CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

SELEÇÃO DE ATRIBUTOS DA BASE DO CENSO ENSINO SUPERIOR BRASILEIRO PARA ANÁLISE DE EVASÃO

Danielle Fontes de Albuquerque

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): Rafaelli Coutinho
Coorientador(a): Diego Brandão

Rio de Janeiro,
Maio de 2022

SELEÇÃO DE ATRIBUTOS DA BASE DO CENSO ENSINO SUPERIOR BRASILEIRO PARA ANÁLISE DE EVASÃO

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.

Danielle Fontes de Albuquerque

Banca Examinadora:

Presidente, Professora D.Sc. Rafaelli Coutinho (CEFET/RJ) (Orientador(a))

Professor D.Sc. Diego Brandão (CEFET/RJ) (Coorientador(a))

Professor D.Sc. Eduardo Ogasawara (CEFET/RJ)

Professor D.Sc. Alessandro Vivas Andrade (UFVJM)

Professor D.Sc. Cristiano Maciel (UFMT)

Rio de Janeiro,

Maio de 2022

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

A345 Albuquerque, Danielle Fontes de
Seleção de atributos da base do censo ensino superior brasileiro para análise de evasão / Danielle Fontes de Albuquerque. — 2022.
97f. : il. (algumas color.), enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2022.
Bibliografia : f. 88-97
Orientador: Rafaelli Coutinho
Coorientador: Diego Brandão

1. Mineração de dados (Computação) – Educação. 2. Evasão escolar. 3. Ensino superior. 4. Banco de dados – Análise. I. Coutinho, Rafaelli. (Orient.). II. Brandão, Diego (Coorient.). III.Título.

CDD 006.312

AGRADECIMENTOS

A Deus, por ter conduzido meu caminho até aqui, me dado forças, coragem e sabedoria para superar meus limites e por ter colocado pessoas ao meu lado que me deram todo apoio necessário.

Aos meus pais, que sempre tiveram a minha educação como prioridade, me deram todo o suporte físico e emocional para que eu evoluísse como pessoa e como profissional. Cuidam de mim até hoje com muito carinho e respeito às minhas escolhas. Minha imensa gratidão por serem os melhores ombros para rir e chorar.

Ao meu marido, que para mim sempre foi uma grande referência de que a ciência vale a pena. Me incentivou a não desistir e a ser uma cientista, profissional e pessoa cada dia melhor. Obrigada por diversas vezes me ajudar a superar desafios, ter paciência e compreensão nos momentos difíceis, me proporcionar muitos dias de alegria e por ser a melhor pessoa que eu poderia ter ao meu lado para compartilhar a vida.

Ao meu irmão e amigos, por tanto ouvirem minhas histórias sobre o mestrado e por estarem sempre presentes, me dando apoio, aconchego e momentos de diversão que fazem a vida valer a pena.

A minha orientadora professora Rafaelli, que me acolheu com muito carinho e cuidado. Acreditou que eu poderia sim realizar a transição de carreira e entrar na área da computação. Me deu um suporte muito maior do que o esperado e me fez evoluir em conteúdo, escrita, posicionamento e atitude se tornando minha grande referência em relação a pessoas que fazem seu trabalho com amor e qualidade. Ao meu coorientador Diego, que me ajudou a superar desafios e limites. Me ensinou muito sobre mineração de dados e me fez entender que essa realmente era a área de estudos que eu queria. Principalmente porque me fez acreditar que eu podia ir muito mais além do que eu imaginava.

A toda a equipe do PPCIC, por terem construído um curso extremamente organizado, prático e útil para a sociedade moderna. Por me fazerem acreditar que eu estava no lugar certo apesar de todos os desafios. Por último, ao CEFET, por ter me proporcionado educação pública e de qualidade durante toda minha carreira.

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

RESUMO

Seleção de atributos da Base do Censo Ensino Superior Brasileiro para Análise de Evasão.

Cada vez mais o setor da educação tem utilizado dados para auxiliar na tomada de decisão dentro das instituições de ensino. Um dos principais problemas enfrentados por essas instituições é a evasão. Ela consiste em um fenômeno preocupante pois gera prejuízos sociais e econômicos tanto para o estudante quanto para a sociedade. Uma maneira de reduzir os impactos da evasão consiste em identificar quais são as possíveis causas do problema por meio das bases de dados disponíveis nas instituições de ensino, podendo ser utilizado para isso técnicas da área de Mineração de Dados Educacionais. Ela é uma área interdisciplinar que usa técnicas computacionais e estatísticas para compreender o cenário educacional a partir das bases de dados das instituições de ensino. A grande quantidade de atributos presentes nessas bases dificultam a construção dos modelos de previsão. Para resolver esse problema, é comum fazer uso da seleção de atributos, que é um conjunto de técnicas capaz de identificar quais são os atributos mais relevantes em uma base de dados extensa e simplificá-la de forma que seja possível expressar a informação com um volume menor de dados. Com isso, é possível realizar análises de bases de dados menores e mais limpas, o que facilita o entendimento do problema e melhora o desempenho computacional tanto em relação ao tempo de processamento quanto à qualidade do modelo gerado. Ademais, identificar os atributos mais importantes é uma forma de compreender quais são as possíveis causas e consequências do problema. Esse trabalho busca encontrar os principais fatores que impactam na evasão do ensino superior brasileiro por meio de uma análise comparativa de técnicas de seleção de atributos utilizando os dados do Censo de Ensino Superior, fornecido pelo governo brasileiro, que reúne informações sobre todos os estudantes de ensino superior do país. Uma nova abordagem para seleção de atributos também foi proposta com algoritmo genético para permitir maior flexibilidade e especificidade no cenário educacional, chamada FlexAG. Os resultados mostram que os atributos ano de ingresso, atividade extracurricular e financiamento estudantil são os mais importantes para o cenário geral da base do Censo de Ensino Superior. Além disso, as técnicas de seleção de atributos se mostraram capazes de melhorar as medidas de desempenho de classificação, a redução na quantidade de atributos e o tempo de classificação.

Palavras-chave: Mineração de Dados Educacionais, Evasão, Educação Superior, Seleção de Atributos

ABSTRACT

Feature Selection from the Brazilian Higher Education Census Base for Evasion Analysis

Increasingly, the education sector is using its extensive data repositories to aid decision-making within universities. One of the main problems these institutions face is the dropout of students. It is a worrying phenomenon because it causes social and economic losses for both students and society. One way to reduce the impact of dropouts is to identify the possible causes of dropouts using the databases available in the institutions using techniques in the area of Educational Data Mining. Is it an interdisciplinary area that uses computational and statistical techniques to understand the educational scenario from the databases of educational institutions. The large number of attributes present in these bases make it difficult to build forecast models. To solve this problem, it is common to make use of feature selection. Feature selection is a set of techniques capable of identifying which are the most relevant attributes in a large database and simplifying it so that it is possible to express the information with a smaller volume of data. With this, it is possible to perform analysis on smaller and cleaner databases, which facilitates the problem understanding and improves computational performance in terms of processing time and the quality of the model generated. Furthermore, identifying the most relevant factors is a way to understand the possible causes and consequences of the problem. This work seeks to find the main factors that impact Brazilian higher education dropout through a comparative analysis of attribute selection techniques using data from Education Census Higher, provided by the Brazilian government, which gathers information about all higher education students in the country. A new approach for feature selection was also proposed with Genetic Algorithm to allow more flexibility and specificity in the educational setting, called FlexAG. The results show that the attributes year of entry, extracurricular activity, and student financing are the most important for the overall base scenario of Education Census Higher. In addition, the feature selection techniques can improve the classification performance measures, and reduce the number of attributes and classification time.

Keywords: Educational Data Mining, Dropout, Higher Education, Feature Selection

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvore De Decisão
AG	Algoritmo Genético
CES	Censo De Ensino Superior
EAD	Educação À Distância
IES	Instituições De Ensino Superior
IM	Informação Mútua
INEP	Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira
KNN	*K*-vizinho Mais Próximos
LDA	*Linear Discriminant Analysis*
MDE	Mineração De Dados Educacionais
NB	*Naive Bayes*
OCDE	Organização Para A Cooperação E Desenvolvimento Econômico
PCA	*Principal Component Analysis*
PROUNI	Programa Universidade Para Todos
QQ	Qui-Quadrado
RD	Redução De Dimensionalidade
RF	*Random Forest*
RL	Regressão Logística
RN	Rede Neural
SA	Seleção De Atributos
SBS	*Sequential Backward Selection*
SEMESP	Secretaria De Modalidades Especializadas De Educação
SFS	*Sequential Forward Selection*
SVM	*Support Vector Machine*

SUMÁRIO

1- Introdução

A educação é um dos principais pilares da realidade econômica, social e política de uma nação. Ela funciona como um recurso estratégico capaz de alavancar o desenvolvimento do país e romper barreiras de classes sociais, pois se comporta como forma de ascensão econômica e social que gera independência financeira e cria oportunidades capazes de elevar o nível cultural e de qualidade de vida do cidadão.

