of Business Intelligence Management: Emerging Trends*, pages 1–12. IGI Global, 2012. 34

A. C. L. de Assis, M. T. Sanabio, C. A. Magaldi, and C. S. Machado. As políticas de assistência estudantil: experiências comparadas em universidades públicas brasileiras. *Revista Gestão Universitária na América Latina-GUAL*, 6(4):125–146, 2013. 15

K. A. De Jong. Analysis of the behavior of a class of genetic adaptive systems. Doctoral dissertation, University of Michigan, 1975. 50

M. DeBerard, G. Spielmans, and D. Julka. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal*, 38(1): 66–80, 2004. 15

N. Delavari, M. R. Beikzadeh, and S. Phon-Amnuaisuk. Application of enhanced analysis model for data mining processes in higher educational system. In *Information Technology Based Higher Education and Training, 2005. ITHET 2005. 6th International Conference on*, pages F4B–1. IEEE, 2005. 19, 35

D. Delen. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506, 2010. 31

N. Diamantidis, D. Karlis, and E. A. Giakoumakis. Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, 116(1-2):1–16, 2000. 47

E. C. Dias, C. R. Theóphilo, and M. A. Lopes. Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de ciências contábeis da universidade estadual de montes claros–unimontes–mg. In *Congresso USP de Iniciação Científica em Contabilidade, São Paulo, SP*, volume 7, 2010. 20

M. Doshi and S. Chaturvedi. Survey of feature selection algorithms in higher education. *International Journal of Computer Applications in Engineering Sciences*, 4(1):5, 2014. 30

C. Drummond and R. C. Holte. Exploiting the cost (in) sensitivity of decision tree splitting criteria. In *ICML*, volume 1, 2000. 53

A. Dutt, M. Ismail, and T. Herawan. A systematic review on education data mining. *IEEE Access*, 5(1):15991–16005, 2017. 27

K. Facelli, A. Lorena, J. Gama, and A. de Carvalho. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC — Livros Técnicos e Científicos Editora Ltda., 2015. ISBN 978-85-216-1880-5. 101

A. K. Farahat, A. Ghodsi, and M. S. Kamel. Efficient greedy feature selection for unsupervised learning. *Knowledge and information systems*, 35(2):285–310, 2013. 63

A. Farissi, H. M. Dahlan, and Samsuryadi. Genetic algorithm based feature selection with ensemble methods for student academic performance prediction. *Journal of Physics: Conference Series*, 1500:012110, apr 2020. URL `https://doi.org/10.1088/1742-6596/1500/1/012110`. 46, 47, 55, 56

T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 105

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996. 53

P. J. Fleming and R. C. Purshouse. Evolutionary algorithms in control systems engineering: a survey. *Control engineering practice*, 10(11):1223–1241, 2002. 51

U. F. Fluminense. *IdUFF - Sistema de Identificação Única da Universidade Federal Fluminense*. Disponível em: `https://inscricao.id.uff.br`, 2014. 37

S. Forrest. Genetic algorithms: principles of natural selection applied to computation. *Science*, 261(5123):872–878, 1993. 49

F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175, July 2012. 61

J. Fürnkranz and P. A. Flach. An analysis of rule evaluation metrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 202–209, 2003. 103

N. d. L. Gaioso. O fenômeno da evasão escolar na educação superior no brasil. *Brasília, DF: Universidade Católica de Brasília*, 2005. 21, 22

D. E. Goldberg and J. H. Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988. 47

D. E. Goldberg et al. A note on boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing. *Complex Systems*, 4(4):445–460, 1990. 51

A. G. H. Gonzalez, R. A. M. Armenta, L. A. M. Rosales, A. G. Barrientos, J. L. T. Xihuitl, and I. Algredo. Comparative study of algorithms to predict the desertion in the students at the itsm-mexico. *IEEE Latin America Transactions*, 14(11):4573–4578, 2016. 30

C. E. L. Guarín, E. L. Guzmán, and F. A. González. A model to predict low academic performance at a specific enrollment using data mining. *Revista Iberoamericana de Tecnologias del Aprendizaje*, 10:119–125, Aug 2015. ISSN 1932-8540. 30, 31

H. A. Henley. Prelude: Uma arquitetura para análise de evasão no ensino superior por meio de aprendizado de máquina relacional. Mestrado em ciências da computação, Universidade Federal Fluminense, Niterói, RJ, 2018. 32, 33

N. Horning. Introduction to decision trees and random forests. *Am. Mus. Nat. Hist*, 2:1–27, 2013. 46

E. R. Hruschka, R. J. Campello, A. A. Freitas, et al. A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(2):133–155, 2009. 51

S. H. Huang. Dimensionality reduction in automatic knowledge acquisition: a simple greedy search approach. *IEEE Transactions on Knowledge and Data Engineering*, 15 (6):1364–1373, 2003. 56, 63

W. H. Inmon and D. Linstedt. *Data architecture: a primer for the data scientist: big data, data warehouse and data vault*. Morgan Kaufmann, 2014. 23, 34, 39