Os ganhos futuros de uma sociedade que se atenta à educação são aparentes e capazes de melhorar a realidade do país a médio e longo prazo, com pessoas que tenham capacidade de exercer melhor sua cidadania e criar boas oportunidades para o país. Nesse cenário, a educação superior tem o papel de produzir, validar, disseminar e utilizar o conhecimento de forma a promover a pesquisa científica, a inovação tecnológica, a criação de políticas públicas e estimular o mercado de trabalho gerando emprego e renda para que a população possa ter uma vida mais digna [Campos et al., 2019].

A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) é uma organização econômica intergovernamental que tem por objetivo estimular o desenvolvimento econômico e divulgar relatórios sobre educação com dados de seus países membros. No relatório da OCDE de 2020, foi destacado que entre os jovens de 25 e 34 anos, apenas 21% tem nível superior no Brasil, enquanto que a média para os países membros da organização está em 40% [OECD, 2020].

No entanto, o mesmo relatório mostra o cenário brasileiro de 2009, onde apenas 12% de pessoas nessa mesma idade tinham ensino superior, o que demonstra um ganho de $9p.p$ em 10 anos na busca pelo acesso a educação superior [OECD, 2020]. Esse movimento de democratização da educação é relevante para o país, pois permite o desenvolvimento de diversos setores econômicos e também do indivíduo, ao possibilitar novas oportunidades de carreira.

Para ajudar a monitorar essa evolução existe o CES, que é a principal fonte de divulgação dos dados das instituições de ensino superior no Brasil, mantido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) [BRASIL, 2020]. Os dados correspondem a características pessoais, como cor, sexo e idade, atributos comportamentais, como participação em atividades extracurriculares e estágios,

e características acadêmicas como fonte de financiamento, curso e localização de todos os alunos matriculados nas instituições públicas e privadas do Brasil. O CES serve como fonte importante de informação para a avaliação do setor, bem como insumo para discussões que levarão a criação de políticas públicas na área.

Uma das informações contidas no CES é sobre o trancamento ou fechamento de matrícula, mais conhecido como evasão acadêmica. Ela representa a saída do estudante de um curso ou do sistema de educação sem concluí-lo com sucesso [1] [Da Silva, 2019; Carminati et al., 2020], sendo um problema preocupante no Brasil e no mundo, principalmente na educação superior. Na Tabela 1 estão as informações retiradas do relatório OECD [2019] em relação a situação do aluno após três anos do prazo ideal de conclusão do curso superior para os países membros da OCDE. A tabela mostra a porcentagem de alunos que conseguiram se graduar nesse tempo, ainda estão cursando ou abandonaram o curso sem concluir. A média mundial de abandono está em 24% enquanto países como a Suíça e o Brasil apresentaram respectivamente 12% e 34% de índice de abandono.

Países	Graduado	Ainda cursando	Não graduado e não matriculado
Reino Unido	85	0	14
Israel	83	3	14
Suiça	81	7	12
Irlanda	81	1	18
Nova Zelândia	77	6	18
Finlândia	73	10	18
Noruega	72	9	19
Australia	70	9	21
Países Baixos	70	12	18
Estados Unidos	69	11	20
Islândia	69	9	22
Média	68	9	24
França	67	12	21
Portugal	65	9	26
Estonia	59	8	33
Austria	58	17	25
Suécia	56	12	32
Chile	54	17	30
Eslovênia	53	8	40
Brasil	50	16	34

Tabela 1 – Situação do aluno três anos após o prazo de integralização do curso para países membros da OCDE, em %. Fonte: OECD [2019].

O desperdício social, acadêmico e econômico da evasão pode ser sentido por todos aqueles que estão envolvidos nesse contexto: o próprio aluno, a família dele, as instituições de ensino e o governo [Silva Filho et al., 2007]. Quando acontece nas universidades públicas, o governo não tem retorno social sobre aquele investimento. Nas

[1] Os estudantes com matrícula trancada podem retomar o curso caso estejam no tempo previsto pela regulamentação.

faculdades privadas os alunos e suas famílias arcam com o prejuízo do investimento financeiro perdido. Além disso, o aluno perde tempo e oportunidades pois não consegue uma qualificação profissional suficiente para alavancar a carreira, assim tendo que lidar com o sentimento de frustração [Silva Filho et al., 2007; Do Couto and De Santana, 2017].

Para melhorar esse cenário é necessário entender quais são as possíveis causas da desistência e identificar qual é o risco de evasão do aluno, afim de tomar medidas preventivas antes que ela ocorra [Delen, 2010]. Segundo o relatório OECD [2020], cerca de 86% das pessoas que se formam pela primeira vez no nível superior tem menos que 30 anos, 58% são mulheres e 25% nos cursos da área de negócios, administração e direito. Dessa forma, será que apenas essas informações são importantes para caracterizar se um aluno concluirá ou não o curso? Além desses fatores básicos (idade, gênero e área do curso), podem existir outros fatores não tão óbvios vindos de mudanças na realidade da população.

A reportagem do UOL [2020] traz uma preocupação em relação à evasão acadêmica no cenário da pandemia do COVID-19, onde 83 mil alunos a mais do que no mesmo período de 2019 desistiram ou trancaram a matrícula no primeiro semestre de 2020 em instituições particulares de ensino superior, de acordo com pesquisa feita pela Secretaria de Modalidades Especializadas de Educação (SEMESP) [UOL, 2020]. A reportagem cita o fator financeiro como principal responsável por esse movimento de evasão nas instituições particulares, o que pode representar um possível impacto da crise econômica no futuro da educação superior.

Dessa forma, os atributos que impactam na decisão de evasão podem ser de diferentes naturezas, como o contexto social, cultural e político no qual o aluno e a instituição estão inseridos [dos Santos Baggi and Lopes, 2011; Santos et al., 2018; Delen, 2010]. Algumas das causas mapeadas por Carminati et al. [2020] estão associadas a atributos pessoais, como: idade, saúde financeira, dúvida na escolha do curso, nível de satisfação em relação ao curso e a IES, baixo desempenho e dificuldade de relacionamento. Além desses, ainda há aspectos internos e externos às IES como taxa de desemprego, prestígio social da profissão, dificuldade de conciliar horários de estudo e trabalho, localização da IES, entre outros.

Neste contexto a Mineração de Dados Educacionais (MDE) tem sido uma ferramenta importante para atender as demandas do setor educacional como um todo, pois com ela é possível mapear diversos indicadores como: o perfil do aluno, o risco de evasão

acadêmica e os principais fatores que impactam no abandono do curso [Colpo et al., 2020; Romero and Ventura, 2013; Baker and Yacef, 2009; Baker et al., 2011]. Por outro lado, analisar bases de dados como CES é uma tarefa complexa devido as dificuldades associadas a alta dimensionalidade dos dados.

Além da alta dimensionalidade e do grande volume de amostras, os dados do CES são esparsos, com ruído, redundantes e muitos podem ser irrelevantes para descobrir informações acerca da evasão. Essas características podem confundir os modelos matemáticos aumentando o risco dos algoritmos de mineração de dados encontrarem padrões errados e gerando *overfitting* [Chandrashekar and Sahin, 2014], isto ocorre quando o modelo está sobre-ajustado a uma parte da base de dados mas não tem boa capacidade de generalização.

Uma forma de amenizar esse problema é realizar um pré-processamento da base de dados onde são identificados quais atributos são importantes para o estudo em questão, neste caso a análise de evasão acadêmica. Dessa forma, é possível reduzir o volume de dados facilitando o processamento, além de que focar apenas no que é relevante traz informações mais adequadas e assertivas para os modelos de classificação. A SA é uma técnica de redução de dados Han et al. [2012] que tem por objetivo reduzir a dimensionalidade da base de dados sem perder informações importantes.

A escolha dos atributos que devem estar presentes no modelo não é simples. Ela foi demonstrada como um problema NP-completo Davies and Russell [1994], devido a complexidade da busca exaustiva de subconjuntos a partir de uma alta dimensionalidade. Por isso, diversos métodos de seleção de atributos já foram propostos para otimizar a escolha dos melhores atributos, usando, por exemplo, métodos estatísticos e meta-heurísticas.

Especificamente no contexto da evasão, a SA pode possibilitar a descoberta dos atributos que estão mais associados ao risco de evasão. Isso permitiria dar mais atenção a esses fatores de impacto, criar medidas práticas de prevenção ao abandono acadêmico mais direcionadas, como programas de apoio aos estudantes e, possivelmente, melhorar a qualidade do modelo de classificação para o risco de evasão, visto que as informações menos relevantes são omitidas. Neste contexto, muitos trabalhos vem estudando técnicas de SA para auxiliar a classificação do risco de evasão a fim de melhorar a assertividade dos modelos e encontrar os principais atributos relacionados a esse problema. [Teodoro and Kappel, 2020; Febro, 2019; Urbina-Nájera et al., 2020; Muchuchuti et al., 2020; Niu

et al., 2018; Gopalakrishnan et al., 2018; Gitinabard et al., 2018].

Um exemplo de aplicação da SA em MDE aparece no trabalho de Teodoro and Kappel [2020], que mostrou através de uma técnica de SA que a identificação do risco de evasão a partir do perfil dos alunos é fundamental para traçar planos que visem mitigar o problema como, por exemplo, criar programas de assistência para fortalecer o vínculo entre o aluno e a instituição de ensino. Alguns dos principais atributos identificados por eles para caracterizar a evasão no ensino superior público brasileiro é a idade, a participação em atividades extracurriculares e a carga horária total do curso. Diversos outros trabalhos também buscaram identificar quais são as causas da evasão acadêmica no Brasil a fim de propor soluções que melhorem esse cenário [da Silva et al., 2019; Santos et al., 2020; Do Couto and De Santana, 2017; Magalhães Hoed et al., 2018; Manhaes et al., 2015].