W. H. Inmon, J. A. Zachman, and J. G. Geiger. *Data stores, data warehousing and the Zachman framework: managing enterprise knowledge*. McGraw-Hill, Inc., 1997. 34

N. Japkowicz. Why question machine learning evaluation methods. In *AAAI Workshop on Evaluation Methods for Machine Learning*, pages 6–11, 2006. 103

S. Kannan, J. H. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, pages 2227–2236, 2018. 63

S. Khan, K. A. Shakil, and M. Alam. Pabed – a tool for big education data analysis. In *2019 IEEE International Conference on Industrial Technology (ICIT)*, pages 794–799, Feb 2019. 30

R. Kimball and J. Caserta. *The data warehouse ETL toolkit*. John Wiley & Sons, 2004. 37

R. Kimball and M. Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011. 34, 39

B. Kitchenham, O. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering –a systematic literature review. *Inf. Softw. Technol.*, 51(1):7–15, 2009. 27

R. Kohavi and F. Provost. Glossary of terms. *Machine Learning*, 30(2):271–274, Feb 1998. ISSN 1573-0565. URL `https://doi.org/10.1023/A:1017181826899`. 101, 104

K. Kohli and S. Birla. Data mining on student database to improve future performance. *International Journal of Computer Applications*, 146(15), 2016. 30

S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007. 73

H. C. Kraemer. Ramifications of a population model forκ as a coefficient of reliability. *Psychometrika*, 44(4):461–472, Dec 1979. ISSN 1860-0980. URL https://doi.org/10.1007/BF02296208. 107

J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977. 107

S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446. 71

S. A. Lira and A. C. Neto. Coeficientes de correlação para variáveis ordinais e dicotômicas derivados do coeficiente linear de pearson. *Ciência & Engenharia*, 15(1/2):45–53, 2006. 106

H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data mining and knowledge discovery*, 6(4):393–423, 2002. 48

Y. Liu, J. Cheng, C. Yan, X. Wu, and F. Chen. Research on the matthews correlation coefficients metrics of personalized recommendation algorithm evaluation. *International Journal of Hybrid Information Technology*, 8(1):163–172, 2015. 106

G. A. T. Luna, J. Chicaiza, M. B. M. Arciniegas, J. P. U. Torres, V. A. S. Faggioni, and M. S. V. Ludeña. Contribution of big data in e-leaming. a methodology to process academic data from heterogeneous sources. In *Computer Science Society (SCCC), 2016 35th International Conference of the Chilean*, pages 1–12. IEEE, 2016. 23

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017. 71

L. Manhães, S. da Cruz, and G. Zimbrão. The impact of high dropout rates in a large public brazilian university. In *Proceedings of the 6th International Conference on Computer Supported Education-Volume 3*, pages 124–129. SCITEPRESS-Science and Technology Publications, Ltda, 2014. 30, 32

L. M. B. Manhães, S. M. S. da Cruz, and G. Zimbrão. Wave: an architecture for predicting dropout in undergraduate courses using edm. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 243–247. ACM, 2014. 17, 30, 32

P. Maschio, M. A. Vieira, N. Costa, S. de Melo, and C. P. Júnior. Um panorama acerca da mineração de dados educacionais no brasil. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1936, 2018. 26

C. Mason, J. Twomey, D. Wright, and L. Whitman. Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education*, 59(3):382–400, 2018. 30

C. M. B. d. Matta, S. M. G. Lebrão, and M. G. V. Heleno. Adaptação, rendimento, evasão e vivências acadêmicas no ensino superior: revisão da literatura. *Psicologia Escolar e Educacional*, 21:583 – 591, 12 2017. ISSN 1413-8557. URL `http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-85572017000300583&nrm=iso`. 20

B. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451, 1975. ISSN 0005-2795. URL `http://www.sciencedirect.com/science/Article/pii/0005279575901099`. 106

V. L. Miguéis, A. Freitas, P. J. Garcia, and A. Silva. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, 115:36–51, 2018. 30

M. C. Monard and J. A. Baranauskas. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1(1):32, 2003. 101

P. Naccache. Matrix elements and correspondence principles. *Journal of Physics B: Atomic and Molecular Physics*, 5(7):1308, 1972. 104

B. Nakhkob and M. Khademi. Predicted increase enrollment in higher education using neural networks and data mining techniques. *Journal of Advances in Computer Research*, 7(4):125–140, 2016. 30

F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani. Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 20(1):343–357, 2016. 73

P. A. M. M. Nascimento and R. E. Verhine. Considerações sobre o investimento público em educação superior no brasil. *Instituto de Pesquisa Econômica Aplicada (IPEA)*, 2017. 15

S. Negash. Business intelligence. *Communications of the association for information systems*, 13(1):15, 2004. 34

OECD. *Education at a glance 2016: OECD indicators*. OECD Publishing Paris, France, 2016. 14

T. M. Ogwoka, W. Cheruiyot, and G. Okeyo. A model for predicting students' academic performance using a hybrid k-means and decision tree algorithms. *International Journal of Computer Applications Technology and Research*, 4(9):693–697, 2015. 30