Com base nessa discussão, o objetivo geral deste trabalho é encontrar os principais fatores que impactam na classificação da evasão no ensino superior brasileiro por meio de técnicas de seleção de atributos. Para isso, uma análise comparativa das técnicas de SA foi realizada por meio da aplicação de diferentes tipos e combinações de métodos de SA com classificadores a fim de compreender a eficiência desses métodos no desempenho da classificação. Foram avaliados também os cenários referentes a alunos de IES públicas e privadas bem como cursos presenciais e EaD para compreender se há diferença no desempenho das técnicas e nos atributos selecionados.

Além disso, como os trabalhos presentes na literatura utilizam apenas técnicas clássicas de SA, observou-se a oportunidade de usar elementos específicos do cenário educacional no processo de seleção dos atributos. Neste sentido, também foi proposto um método de SA, chamado FlexAG, que é uma adaptação da abordagem clássica de *wrapper* com Algoritmo Genético (AG) [Guyon and Elisseeff, 2003; Galvão, 2007; Kohavi and John, 1997]. Nele há um fator que leva em consideração o agrupamento de atributos educacionais, com a intenção de descobrir se priorizar determinados grupos impacta positivamente na busca pelo conjunto otimizado de atributos que melhor descreve a evasão acadêmica.

Os grupos podem ser de atributos identificados como demográficos, comportamentais, acadêmicos, relacionados ao corpo docente ou a instituição, por exemplo. Dessa forma, o método possibilita que especialistas em educação tenham maior flexibilidade para inserirem informações empíricas no processo de SA, de acordo com as diversas

necessidades e realidades que o ambiente educacional apresenta, permitindo que características intrínsecas de cada ambiente escolar sejam melhor exploradas. Caso o especialista entenda que as condições da instituição sejam de maior impacto no problema, ele pode dar uma prioridade maior para esse grupo e assim comprovar a sua tese e facilitar a escolha do algoritmo na busca do melhor conjunto de atributos.

Os objetivos específicos deste trabalho são:

- Identificar quais são as combinações de algoritmos de classificação e técnicas de SA que melhor auxiliam na seleção dos principais atributos associados a evasão no contexto do ensino superior brasileiro e compreender quais são esses atributos mais importantes.
- Avaliar se as combinações mais assertivas de métodos de SA e classificadores e os atributos selecionados diferem quando analisamos de forma separada os dados contidos da base do CES para o ensino público *versus* privado, e do ensino presencial *versus* da EaD.
- Compreender se a priorização de grupos de atributos proposta pelo método FlexAG no processo de seleção influência o desempenho dos métodos de classificação no contexto de dados educacionais.
- Avaliar se os principais atributos selecionados para um determinado ano do CES geram resultados semelhantes ao serem aplicados nas bases de dados de outros anos.

Além deste capítulo, este trabalho está estruturado como segue. O Capítulo 2 apresenta os conceitos fundamentais de mineração de dados educacionais e das técnicas de SA. O Capítulo 3 aborda os principais trabalhos encontrados na literatura sobre seleção de atributos em dados educacionais, o Capítulo 4 detalha a metodologia feita nessa pesquisa e o Capítulo 5 apresenta os resultados obtidos. Finalmente, o Capítulo 6 mostra as principais conclusões do estudo.

2- Referencial Teórico

Neste capítulo será apresentada uma visão geral [Grant and Booth, 2009] sobre conceitos fundamentais de Mineração de Dados Educacionais (MDE), técnicas de Seleção de Atributos(SA) e métodos de classificação necessários para predizer a evasão universitária na base do CES. A Seção 2.1 apresenta o contexto e os métodos de MDE. As Seções 2.2, 2.3 e 2.4 descrevem as abordagens para SA e os algoritmos de classificação, respectivamente, que são utilizados na metodologia proposta.

2.1- Mineração de dados educacionais

O advento da Internet e a digitalização dos processos levaram o cenário educacional para uma nova realidade nos últimos anos. Cursos no formato a distância, *softwares* educacionais, bancos de dados públicos e sistemas de gestão escolar computadorizados são alguns exemplos das tecnologias que criaram grandes repositórios de dados capazes de gerar informações que podem auxiliar na descrição da jornada estudantil de um aluno [Romero and Ventura, 2013].

Uma importante ferramenta para explorar essa nova realidade é a mineração de dados, também conhecida como busca de conhecimento através de dados, do inglês *Knowledge Discovery from Data* (KDD). Ela consiste em formas de extrair conhecimento a partir de grande volume de dados [Han et al., 2012], usando conceitos de inteligência artificial, estatística, banco de dados, visualização de informações e modelagem de dados. As etapas do processo de KDD estão ilustradas na Figura 1.

A primeira etapa consiste no pré-processamento, que a partir da base de dados bruta cria uma nova base mais apropriada para o uso das técnicas de mineração de dados. A necessidade de fazer um tratamento prévio dos dados acontece pois muitas das vezes os dados brutos podem conter inconsistências, ruídos, valores faltantes ou não estarem no formato ideal para serem usados. As etapas do tratamento de dados são: integração, limpeza, seleção e transformação [Han et al., 2012].

A integração serve para unificar bases de dados separadas, de maneira coesa por meio das dos campos-chave que as conectam. A limpeza auxilia no tratamento de dados faltantes, inconsistentes ou com ruído, eliminando ou alterando os dados conforme a estratégia adotada. Já a seleção é uma maneira de reduzir a base de dados, selecionando para uso apenas aquilo que será útil para o objetivo do estudo. A transformação dos dados consiste em alterações feitas para adequar os dados para o uso de determinadas metodologias. Dentre elas podemos citar a categorização dos dados, transformação em dados binários, normalização, criação de novos atributos por meio de agrupamentos, entre outros.

Após o pré-processamento, os dados tratados são submetidos a métodos de mineração de dados que tem como objetivo extrair informação útil deles, como por exemplo descobrir padrões. Esses métodos podem ser descritivos ou preditivos, de natureza estatística ou de aprendizado de máquina. Eles geram resultados que podem ser avaliados por meio de métricas bem conhecidas, como acurácia, f_1, dentre outros, sendo necessário ao fim avaliar o que tais resultados dizem sobre o fenômeno avaliado. No Capítulo 4 é possível observar a aplicação dessas etapas da mineração de dados na base do CES.

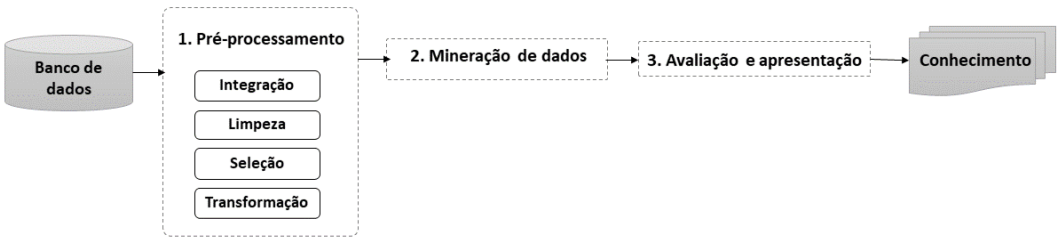

Figura 1 – Etapas da mineração de dados no processo de busca do conhecimento. Adaptado de Han et al. [2012].

A MDE é uma subárea da mineração de dados que explora técnicas computacionais para estudar dados educacionais. Os principais benefícios da MDE consistem em: resolver questões do ambiente educacional, identificar padrões no perfil dos estudantes, otimizar os recursos computacionais e tornar as decisões do contexto educacional mais assertivas [Romero and Ventura, 2010; Baker and Yacef, 2009].

Os dados educacionais podem ser de diversos tipos e usados com diferentes objetivos. Eles podem ter origem em escolas, faculdades, governos ou qualquer outra instituição ligada à educação. Podem ser de natureza comportamental, geográfica, administrativa, demográfica, social, econômica, psicológica, pedagógica, entre outras. A divisão mais utilizada para dados educacionais é acadêmica, social e econômica [Colpo et al., 2020], provavelmente porque esses grupos conseguem dar respostas acerca da história do aluno na universidade, de suas condições particulares e de fatores externos, como por exemplo crises econômicas. Essas informações podem ainda ser referentes aos alunos, as suas famílias, às instituições de ensino, aos professores ou ao curso.

Em Baker and Yacef [2009] foram descritas as principais técnicas de MDE: predição, agrupamento, mineração de relações, uso dos dados para julgamento e descobertas com modelos. Dentre as técnicas de predição, a regressão e a classificação são utilizadas para dados contínuos e discretos, respectivamente. No contexto de MDE a regressão pode aparecer como forma de prever as notas futuras dos alunos, enquanto a classificação é usada para prever situações como se o estudante vai evadir ou não evadir, aprovar ou reprovar. O agrupamento é uma forma de criar grupos de alunos por suas características e descobrir como se assemelham e se diferenciam entre si. Já a mineração de relações busca encontrar relacionamentos de causa e efeito entre os fatores educacionais através de regras de associação e correlações. O uso dos dados para julgamento emprega técnicas de visualização de dados para dar apoio na tomada de decisão [Baker et al., 2011], enquanto na descoberta com modelos, um modelo matemático pode ser criado através de qualquer processo para definir algum fenômeno educacional que possa ser validado, produzindo análises sobre o assunto.