J. G. Oliveira Júnior, R. V. Noronha, and C. A. A. Kaestner. Creation and selection of applied attributes in preventing course evasion in undergraduate students (in portuguese). *VIII Computer on the Beach Congress*, pages 061–070, 2017. 17, 30, 32, 33

C. M. Olszak and E. Ziemba. Approach to building and implementing business intelligence systems. *Interdisciplinary Journal of Information, Knowledge, and Management*, 2(1): 135–148, 2007. 34

M. O'Neill, L. Vanneschi, S. Gustafson, and W. Banzhaf. Open issues in genetic programming. *Genetic Programming and Evolvable Machines*, 11(3-4):339–363, 2010. 51

T. M. Padmaja, N. Dhulipalla, R. S. Bapi, and P. R. Krishna. Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. In *15th International Conference on Advanced Computing and Communications (ADCOM 2007)*, pages 511–516. IEEE, 2007. 66

Z. Papamitsiou and A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49, 2014. 30

A. S. Paredes. *A evasão do terceiro grau em Curitiba*. NUPES, 1994. 21, 22

M. B. Patel and J. Dharwa. Selection of optimal classification algorithms in education data mining. *Imperial Journal of Interdisciplinary Research*, 3(1), 2016. 30

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011. 57

W. Peng, J. Chen, and H. Zhou. An implementation of id3-decision tree learning algorithm. *From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May*, 13, 2009. 52

A. G. Picciano. The evolution of big data and learning analytics in american higher education. *Journal of asynchronous learning networks*, 16(3):9–20, 2012. 23

D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2:37–63, 2011. ISSN 2229-3981. URL `http://www.bioinfo.in/contents.php?id=51`. 104

R. Prati, G. Batista, and M. Monard. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, 6(2):215–222, 2008. 105

PROEG. Estudo preliminar sobre evasão, retenção e taxa de sucesso na ufrr. Technical report, Pró-Reitoria de Ensino e Graduação - UFRR, 2016. 16

S. Purushotham and B. Tripathy. Evaluation of classifier models using stratified tenfold cross validation techniques. In *International Conference on Computing and Communication Systems*, pages 680–690. Springer, 2011. 47

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986. 54

J. R. Quinlan. C4.5: Programs for empirical learning, 1993. 54

J. R. Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014. 55

E. Ramentol, J. Madera, and A. Rodríguez. Early detection of possible undergraduate drop out using a new method based on probabilistic rough set theory. In *Uncertainty Management with Fuzzy and Rough Sets*, pages 211–232. Springer, 2019. 30

J. Ranjan. Business intelligence: Concepts, components, techniques and benefits. *Journal of Theoretical and Applied Information Technology*, 9(1):60–70, 2009. 34

A. Rodriguez. Fatores de permanência e evasão de estudantes do ensino superior brasileiro–um estudo de caso. *Caderno de Administração. Revista da Faculdade de Administração da FEA*, 5(1), 2011. 15

L. Rokach and O. Maimon. Chapter 9 decision trees. In *Data mining and knowledge discovery handbook*, pages 165–192. Springer, 2005. 52, 53, 54

C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40 (6):601–618, 2010. 23

C. Romero and S. Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013. 15, 19, 30, 31

A. Rudra and E. Yeo. Key issues in achieving data quality and consistency in data warehousing among large organisations in australia. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, pages 8–pp. IEEE, 1999. 34

L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda. The cart decision tree for mining data streams. *Information Sciences*, 266:1–15, 2014. 53, 54

L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda. A new method for data stream mining based on the misclassification error. *IEEE transactions on neural networks and learning systems*, 26(5):1048–1059, 2015. 53

G. A. S. Santos, A. L. Bordignon, S. L. G. Oliveira, D. B. Haddad, D. N. Brandão, and K. T. Belloze. A brief review about educational data mining applied to predict student's dropout. In *Anais da V Escola Regional de Sistemas de Informação do Rio de Janeiro*, pages 86–91, Porto Alegre, RS, Brasil, 2018. SBC. URL `http://portaldeconteudo.sbc.org.br/index.php/ersi-rj/Article/view/4660`. 24, 30

G. A. S. Santos, A. Bordignon, D. Haddad, D. Brandão, L. Tarrataca, and K. T. Belloze. Data warehouse educacional: Uma visão sobre a evasão no ensino superior. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 235–240, Porto Alegre, RS, Brasil, 2019. SBC. URL `https://sol.sbc.org.br/index.php/sbbd/article/view/8829`. , 35

L. W. Santoso et al. The analysis of student performance using data mining. In *Advances in Computer Communication and Computational Sciences*, pages 559–573. Springer, 2019. 30

A. Sarra, L. Fontanella, and S. Di-Zio. Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research*, pages 1–20, 2018. 30, 31

Y. Sasaki et al. The truth of the f-measure. *Teach Tutor mater*, 1(5):1–5, 2007. 103

O. Seiffert and S. M. Hage. Políticas de ações afirmativas para a educação superior no brasil: da intenção à realidade. *Educação superior no Brasil*, 10:137–162, 2008. 15

H. Sharma and S. Kumar. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, 5(4):2094–2097, 2016. 18, 52, 55

L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T.-M. Chu, F. M. Goodsaid, L. Pusztai, et al. The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8):827, 2010. 106