As motivações das aplicações práticas dessas técnicas na MDE foram sugeridas por [Baker and Yacef, 2009] como sendo quatro grandes áreas:

- Modelagem do aluno: apresenta conhecimento sobre as características do aluno, seu processo de desenvolvimento e desempenho acadêmico e o nível de motivação para continuar no curso. Essa aplicação é muito utilizada para previsão de desempenho de notas e de risco de evasão.

- Aprimoramento de modelos na estrutura do conhecimento: cria novos métodos de ensino e formas de monitoramento do processo de aprendizagem, como aplicações complexas que interagem com o aluno, complementam a educação tradicional e

coletam dados da experiência.

- Suporte pedagógico: utiliza novos ambientes de aprendizado para descobrir métodos eficazes de apoio, como por exemplo, aplicações de tutoria online e *feedbacks* para professores.

- Compreensão dos principais fatores que impactam a aprendizagem: busca entender quais fatores tem maior impacto no processo pedagógico a fim de comprovar ou questionar teorias educacionais, como por exemplo, se há diferença de aprendizado entre meninos e meninas.

Em Colpo et al. [2020] foi realizada uma análise metodológica de 23 artigos que usaram técnicas de MDE aplicadas à evasão educacional. Todos usaram técnicas de classificação para prever o risco de abandono ou encontrar atributos importantes. A maioria dos trabalhos usaram dados referentes à graduação e ao sistema público de ensino na modalidade presencial. Cerca de 22% dos trabalhos usaram técnicas de redução de dimensionalidade. A Redução de Dimensionalidade (RD) é o processo de reduzir a quantidade de atributos utilizada [Han et al., 2012]. Ela funciona como solução para a "Maldição da dimensionalidade" [Han et al., 2012], que está relacionada ao quanto um conjunto de dados de alta dimensão pode ser prejudicial para a qualidade dos resultados de algoritmos.

Isso acontece porque dentro do conjunto de dados podem haver atributos que sejam irrelevantes, redundantes e com ruídos que atrapalham a extração de padrões pois distorcem a realidade dos dados, criam ambiguidades e tornam o modelo menos generalizável [Ullah et al., 2017; Venkatesh and Anuradha, 2019]. Para resolver esse problema, a RD cria uma nova base de dados usando o conjunto de atributos que melhor descreve determinada situação a ser avaliada. Os principais benefícios da RD são: diminuir a necessidade de espaço de armazenamento, tornar os algoritmos mais rápidos e eficientes, melhorar a precisão da classificação, melhorar a qualidade dos dados e facilitar a visualização dos resultados [Velliangiri et al., 2019]. A RD pode ser realizada através de técnicas de extração e de SA.

Os métodos de extração combinam os atributos para criar novos atributos em uma nova base de dados com baixa dimensão que consiga representar as informações da base original. As principais abordagens de extração são: Análise de Componentes Principais (*Principal Component Analysis* (PCA)), Análise de Correlação Canônica e

Análise de Discriminante Linear (*Linear Discriminant Analysis* (LDA)). Já a seleção de atributos escolhe um subconjunto de atributos do conjunto de dados com o objetivo de minimizar a redundância e maximizar a relevância dos atributos em relação a classe [Ullah et al., 2017; Aggarwal, 2014]. A seleção ainda pode ser subdividida em quatro abordagens, conforme visto na Figura 2. Na próxima seção as técnicas de SA serão melhor descritas.

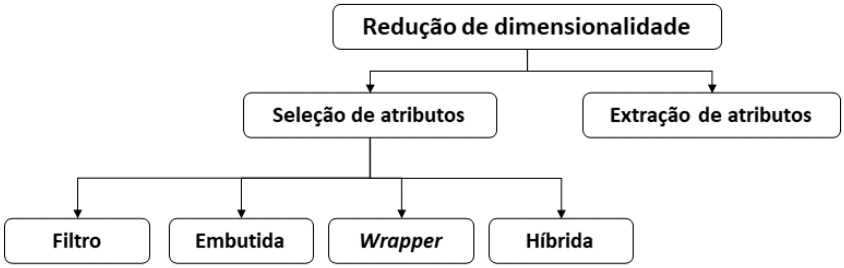

Figura 2 – Tipos de Redução de Dimensionalidade [Ullah et al., 2017; Chandrashekar and Sahin, 2014]

2.2- Seleção de atributos

A seleção de atributos busca definir um subconjunto de atributos obtidos a partir da base original, de maneira a representar a melhor distribuição das classes dos dados. Inicialmente, os subconjuntos podem estar vazios, com todos os atributos ou com alguns selecionados aleatoriamente. A estratégia consiste em selecionar atributos a partir da remoção, inclusão ou troca de elementos dentro desses subconjuntos [Velliangiri et al., 2019; Venkatesh and Anuradha, 2019].

Um critério de parada precisa ser definido para dar fim ao processo. Esse critério pode ter características diferentes, como: quantidade de atributos pré-definida, número fixo de interações ou se o conjunto de atributos conseguiu alcançar determinada meta no

critério de parada (como por exemplo, encontrar o ponto máximo de uma função objetivo).

No final, a nova base de dados com os atributos selecionados deve ser validada, a fim de compreender se o resultado foi satisfatório [Dash and Liu, 1997]. Esse processo pode ser visto na Figura 3. Neste contexto, os métodos de seleção de atributos podem ser classificados em quatro abordagens: filtro, *wrapper*, embutida e híbrida, conforme serão descritos a seguir.

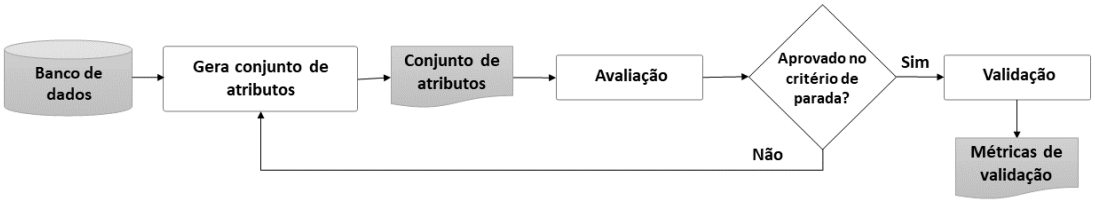

Figura 3 – Processo de seleção de atributos. Adaptado de Dash and Liu [1997].

2.2.1 Abordagem de Filtro

A abordagem de filtro usa propriedades estatísticas intrínsecas dos dados para selecionar atributos, como: distância, dependência, ganho de informação, consistência e correlação. Essa abordagem é vantajosa pois gera maior escalabilidade e menor tempo computacional do que as abordagens *wrapper* e embutida. No entanto, costuma ter menor precisão quando usada em um modelo preditivo, por não usar um algoritmo classificador no processo de escolha [Venkatesh and Anuradha, 2019].

Nessa abordagem há duas formas de fazer a seleção: multivariada e univariada. Na multivariada, os atributos são avaliados em subconjuntos e na univariada os atributos são avaliados individualmente [Liu and Motoda, 2008]. O Qui-Quadrado (QQ) e a Correlação de Pearson são métodos de filtro para seleção de atributos que criam um ranqueamento onde cada atributo é classificado de forma independente. Esses modelos univariados têm a vantagem de possuir maior escalabilidade em relação aos multivariados,

todavia ignoram as redundâncias por não comparar os atributos entre si [Ullah et al., 2017].

A Correlação de *Pearson* é um método estatístico que compara a relação dos atributos dois a dois. A correlação usada na SA é a medida de cada atributo em relação ao atributo alvo (classe). Essa medida é calculada conforme a Equação 1, onde X é o atributo analisado e Y é a classe a ser predita. Na fórmula são usados a covariância entre eles ($Cov(X, Y)$) e o desvio padrão de cada um deles σ_X e σ_Y [Prabha et al., 2019].

$$P(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

O coeficiente de correlação indica o quanto uma variável está associada a classe, isso significa que quanto maior o módulo do valor, mais relacionadas elas são e quanto mais próximo de zero, mais independentes elas são. Caso esse relacionamento seja diretamente proporcional, o valor será positivo e caso seja inversamente proporcional, o valor será negativo. Muitas das vezes usa-se apenas o módulo do valor do coeficiente para criar um ranqueamento de variáveis com maior relação linear.