J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson. Past, present, and future of decision support technology. *Decision support systems*, 33(2): 111–126, 2002. 34

M. M. Shoukri. *Measures of interobserver agreement and reliability*. CRC press, 2010. 107

M. M. Silva. Uma abordagem multiobjetiva para construção automática de algoritmos de indução de árvores de decisão. Dissertação de mestrado, Universidade Federal de São Paulo (UNIFESP), 2015. 55

R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, and M. B. C. M. Lobo. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37(132):641–659, 2007. 15, 19, 20

K. Sin and L. Muthu. Application of big data in educational data mining and learning analytics – a literature review. *ICTACT Journal on soft computing*, 5(4), 2015. 23, 30, 31

S. Singh and P. Gupta. Comparative study id3, cart and c4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, 27(27):97–103, 2014. 18, 53, 55, 57

R. Sivaraj and T. Ravichandran. A review of selection methods in genetic algorithm. *International journal of engineering science and technology*, 3(5):3792–3797, 2011. 50

K. A. Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 160–163, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc. ISBN 1-55860-036-1. URL `http://dl.acm.org/citation.cfm?id=102118.102172`. 105

P. Speller, F. Robl, and S. M. Meneghel. Desafios e perspectivas da educação superior brasileira para a próxima década. *Oficina de Trabalho. p. 164, 2012. ISBN: 978-85-7652-171-6*, 2012. 21, 22

S. V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997. 104

G. Stein, B. Chen, A. S. Wu, and K. A. Hua. Decision tree classifier for network intrusion detection with ga-based feature selection. In *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, pages 136–141. ACM, 2005. 46

S. Sultana, S. Khan, and M. A. Abbas. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering Education*, 54(2):105–118, 2017. 30, 31

P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Introduction to Data Mining (Second Edition)*. Pearson, 2018. 53

N. Tasnim, M. K. Paul, and A. S. Sattar. Identification of drop out students using educational data mining. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–5. IEEE, 2019. 30, 31

A. Tekin. Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research*, 54:207–226, 2014. 30, 31

P. Thakar and A. Mehta. Performance analysis and prediction in educational data mining: A research travelogue. *International Journal of Computer Applications*, 975:8887, 2015. 30, 31

L. M. V. Tigrinho. Evasão escolar nas instituições de ensino superior. *Revista Gestão Universitária*, 173:01–14, 2008. 15, 21

R. Timofeev. Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*, 2004. 18, 46, 53, 55

J. Vasconcelos, J. A. Ramirez, R. Takahashi, and R. Saldanha. Improvements in genetic algorithms. *IEEE Transactions on magnetics*, 37(5):3414–3417, 2001. 50

A. H. Wright. Genetic algorithms for real parameter optimization. In *Foundations of genetic algorithms*, volume 1, pages 205–218. Elsevier, 1991. 50

E. Yukselturk, S. Ozekes, and Y. Türel. Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning*, 17(1):118–133, 2014. 31

M. Zaffar, K. Savita, M. A. Hashmani, and S. S. H. Rizvi. A study of feature selection algorithms for predicting students academic performance. *Int. J. Adv. Comput. Sci. Appl*, 9(5):541–549, 2018. 30

M. Zambon, R. Lawrence, A. Bunn, and S. Powell. Effect of alternative splitting rules on image processing using classification tree analysis. *Photogrammetric Engineering & Remote Sensing*, 72(1):25–30, 2006. 54

Y. Zhang, S. Lu, X. Zhou, M. Yang, L. Wu, B. Liu, P. Phillips, and S. Wang. Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. *Simulation*, 92(9):861–871, 2016. 53

A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, 1(1):32–49, 2011. 49, 51

H. Zhu, L. Jiao, and J. Pan. Multi-population genetic algorithm for feature selection. In *International Conference on Natural Computation*, pages 480–487. Springer, 2006. 48