O filtro QQ avalia os atributos individualmente usando a medida estatística χ^2 com relação à classe. O cálculo utiliza as frequências observadas e estimadas entre o atributo e a classe como na Equação 2, onde n é a quantidade de categorias possíveis para aquele atributo, k é a quantidade de classes, A_{ij} é a frequência observada da categoria i na classe j e E_{ij} é a frequência estimada, calculada pela distribuição dos dados no atributo e na classe [Liu and Setiono, 1995]. Quanto maior o valor de QQ, maior a relação entre o atributo e a classe. Dessa forma, é possível criar um ranqueamento dos atributos e escolher os k que apresentaram maior coeficiente de QQ.

$$QQ = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

2.2.2 Abordagem *Wrapper*

A abordagem de *wrapper* seleciona atributos através de um algoritmo construtor de conjuntos, aplica cada um desses conjuntos no classificador, avalia o resultado da

predição e verifica se o critério de parada foi atendido. Caso positivo, ele indica os atributos presentes nesse conjunto como selecionados, se o critério de parada não for atingido, propõe outro conjunto de atributos [Guyon and Elisseeff, 2003; Galvão, 2007; Kohavi and John, 1997]. Para encontrar o melhor desempenho do classificador é necessário testar vários conjuntos, o que eleva o custo computacional [Kohavi and John, 1997].

Alguns algoritmos, como heurísticas e metaheurísticas, são utilizados para facilitar a escolha das combinações de atributos selecionados para cada conjunto. Dentre tais algoritmos, o Algoritmo Genético (AG) é utilizado para escolher de forma otimizada os atributos que serão testados, reduzindo o custo computacional dessa etapa [Bouaguel, 2016; Lilian et al., 2008].

O AG é um algoritmo de pesquisa aleatória e otimização adaptativa baseado em modelos de evolução natural composto por: cromossomos, populações, gerações, função objetivo, cruzamento e mutação. O cromossomo representa cada solução candidata e a população é um conjunto de cromossomos. Os cromossomos são variados a cada geração por meio de alterações provocadas por operadores genéticos de *crossover* e mutação [Bouaguel, 2016; Lilian et al., 2008].

Na Figura 4(a) está demonstrada a ordem de como os atributos A, B, C e D estão organizados no cromossomo, na mesma ordenação da base de dados. Na Figura 4(b) está a representação fiel do cromossomo que será submetido ao processo de otimização. Ela corresponde a uma sequência de números binários, que indica se o atributo está (valor 1) ou não (valor 0) na solução. Para esse exemplo apenas os atributos C e D foram selecionados.

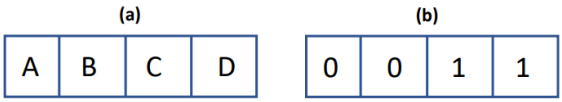

Figura 4 – (a) Representação simbólica de um cromossomos com os atributos A, B, C e D, na ordem que aparecem na base de dados. (b) Representação de um cromossomo real onde apenas C e D são selecionados.

O *crossover* imita o processo evolutivo de recombinação genética, pois recombina e repassa características dos cromossomos de sucesso identificados nas gerações anteriores para as soluções futuras. A mutação também é um fator que impacta na criação da nova população, pois faz a alteração arbitrária de alguns componentes dos cromossomos a fim de manter a diversidade genética da população [Lei, 2012].

A função objetivo, também conhecida como função de *fitness*, avalia a qualidade das soluções propostas nos cromossomos e traz estatísticas para representar a qualidade da população a cada geração, sendo possível compará-las [Bouaguel, 2016].

No início os cromossomos da primeira geração são criados de maneira aleatória, representando as primeiras soluções candidatas geradas. Enquanto o critério de parada não for atingido, o processo evolutivo acontece na concepção de novas gerações com soluções diferentes através das operações de *crossover* e da mutação. O critério de parada pode ser a conclusão da quantidade de gerações dada como parâmetro. Nesse caso, a melhor solução avaliada pela função objetivo é utilizada, indicando os atributos selecionados. A Figura 5 demonstra o funcionamento do AG quando usado como algoritmo de construção de conjuntos na abordagem *wrapper* de seleção de atributos.

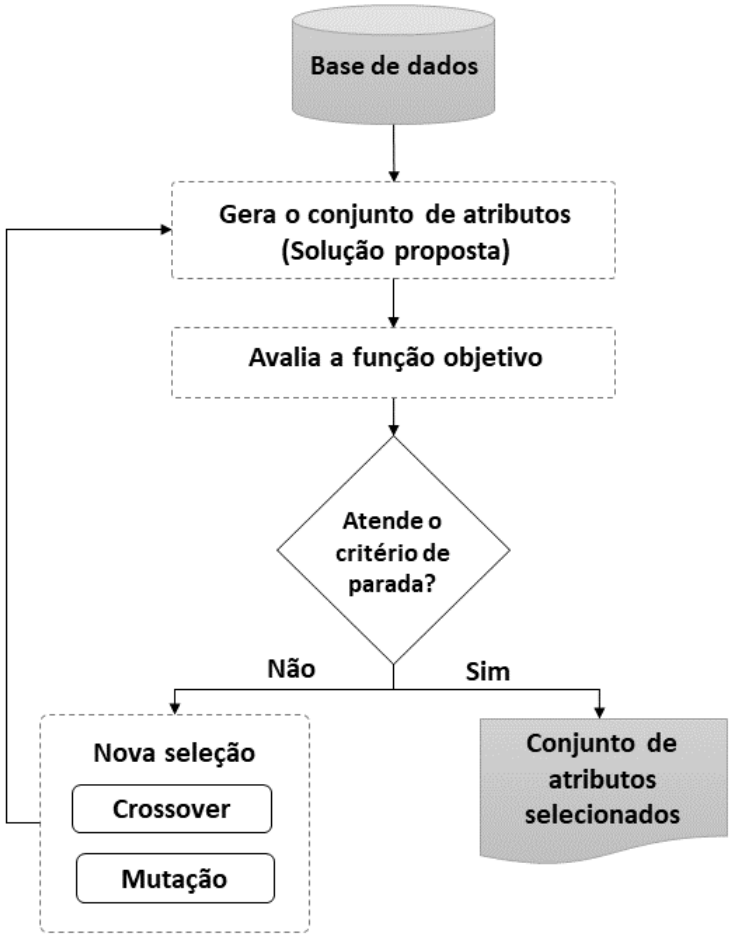

Figura 5 – Processo de seleção de atributos com Algoritmo Genético. Fonte: Elaboração própria, adaptado de Bouaguel [2016].

2.2.3 Abordagem Embutida

A abordagem embutida consiste na utilização de algoritmos de aprendizado de máquina que são capazes de fazer a predição e selecionar atributos simultaneamente. Alguns classificadores como a Árvore de Decisão (AD), *Random Forest* (RF) e Regressão Logística (RL) [Guyon and Elisseeff, 2003] são capazes de selecionar um conjunto de atributos ideal que maximize o desempenho da classificação durante a etapa de treinamento. Essa abordagem é vantajosa pois utiliza o critério estatístico da precisão da própria classificação, sem precisar do custo computacional elevado utilizado pelos algoritmos indutores da abordagem de *wrapper* e sem precisar refazer o treinamento depois da seleção [Ullah et al., 2017].

A AD tem seleção do tipo *para frente*, ou seja, ela vai acrescentando atributos no conjunto dos selecionados. Essa escolha é feita com o critério de impureza utilizado na divisão da árvore (índice de Gini ou Entropia). Esse critério resulta em um ranqueamento dos atributos mais importantes, que é utilizado no processo de treinamento da árvore [Lal et al., 2006]. O algoritmo de CART é um dos algoritmos mais conhecidos de AD e utiliza o ganho de informação do índice de Gini como critério de seleção de atributos. A RF é um método que combina várias árvores de decisão utilizando a média da importância dos atributos encontradas em cada AD [Hasan et al., 2016; Saeys et al., 2008]. Já a RL, usa os coeficientes referentes a cada atributo na equação do modelo de regressão como forma de mensurar o impacto que cada um deles tem na classe que será prevista. Os classificadores AD, RF e RL serão detalhados na Seção 2.3.

2.2.4 Abordagem Híbrida

A abordagem híbrida para SA consiste na utilização de mais de uma abordagem de SA, que pode ser filtro, *wrapper* ou embutida. Ela é uma boa alternativa pois utiliza combinações de técnicas de diferentes naturezas. O método filtro geralmente é o primeiro aplicado por ser um método mais simples e rápido. Depois, sua seleção é usada como entrada para um método de SA de custo computacional mais elevado, como os algoritmos

da abordagem *wrapper* [Wang et al., 2018]. A Figura 6 mostra esse processo, onde os filtros de correlação ou QQ são usados, criando novos conjuntos de atributos menores do que o original. A partir deles, a base de dados é selecionada com esses atributos e submetida ao processo de otimização com a abordagem *wrapper*, que no exemplo está representado pelo AG. Os resultados são conjuntos de atributos ainda menores do que os anteriores para serem utilizados. Assim, é possível combinar a alta precisão dos algoritmos de otimização com menor custo computacional, visto que a base de dados aplicada é menor por já ter sofrido uma pré-seleção da abordagem de filtro, que costuma ter baixo tempo de processamento. O AG é muito utilizado na abordagem híbrida como algoritmo otimizador da abordagem *wrapper*, assim como o filtro de correlação, que também foi encontrado em trabalhos na literatura como sendo uma importante técnica aplicada na abordagem híbrida [Wang et al., 2018; Ruiz et al., 2012; Bermejo et al., 2011]. Outras combinações podem ser feitas, como abordagem embutida e *wrapper* ou filtro e embutida, mas essas foram menos utilizadas na literatura.