A- Resultado do Modelo de Predição gerado pelo *EvolveDTree*

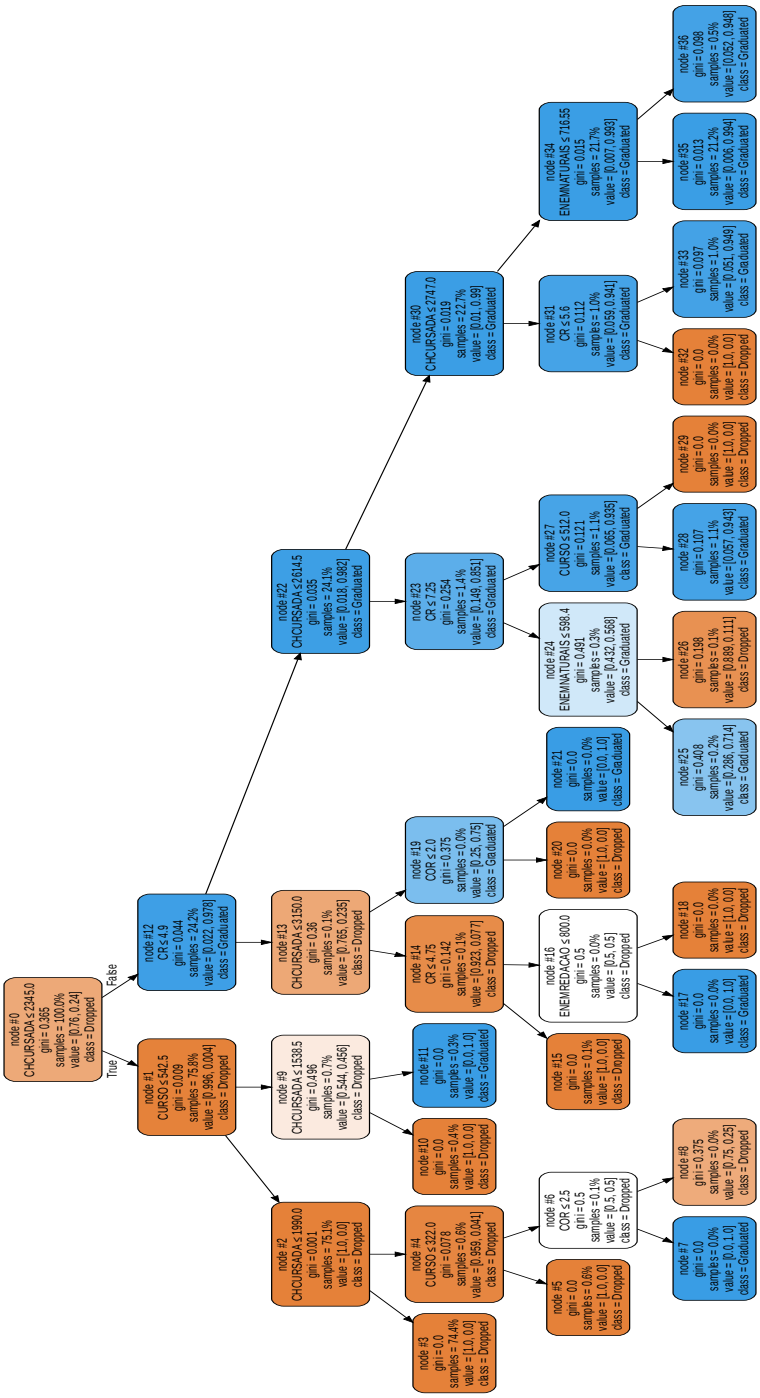


Figura 22 – Saída completa do Fluxograma baseado em Árvore gerada pelo *EvolveDTree*

B- Conjunto de Dados

Tabela 23 – Descrição do Conjunto de Dados

Atributo	Descrição	Tipo
AcaoAfirmativa	Atributo referente à participação em ação afirmativa	Categórico
AnoDesvinculacao	Ano de desvinculação do aluno	Numérico
AnoIngresso	Ano de ingresso do aluno	Numérico
Bairro	Bairro em que o aluno reside	Categórico
CHCursada	Carga horária total cumprida pelo aluno	Numérico
CEP	Código postal do aluno	Categórico
Cidade	Cidade de residência do aluno	Categórico
CodTurnoAtual	Identificador do turno atual do aluno	Numérico
CodTurnoIngresso	Identificador do turno de ingresso do aluno	Numérico
Cor	Atributo referente à cor da pele ou raça do aluno	Categórico
Curso	Identificador do curso de graduação	Numérico
CR	Coeficiente de rendimento do aluno	Numérico
EnemNaturais	Nota de entrada do Enem em ciências naturais	Numérico
EnemHumanas	Nota de entrada do Enem em humanas	Numérico
EnemLinguagem	Nota de entrada do Enem em linguagem	Numérico
EnemMatematica	Nota de entrada do Enem em matemática	Numérico
EnemRedacao	Nota de entrada do Enem em redação	Numérico
EstadoCivil	Estado civil do estudante	Categórico
Idade	Idade do aluno	Numérico
IdAluno	Código de identificação do aluno	Numérico
Mobilidade	Atributo que indica participação em intercâmbio	Numérico
TempoPermanencia	Número de anos matriculados no curso pelo aluno	Numérico
Trancamentos	Número de trancamentos do aluno no curso	Numérico
TurnoCurso	Descrição do turno do curso	Categórico
SemestreDesvinculacao	Semestre de desvinculação do aluno: (1,2)	Numérico
SemestreIngresso	Semestre de ingresso do aluno: (1,2)	Numérico
Sexo	Sexo do aluno: (M,F)	Categórico
StatusFormacao	Classe que indica se o aluno graduou ou evadiu: (1,0)	Numérico

C- Conjunto de atributos transformado por *One-Hot Encoding*

O conjunto total dos atributos, após a transformação das informação de 'ACAOAFIRMATIVA', 'COR', 'CURSO' e 'TURNOATUAL', é apresentado por atributos selecionados e descartados.

Desse modo, seguem os 64 atributos selecionados pelo AG: ENEMMATEMATICA, ENEMREDACAO, CR, ANOINGRESSO, SEMESTREDESVINCULACAO, BAIRRO, CHCURSADA, ESTADOCIVIL, TEMPOPERMANENCIA, ACAOAFIRMATIVA_AC, ACAOAFIRMATIVA_L1, ACAOAFIRMATIVA_L2, ACAOAFIRMATIVA_L3, ACAOAFIRMATIVA_L4, TURNOATUAL_NOTURNO, COR_NEGRA, CURSO_2, CURSO_4, CURSO_5, CURSO_10, CURSO_14, CURSO_15, CURSO_16, CURSO_18, CURSO_21, CURSO_22, CURSO_23, CURSO_24, CURSO_30, CURSO_31, CURSO_32, CURSO_33, CURSO_36, CURSO_39, CURSO_42, CURSO_43, CURSO_47, CURSO_48, CURSO_49, CURSO_52, CURSO_53, CURSO_54, CURSO_60, CURSO_61, CURSO_63, CURSO_65, CURSO_101, CURSO_201, CURSO_221, CURSO_222, CURSO_241, CURSO_243, CURSO_261, CURSO_264, CURSO_287, CURSO_342, CURSO_343, CURSO_382, CURSO_402, CURSO_403, CURSO_422, CURSO_462, CURSO_522, CURSO_195,.

E os atributos descartados: ENEMLINGUAGEM, ENEMHUMANAS, ENEMNATURAIS, SEMESTREINGRESSO, ANODESVINCULACAO, IDADE, CEP, CIDADE, MOBILIDADE, TRANCAMENTOS, SEXO, TURNOATUAL_MATUTINO, TURNOATUAL_VESPERTINO, COR_BRANCA, COR_INDIGENA, COR_NAODECLARADO, COR_PARDA, CURSO_6, CURSO_7, CURSO_9, CURSO_17, CURSO_20, CURSO_25, CURSO_26, CURSO_27, CURSO_28, CURSO_29, CURSO_34, CURSO_35, CURSO_37, CURSO_38, CURSO_40, CURSO_41, CURSO_44, CURSO_45, CURSO_46, CURSO_50, CURSO_51, CURSO_55, CURSO_56, CURSO_57, CURSO_58, CURSO_59, CURSO_62, CURSO_64, CURSO_102, CURSO_194, CURSO_242, CURSO_244, CURSO_245, CURSO_262, CURSO_263, CURSO_282, CURSO_286, CURSO_288, CURSO_302, CURSO_322, CURSO_362, CURSO_502, CURSO_523, CURSO_3.

D- Métricas de Avaliação

Durante a avaliação dos algoritmos de AM, o conhecimento obtido nos domínios investigados é fornecido pelo conjunto de dados. Na maioria dos casos, as características das técnicas existentes e do problema que está sendo solucionado são consideradas na escolha do modelo algorítmico a ser utilizado [Facelli et al., 2015].

Diante disso, a avaliação de um algoritmo de AM supervisionado é normalmente realizada por meio da análise de desempenho do preditor, com base na rotulação gerada por ele às instâncias analisadas [Monard and Baranauskas, 2003]. Essa avaliação é feita através das métricas de avaliação.

As métricas de avaliação concedem uma visão qualitativa dos resultados e permitem uma melhor análise perante a execução do algoritmo escolhido. Para compreender as métricas de avaliação, é necessário entender as seguintes terminologias e definições utilizadas na classificação de um modelo [Kohavi and Provost, 1998]:

- Verdadeiro Positivo (TP, do inglês *True Positive*): significa uma avaliação correta da classe positiva. Por exemplo, a classe dos alunos graduados é a classe positiva (1) e o modelo classificou-os corretamente como graduados;

- Verdadeiro Negativo (TN, do inglês *True Negative*): significa uma avaliação correta da classe negativa. Por exemplo, a classe dos alunos evadidos é a classe negativa (0) e o modelo classificou-os corretamente como evadidos;

- Falso Positivo (FP, do inglês *False Positive*): significa uma avaliação errada da classe positiva. Ou seja, o modelo classificou alunos graduados erroneamente como evadidos;

- Falso Negativo (FN, do inglês *False Negative*): significa uma avaliação errada da classe negativa. Ou seja, o modelo classificou alunos evadidos erroneamente como graduados.

Para analisar as técnicas de classificação, existem diversas métricas que serão apresentadas nas seções a seguir.

D.1- Acurácia

A acurácia refere-se a uma métrica que visa avaliar o nível de predição de um modelo perante à realidade que está sendo representada. O termo acurácia é aplicado no contexto de modelos de classificação. Sendo assim, a acurácia de um classificador pode ser definida como a relação entre o número de instâncias corretamente classificadas e o número total de instâncias avaliadas. A métrica de acurácia é calculada segundo a Equação 4.

$$Acuracia = (TP + TN)/(TP + FP + TN + FN) \tag{4}$$

É possível perceber que a acurácia avalia numericamente em como o classificador se desempenhou de uma maneira geral, pois ela mede a quantidade de acertos sobre o total de instâncias. Dessa forma, sabe-se que a acurácia avalia o percentual de instâncias classificadas corretamente.