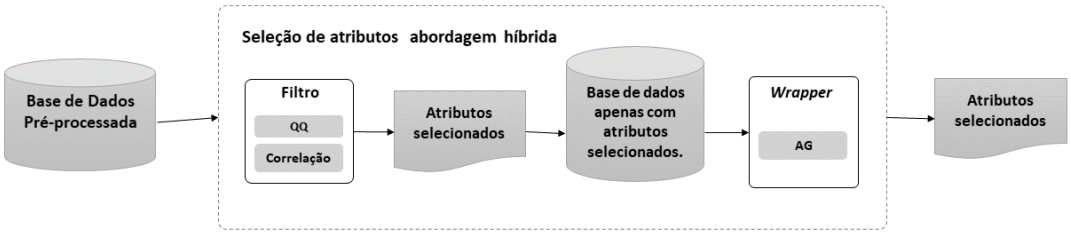

Figura 6 – Processo de seleção de atributos para abordagem híbrida. Fonte: Elaboração própria.

2.3- Aprendizado Supervisionado: Classificação

Estudar como os dados se comportam é uma oportunidade de compreender como um determinado fenômeno funciona e como podemos alterá-lo dentro do processo de tomada de decisão. A mineração de dados é um conjunto de métodos utilizados

na coleta, armazenamento, processamento e visualização dos dados [Evsukoff, 2020]. Dentro dela, encontramos o aprendizado de máquina, que cria modelos matemáticos generalizáveis capazes de prever novas amostras de dados referentes a uma situação específica [Kotsiantis, 2007]. Essa aplicação também acontece no setor educacional por meio da MDE.

Quando um alvo a ser atingido pelo modelo é fornecido, chamamos de aprendizado supervisionado e quando não conhecemos o alvo, chamamos de aprendizado não-supervisionado. Os modelos supervisionados podem ser de regressão, onde a variável de saída é um número real, ou de classificação onde a variável de saída é uma classe.

O objetivo da classificação é que o modelo represente o relacionamento dos atributos e da classe de forma a classificar as observações dentro de alguma das classes pré-definidas [Evsukoff, 2020; Chen et al., 1997]. As etapas desse processo são descritas na Figura 7. A primeira etapa consiste em coletar os dados necessários que são capazes de descrever o problema. A segunda compreende o pré-processamento dos dados, onde são tratados dados faltantes e inconsistência nos dados. Nessa etapa, que foi melhor abordada na seção 2.1, também é realizada a transformação dos dados de forma que eles fiquem apropriados para o uso das técnicas de aprendizado de máquina.

A etapa de divisão da base deve contemplar uma parte para o treinamento do modelo e outra para teste, para compreender o quanto o modelo está aderente a novos dados [Kotsiantis, 2007]. Essa divisão pode ser feita dividindo os dados aleatoriamente em proporções pré definidas, como por exemplo 70% para treinamento e 30% para teste ou usar o modelo de validação cruzada, que divide a base de dados em k conjuntos menores de treinamento e teste. Essa alternativa é uma boa solução para problemas de *overfitting*.

Overfitting é um termo usado para descrever o sobreajuste do modelo de classificação à base de treinamento. Isso significa que ele apresenta anomalias muito específicas da base de treinamento que não estão presentes na base de teste, o que afeta o nível de generalização do modelo para novos dados [Han et al., 2012].

Uma forma de evitar o *overfitting* é realizar a validação cruzada. Ela é uma técnica utilizada para avaliar se um modelo desenvolvido tem capacidade de generalização. Para tanto, ela realiza uma divisão da base onde o total de observações é dividido em k subconjuntos. Para cada modelo gerado, um subconjunto será o de teste enquanto os demais serão utilizado para o treinamento. É realizada a troca de forma que todos os

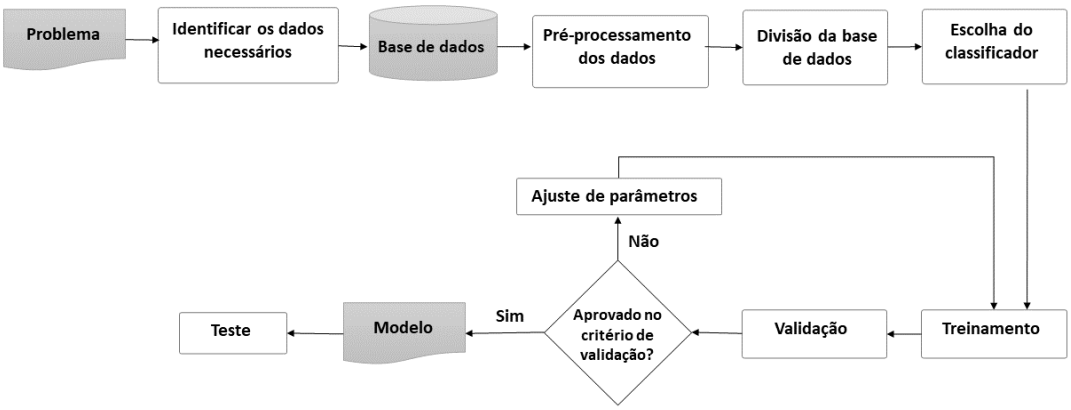

Figura 7 – Processo de Classificação. Fonte: Elaboração própria, Adaptado Kotsiantis [2007].

subconjuntos tenham sido o grupo de teste uma vez, gerando k modelos de classificação, onde o que tiver a melhor validação poderá ser escolhido [Kotsiantis, 2007; Farissi et al., 2020]. Outra forma de obter os resultados é utilizar a média das métricas de desempenho obtidas na validação cruzada. Na Tabela 2 tem-se uma ilustração do modelo de validação cruzada com $k = 5$, onde cada linha representa um modelo com uma das partições sinalizada com x, sendo esta a utilizada para teste.

Grupo de teste da rodada	**k partições com $k = 5$**				
1	x				
2		x			
3			x		
4				x	
5					x

Tabela 2 – Modelo de validação cruzada. Fonte: Elaboração própria, adaptado Farissi et al. [2020]

A escolha de qual tipo de classificador será usado pode ser pautada nas características de cada um deles, podendo usar diferentes combinações de parâmetros para entender qual se adéqua melhor ao problema. Na etapa de treinamento, o modelo é criado com a aplicação do algoritmo classificador na base de dados de treino.

A etapa de avaliação com a base de teste é feita por meio da aplicação do modelo na base de dados separada para teste a fim de compreender se o modelo consegue prever com assertividade em dados novos. Essa classificação de teste tem como resultado os valores dos cálculos das métricas de desempenho, que serão melhor abordadas na

Seção 5. De acordo com as métricas, é possível identificar se elas estão em um nível satisfatório e assim considerar o modelo validado para utilização. Caso as métricas não estejam satisfatórias, é possível fazer ajustes nos parâmetros de treinamento para alcançar modelos mais assertivos.

2.4- Classificadores utilizados

Neste trabalho foram utilizados os seguintes métodos de aprendizado supervisionado: AD, RF e RL. Os mesmos foram usados porque são capazes de criar ranqueamentos referentes à importância desses atributos no processo de classificação. Isso torna possível a aplicação da abordagem embutida e a comparação dela com os demais métodos de SA.

A AD é um algoritmo de busca que gera uma árvore direcionada onde cada nó interno refere-se a uma regra de decisão de particionamento e cada nó folha é referente a uma classe. Os itens percorrem as árvores e passam pelas regras de divisão que particionam os atributos até que esse direcionamento leve a uma resposta sobre a classe, conforme pode ser visto na Figura 8 [Aggarwal, 2014].

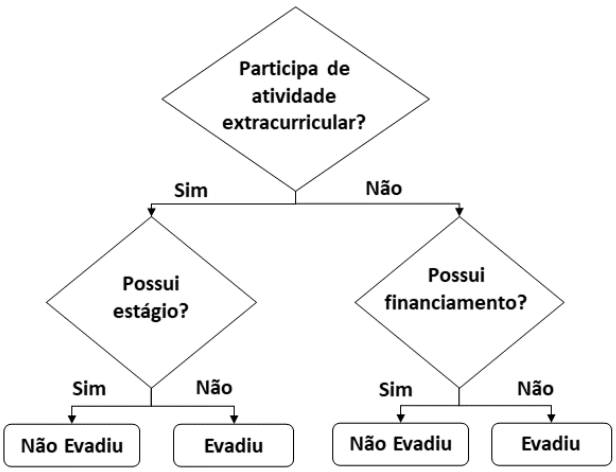

Figura 8 – Exemplo de esquema de uma árvore de decisão. Fonte: Elaboração própria.