D.2- Precisão

Uma visão simplificada quanto ao desempenho de um sistema de classificação é dada pela métrica de precisão, também conhecida como especificidade. Esta métrica corresponde ao número de vezes que a classe positiva foi predita corretamente, dividido pelo número total de vezes que a classe positiva foi predita, corretamente ou erroneamente. Diante disso, a precisão é calculada de acordo com a Equação 5.

$$Precisao = TP/(TP + FP) \tag{5}$$

D.3- *Sensibilidade*

A métrica de sensibilidade em AM é sinônimo de uma taxa positiva verdadeira. Essa métrica também é conhecida como *recall* e representa o número de predições corretas de uma classe dividido pelo número de vezes que a classe aparece no dado de teste. Ou seja, a sensibilidade avalia em que medida todos os exemplos que precisavam ser classificados como positivos foram classificados corretamente [Japkowicz, 2006]. Essa métrica é definida pela Equação 6.

$$Sensibilidade = TP/(TP + FN) \tag{6}$$

Em princípio, espera-se que a sensibilidade resulte em um bom quantitativo de casos relevantes. A sensibilidade possui destaque por ser reconhecida como a métrica mais difundida para avaliar regras e observar a proporção de exemplos positivos na identificação de padrões frequentes. Em mineração de dados, ela é chamada de confiança [Fürnkranz and Flach, 2003].

D.4- *F–Score*

Na análise estatística da classificação binária, a métrica *F–Score* (também conhecida por *F* ou F1) considera tanto a $Precisao$ quanto a $Sensibilidade$ do teste a ser calculado, onde o número de resultados positivos corretos é dividido pelo número de todos os resultados positivos, e o número de resultados positivos corretos é dividido pelo número de resultados positivos que deveriam ter sido retornados. *F–Score* pode ser interpretada como uma média ponderada da $Precisao$ e da $Sensibilidade$, em que F atinge seu melhor valor em 1 e o pior em 0 [Sasaki et al., 2007]. A *F-Score* é calculada segundo a Equação 7

$$F = 2 \times ((Precisao \times Sensibilidade)/(Precisao + Sensibilidade)) \tag{7}$$

Por meio da *F–Score*, é possível ser mais preciso sobre o desempenho do classificador, já que se avalia a eficiência de classificação [Powers, 2011].

D.5- Matriz de Confusão

No campo de Aprendizado de Máquina e, especificamente, para a classificação binária, a matriz de confusão representa uma tabela específica que permite a visualização do desempenho de um algoritmo, tipicamente usado em Aprendizado Supervisionado [Stehman, 1997]. Outro contexto de uso desta métrica é em Aprendizado Não Supervisionado e, nesse caso, essa matriz é conhecida como matriz de correspondência [Naccache, 1972].

A matriz de confusão é um tipo especial de tabela, com duas dimensões (“real” e “predita”’) e suas classes, capaz de resumir o desempenho de um classificador para as tarefas de classificação binária avaliadas. A Matriz de Confusão é apresentada da seguinte forma [Kohavi and Provost, 1998]:

		Valor Predito		
		Positivo	Negativo	
Valor Real	Positivo	*TP*	*FP*	*TP+FP*
	Negativo	*FN*	*TN*	*FN+TN*
		TP+FN	*FP+TN*	

Tabela 24 – Matriz de Confusão

Cada linha da matriz de confusão retrata as instâncias em uma classe real, enquanto cada coluna refere-se às instâncias em uma classe predita, conforme pode ser visto na Tabela 24. O nome deriva do fato de que fica mais fácil ver se o sistema está confundindo duas classes [Powers, 2011].

D.6- Curva ROC

A curva ROC (do inglês, *Receiver Operating Characteristics*) é útil para avaliar um classificador. Os gráficos ROC são comumente usados na tomada de decisões médicas e, nos últimos anos, têm sido utilizados em pesquisas de AM [Fawcett, 2006]. Um dos primeiros a adotar gráficos de curva ROC em AM foi o pesquisador Spackman [1989], demonstrando a aplicação de curvas ROC na avaliação e comparação de algoritmos de AM. Na Figura 23, há uma representação ilustrativa de como essa métrica é apresentada.