A impureza de um nó significa o quanto as classes estão balanceadas nele. O

índice de Gini é a medida de impureza adotada no algoritmo CART [Breiman et al., 2017] de árvore de decisão. Ele mede quantas vezes um elemento escolhido aleatoriamente poderia ser rotulado incorretamente ao seguir a distribuição da base de dados com a presença de determinados atributos. A Equação 3 representa o cálculo do índice de Gini, onde p_i é a probabilidade de cada classe [Rutkowski et al., 2014; Raileanu and Stoffel, 2004].

$$Gini = \sum_{i=1}^{c} {p_i}^2 \tag{3}$$

O ganho de informação refere-se sobre a informação aprendida das classes de um nó para o outro e serve como critério de decisão no particionamento de uma árvore de decisão. A Equação 4 representa o ganho de informação com o índice de Gini, onde $Gini(A)$ é o índice antes da partição do nó e $Gini(P)$ é o índice ponderado dos seus nós filhos [Rutkowski et al., 2014].

$$\Delta Gini = Gini(A) - Gini(P) \tag{4}$$

Já os métodos do tipo *ensemble* são algoritmos que utilizam uma coleção de modelos para criar um único que melhore a capacidade de generalização da classificação [Aggarwal, 2014]. O aprendizado com múltiplos modelos pode transformar classificadores mais simples em outros com maior capacidade de classificação, pois são usadas diferentes amostras de observações e atributos para treinar os diferentes modelos. O resultado dessa combinação pode ser muito diferente das respostas dadas por um único modelo, o que cria maior flexibilidade para o resultado.

O *Bagging* é um tipo de *ensemble* que gera novos subconjuntos de dados com diferentes amostragens para treinar os modelos, fazendo uma combinação por votação majoritária como método de unificação. A Random Forest(RF) é um exemplo desse modelo feito pelo algoritmo de árvore de decisão. Além das características básicas de *Bagging*, a RF usa uma aleatoriedade para promover a diversidade dos dados [Aggarwal, 2014].

A Regressão Logística(RL) é um método de aprendizado de máquina usado em problemas de classificação onde a classe é uma variável do tipo booleana, ou seja, $Y \in \{0, 1\}$. Considerando $X = \langle X_1, \ldots, X_d \rangle$ como sendo um vetor de atributos contendo variáveis discretas ou contínuas, a RL busca $P(Y \mid X)$, ou seja, a probabilidade de Y

acontecer caso X aconteça. Formalmente, o modelo de regressão logística é definido na Equação 5 [Aggarwal, 2014].

$$p(Y = 1 \mid X) = g\left(\theta^T X\right) = \frac{1}{1 + e^{-\theta^T X}} \tag{5}$$

Onde $g(z) = \frac{1}{1+e^{-z}}$ é conhecida como a função logística, ou função sigmoide, $z = \theta^T X$ e $\theta^T X = \theta_0 + \sum_{i=1}^{d} \theta_i X_i$, ou seja, uma função de variáveis e parâmetros. Observe que $g(z)$ tende para 1 quando $z \to \infty$, e $g(z)$ tende para 0 quando $z \to -\infty$.

Como a soma das probabilidades deve ser igual a 1, $p(Y = 0|X)$ pode ser então estimado usando a Equação 6.

$$p(Y = 0 \mid X) = 1 - g\left(\theta^T X\right) = \frac{e^{-\theta^T X}}{1 + e^{-\theta^T X}} \tag{6}$$

Para cada ponto de treinamento temos um vetor de características X e uma classe observada Y. A probabilidade da classe é $g\left(\theta^T X\right)$, se $y_k = 1$, ou é $1 - g\left(\boldsymbol{\theta}^T X\right)$, se $y_k = 0$. Na classificação do vetor X, devemos atribuir um valor y_k a partir do vetor de parâmetros θ de forma que a função de verossimilhança de θ seja maximizada. Para estimar o vetor de parâmetros θ, o log da função de verossimilhança pode ser maximizado utilizando o método do gradiente ascendente, conforme visto em Aggarwal [2014].

3- Trabalhos Relacionados

Este capítulo foi realizado a partir do desenvolvimento de um mapeamento sistemático [Grant and Booth, 2009] de trabalhos encontrados na literatura sobre seleção de atributos em dados educacionais. O objetivo foi entender qual é o estado da arte do tema proposto, perceber que tipos de trabalhos estão sendo realizados e identificar lacunas de conhecimento que possam ser preenchidas.

A pesquisa foi feita na base *Scopus*, com as seguintes *strings* de busca: TITLE-ABS-KEY ("*Educational data mining*" AND ("*Feature selection*" OR "*Dimensionality reduction*")), tendo o retorno de 119 artigos de revistas e conferências. Foram selecionados 34 artigos que tinham como objetivo principal a seleção de atributos em dados educacionais e utilizaram pelo menos uma das abordagens apresentadas na Seção 2.

A partir dos trabalhos encontrados, observou-se que existem vários métodos de SA usados em pesquisas na área de MDE. Muitos utilizam métodos de filtro com estatísticas de ranqueamento pela facilidade do uso [Ahmed et al., 2019; Govindasamy and Velmurugan, 2019; Jalota and Agrawal, 2021; Febro, 2019; Triayudi and Fitri, 2021; Muchuchuti et al., 2020]. Outros utilizam a abordagem *wrapper* com heurísticas que demandam mais poder computacional [Farissi et al., 2020; Santos et al., 2020; Zaffar et al., 2018a; Thaher et al., 2021]. Há ainda aqueles que utilizam recursos do próprio classificador na abordagem embutida [Hassan et al., 2019; Teodoro and Kappel, 2020; Gitinabard et al., 2018]. Além das formas tradicionais, há a possibilidade de combinações de métodos para potencializar a assertividade da escolha [Sokkhey and Okazaki, 2020]. Se a combinação for realizada com métodos de diferentes abordagens de SA, consideramos como abordagem híbrida para SA [Zaffar et al., 2018a, 2021; Saeed et al., 2021]. Os diferentes métodos e abordagens podem ser comparados a fim de compreender qual(is) deles seria(m) o(s) mais adequado(s) ao cenário específico do estudo [Ahmed et al., 2020; Garcia Torres et al., 2020; Hashemi et al., 2018; Memon et al., 2021].

A aplicação dos métodos de SA também pode variar, sendo aplicado a apenas um classificador [Teodoro and Kappel, 2020] ou a um conjunto de classificadores afim de compreender se há diferença de ganho de desempenho entre eles [Zaffar et al., 2018b; Ajibade et al., 2019; Jalota and Agrawal, 2021]. As bases de dados utilizadas nos trabalhos

encontrados são de diversas naturezas, como: questionários respondidos pelos alunos [Ahmed et al., 2019; Chaudhury and Tripathy, 2020], dados de plataformas *e-learning* [Govindasamy and Velmurugan, 2019; Jalota and Agrawal, 2021], bancos de dados de escolas [Abid et al., 2018] e departamentos de cursos universitários [Santos et al., 2020]. A Tabela 3 mostra trabalhos publicados que tiveram a SA em dados educacionais como foco, atuando no ensino básico ou superior, com bases de acesso aberto ou fechado para problemas de evasão e desempenho do aluno.

Autor	Filtro	Wrap.	Emb.	Híbrido	Escolaridade	Base	Problema
Ramaswami and Bhaskaran [2009]	x				Básico	Fechado	Desempenho
Gitinabard et al. [2018]			x		Superior	Fechado	Evasão
Gopalakrishnan et al. [2018]	x				Superior	Fechado	Evasão
Niu et al. [2018]			x		Superior	Fechado	Evasão
Punlumjeak and Rachburee [2015]	x				Superior	Fechado	Desempenho
Abid et al. [2018]	x				Básico	Aberto	Desempenho
Hashemi et al. [2018]	x	x			Superior	Fechado	Desempenho
Zaffar et al. [2018b]	x				Básico	Aberta	Desempenho
Febro [2019]	x				Superior	Fechada	Evasão
Govindasamy and Velmurugan [2019]	x				Superior	Fechado	Desempenho
Ahmed et al. [2019]	x				Básico	Fechado	Desempenho
Wafi et al. [2019]		x			Básico	Aberto	Desempenho
Hassan et al. [2019]			x		Superior	Fechado	Desempenho
Ajibade et al. [2019]	x	x			Básico	Aberta	Desempenho
Dimic et al. [2019]	x				Superior	Fechada	Desempenho
Das et al. [2020]	x				Superior	Fechada	Desempenho
Almasri et al. [2020]	x	x			Superior	Fechada	Desempenho
Chaudhury and Tripathy [2020]	x				Superior	Fechado	Desempenho
Enaro and Chakraborty [2020]	x				Básico	Aberta	Desempenho
Garcia Torres et al. [2020]		x			Superior	Fechada	Desempenho
Sokkhey and Okazaki [2020]	x				Básico	Fechado	Desempenho
Teodoro and Kappel [2020]			x		Superior	Aberta	Evasão
Farissi et al. [2020]		x			Básico	Aberta	Desempenho
Santos et al. [2020]		x			Superior	Fechada	Evasão
Muchuchuti et al. [2020]	x				Superior	Fechada	Desempenho
Ahmed et al. [2020]	x	x			Superior	Fechada	Desempenho
Urbina-Nájera et al. [2020]	x				Superior	Fechada	Evasão
Triayudi and Fitri [2021]	x				Superior	Fechado	Desempenho
Thaher et al. [2021]		x			Básico	Aberto	Desempenho
Alam et al. [2021]	x	x			Básico	Aberto	Desempenho
Saeed et al. [2021]				x	Básico	Aberto	Desempenho
Memon et al. [2021]	x	x			Básico	Aberto	Desempenho
Zaffar et al. [2021]				x	Básico	Aberto	Desempenho
Jalota and Agrawal [2021]	x	x			Básico	Aberta	Desempenho

Tabela 3 – Trabalhos que usaram seleção de atributos em dados educacionais.