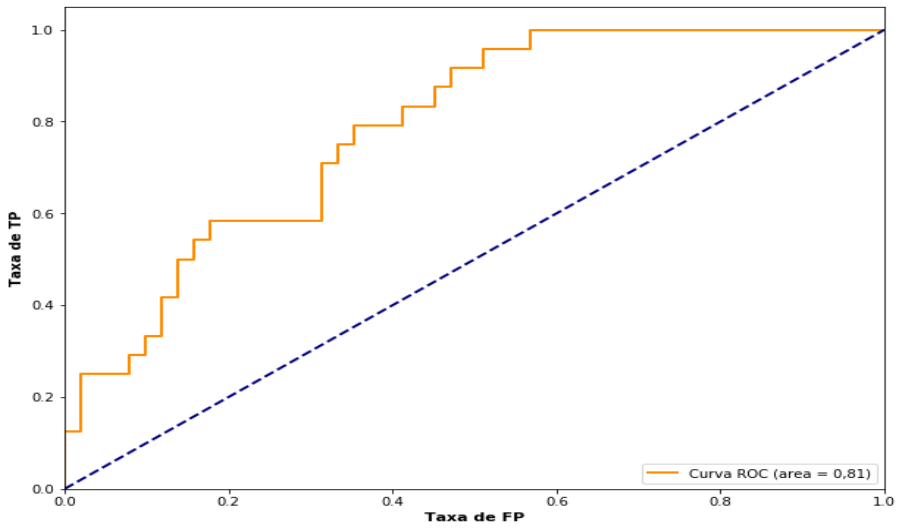


Figura 23 – Imagem ilustrativa da Curva ROC

Sistemas de classificação geralmente produzem um resultado situado no intervalo contínuo de $[0; 1]$. Sendo assim, é preciso definir um ponto de separação para identificar o número de predições positivas e negativas.

Os pontos de separação de uma Curva ROC são calculados através dos valores de precisão e sensibilidade. Esses valores são definidos no gráfico, confome pode ser visto na Figura 23 pela “Taxa de TP” e pela “Taxa de FP”, que é o complemento da precisão ($1 - Precisao$).

Nos últimos anos, houve um aumento no uso de gráficos ROC na comunidade de aprendizado de máquina, devido, em parte, à realização dessa precisão de classificação de forma simples e assertiva [Fawcett, 2006]. Além de ser um método de representação gráfica geralmente útil, os gráficos ROC têm propriedades que os tornam especialmente úteis para domínios com classes binárias desbalanceadas [Prati et al., 2008].

D.7- Coeficiente de Correlação de Matthews

A métrica do Coeficiente de Correlação de Matthews (MCC) foi introduzida pela primeira vez para avaliar o desempenho da predição de uma estrutura secundária de proteína [Matthews, 1975]. Devido ao seu bom desempenho, o MCC tornou-se uma métrica amplamente utilizada em vários estudos [Chicco, 2017; Boughorbel et al., 2017; Liu et al., 2015]. O MCC obteve grande destaque quando foi escolhido como uma das principais métricas de avaliação, no projeto de pesquisa chamado MAQC-II, liderado pela agência federal norte-americana FDA (*Food and Drug Administration*), que visava alcançar um consenso sobre as melhores práticas para desenvolvimento e validação de modelos preditivos para atendimento personalizado na medicina [Shi et al., 2010]. O MCC é calculado pela Equação 8

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{8}$$

Durante o seu processo de cálculo, um valor resultante entre -1 e $+1$ é apresentado como solução:

- Desempenho inaceitável: $[\, 0 > MCC \geqslant -1]$
- Desempenho aleatório: $MCC \approx 0$
- Desempenho significativo: $[\, 0 < MCC \leqslant 1]$

Conforme os cenários de desempenho mencionados, é válido destacar que o resultado $+1$ representa uma predição perfeita, 0 corresponde a uma predição aleatória média, e -1 caracteriza uma predição inversa. Este comportamento estatístico é equivalente ao coeficiente de Pearson [Baldi et al., 2000; Lira and Neto, 2006], o qual tenta resumir a qualidade da tabela de contingência em um único valor numérico possível de comparação.

O MCC é essencialmente a relação entre o observado e o previsto na classificação binária que retornando um intervalo de valores em -1 até 1. No processo de cálculo deste coeficiente, é possível também usar a matriz de confusão para obter o valor relativo da classificação de resultados recomendada[Liu et al., 2015].

D.8- Coeficiente de Kappa

O coeficiente $Kappa$, proposto por Cohen [1968], é um método estatístico para avaliar o nível de concordância ou reprodutibilidade entre dois avaliadores. É bastante utilizado, por exemplo, para analisar questionários em fase de validação [Kraemer, 1979]. O Coeficiente $Kappa$ é obtido pela Equação 9 ([Cohen, 1960]).

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{9}$$

O coeficiente $Kappa$ pode, então, ser definido como a proporção de acordo entre a proporção de acordo devido ao acaso, apresentada na fórmula, onde p_o é a taxa de concordância observada e p_e é a taxa de concordância hipotética. Sendo assim, quando a concordância é total entre os avaliadores, tem-se $\kappa = 1$. De acordo com Landis and Koch [1977], as taxas de concordância pode ser classificadas conforme a Tabela 25.

Tabela 25 – Niveis de Concordancia de $Kappa$

Valor de κ	**Nivel de Concordancia**
< 0	Não há
0 - 0,20	Mínima
0,21 - 0,40	Razoável
0,41 - 0,60	Moderada
0,61 - 0,80	Substancial
0,81 - 1,00	Perfeita

Segundo Shoukri [2010], apesar da popularidade do coeficiente Kappa como medida de concordância entre avaliadores, esta métrica apresenta limitações. Por exemplo, na avaliação dos marcadores de diagnósticos, é sabido que certos testes clínicos, mesmo com uma alta sensibilidade e precisão, podem apresentar baixa capacidade preditiva. Analogamente, dois avaliadores que conseguem apresentar uma alta concordância podem produzir um valor baixo para o coeficiente $Kappa$, conforme apresentado em [Kraemer, 1979].