Os métodos de filtro costumam ser mais simples e rápidos, por isso há mais trabalhos na literatura que usam essa abordagem para seleção de atributos. Gopalakrishnan et al. [2018] quiseram entender quais eram os principais fatores que impactavam no abandono de curso em uma universidade nos Estados Unidos. Eles propuseram testes com algoritmos de filtro do QQ, ganho de informação, correlação, *relief*, máxima relevância e mínima redundância (mRMR) e teste de Kruskal-Wallise. Cada um dos algoritmos gerou um resultado diferente no ranqueamento de atributos, sendo o nível de escolaridade da mãe e o nível inicial de habilidade em matemática do aluno os mais

importantes. Febro [2019] também usou de alguns desses mesmos métodos de filtro para mostrar que a renda familiar e a nota no vestibular são fatores importantes no índice de abandono de alunos universitários nas Filipinas. Enquanto no México, Urbina-Nájera et al. [2020] identificaram através de métodos de filtro que a falta de aconselhamento, falta de um ambiente adequado e de acompanhamento acadêmico eram as três principais causas da evasão em alunos universitários.

Ainda sobre evasão universitária, Dimic et al. [2019] utilizaram dados de uma plataforma de gestão de notas em uma faculdade de engenharia na Sérvia para prever o desempenho do aluno com os métodos de filtro do QQ, *relief*, ganho de informação e baseado em correlação. Outros autores como Ramaswami and Bhaskaran [2009] demonstraram ganho em tempo computacional na classificação do desempenho escolar do aluno, ao usar as técnicas de QQ, ganho de informação, *relief*, taxa do ganho de informação e incerteza simétrica em dados demográficos e socieconômicos vindos de questionários de estudantes do ensino médio na Índia.

Outros trabalhos [Rachburee and Punlumjeak, 2015; Enaro and Chakraborty, 2020; Das et al., 2020] fizeram diversas combinações entre classificadores e métodos de seleção de filtro, para entender qual método se adapta melhor ao algoritmo testado. Sokkhey and Okazaki [2020] avaliaram outro tipo de combinação, um método de seleção com dois algoritmos de ranqueamento, o QQ e a Informação Mútua (IM). O nome dado a nova técnica foi CHIMI e foi aplicada para identificar os fatores mais relevantes no desempenho escolar.

A abordagem de *wrapper* foi utilizada por Santos et al. [2020] com dados de um sistema de gestão de notas de uma universidade brasileira, usando informações pessoais e de desempenho acadêmico dos alunos. Eles propuseram um classificador de árvore de decisão otimizado com o auxílio do AG para prever o risco de evasão dos alunos.

No contexto do desempenho escolar, a abordagem de *wrapper* foi utilizada por Farissi et al. [2020], também com o uso do AG como algoritmo construtor dos conjuntos de atributos. O trabalho teve por objetivo classificar o desempenho de alunos de uma escola em baixo, médio e alto através de atributos demográficos, comportamentais e de formação acadêmica. Os resultados sugerem que houve melhora nas métricas de avaliação dos classificadores com a seleção de atributos feita pelo AG.

Ainda com o uso de meta-heurísticas em classificação de desempenho escolar, Wafi et al. [2019] fizeram a combinação do AG e com o classificador KNN, o que de-

monstrou melhora significativa da acurácia em relação ao modelo sem SA. Contudo foi observado que o tempo de processamento da SA pode ser alto dependendo do volume de dados. Almasri et al. [2020] avaliaram diferentes algoritmos de indução de amostragem de conjuntos usados na abordagem *wrapper* para prever o desempenho de alunos de uma universidade na Jordânia, tendo os melhores resultados com *Bat Search*, *Harmony Search* e *Ant Search*.

A abordagem embutida foi utilizada em Gitinabard et al. [2018]; Niu et al. [2018] por meio do classificador AD para definir os principais fatores que levam a conclusão ou desistência de cursos livres disponíveis em plataformas *online*. Gitinabard et al. [2018] identificaram que assistir aos vídeos é um atributo importante para o aluno conseguir o certificado de cursos ofertados na plataforma online `Coursera` e `EdX`, enquanto que Niu et al. [2018] descobriram que a diferença nos intervalos de acesso é o principal atributo para prever o abandono de cinco cursos multidisciplinares *online* da plataforma `Icourse163`.

O ranqueamento dos atributos mais relevantes para o desempenho de alunos do classificador RF foi usado por Hassan et al. [2019] em uma base de dados de um curso *e-learning* de uma universidade pública de engenharia na Malásia, tendo como principal atributo encontrado a quantidade de vezes que curso foi visualizado. Teodoro and Kappel [2020] também fizeram a seleção de atributos de forma embutida a partir da RF com a base de dados aberta do CES, fornecida pelo governo brasileiro, com o objetivo de encontrar quais são as principais características que levam a evasão no ensino superior. O resultado sugeriu que os atributos de idade, participação em atividades extracurriculares e carga horária total do curso foram os mais relevantes na classificação de evasão.

A abordagem híbrida representa o uso sequencial de duas ou mais abordagens de SA para alavancar o custo benefício de desempenho computacional e precisão de classificação. Ela foi usada por Zaffar et al. [2021] em quatro bases de dados de referência em MDE disponíveis ao público para prever o desempenho escolar de alunos. Foram usados o filtro do QQ e o algoritmo otimizador *Sequential Forward Selection* (SFS) como método de *wrapper*. De forma geral, a combinação híbrida proposta superou os resultados em relação a métodos usados na literatura de filtro e de *wrapper*. Outro trabalho que usou uma abordagem híbrida foi o de Saeed et al. [2021]. Foram usados os métodos de filtros de correlação e informação mútua e o otimizador de enxame de partículas, do

inglês *Particle Swarm Optimization* (PSO). Neste caso os autores usaram uma base de dados disponível publicamente referente a um questionário de 36 perguntas sobre o nível de satisfação de professores durante o período de aulas remotas decorrente da pandemia de COVID-19. Apesar do alvo da classificação não ser diretamente sobre o desempenho do aluno, indiretamente esse tema faz parte das dificuldades enfrentadas no ambiente educacional e reflete o desempenho escolar. Para esse experimento, a proposta híbrida também conseguiu superar os resultados de precisão alcançados por técnicas clássicas de SA.

Outros trabalhos testaram diferentes combinações de abordagens de SA e classificadores para fazer comparações entre os ganhos obtidos em cada abordagem. Ahmed et al. [2020] estudaram o desempenho acadêmico de alunos com informações do departamento de ciência da computação de uma universidade em Bangladesh. Foram feitas combinações da abordagem *wrapper* e técnicas de filtro com os classificadores *k*-vizinho mais próximos (KNN), *Naive Bayes* (NB), *Bagging*, RF e AD, onde a combinação AG + KNN foi a que apresentou melhor desempenho. Garcia Torres et al. [2020] também utilizaram a abordagem de filtro de correlação e duas meta-heurísticas para gerar os conjuntos na abordagem de *wrapper*. A base de dados usada foi de uma plataforma *online* de gestão de informação do curso de física em uma universidade na Espanha, a fim de predizer se o aluno seria aprovado no curso.

Ajibade et al. [2019] e Jalota and Agrawal [2021] utilizaram a mesma base escolar de desempenho de alunos para traçar comparativos entre técnicas de filtro, *wrapper*. No primeiro trabalho, Ajibade et al. [2019] usaram os algoritmos de evolução diferencial, SFS e *Sequential Backward Selection* (SBS) como opções de algoritmos construtores de conjuntos para a abordagem *wrapper*, e concluíram que esses métodos foram mais eficientes do que o uso da correlação. Enquanto Jalota and Agrawal [2021] detectaram que o modelo de NB teve melhor desempenho aliado a abordagem *wrapper*, e a árvore de decisão se adaptou melhor a abordagem de filtro com correlação para seleção de atributos.

Hashemi et al. [2018] também utilizaram as abordagens filtro e *wrapper* de forma comparativa para prever a aceitação em universidades a partir de um exame nacional de educação do Irã, obtendo resultados muito semelhantes para as duas abordagens. Enquanto Punlumjeak and Rachburee [2015] analisaram a SA para prever o desempenho de alunos em uma universidade de engenharia na Tailândia por meio de dados do

departamento, utilizando as abordagens *wrapper*, filtro e embutida, com o classificador *Support Vector Machine* (SVM). Concluíram que o método de filtro de máxima Relevância e Mínima Redundância (mRMR) apresentou o melhor resultado. Por último, com o objetivo de analisar e propor alternativas às possíveis causas de baixo desempenho do aluno, o trabalho de Alam et al. [2021] comparou os conjuntos de atributos selecionados por métodos das abordagens filtro e *wrapper* a fim de entender se as duas soluções levam a resultados semelhantes. Concluiu que a localização das escolas em relação a zonas urbanas e rurais eram um fator importante na previsão do desempenho do aluno para a maioria das soluções.

Esses estudos apontam que as técnicas de SA são capazes de melhorar o desempenho de algoritmos de aprendizado de máquina no contexto da MDE e ajudam a compreender melhor quais são os principais fatores que levam a problemas de evasão e baixo desempenho escolar. Em relação as bases de dados usadas, apenas um artigo com o tema de evasão usou uma base aberta e apenas com uma das abordagens de seleção de atributos descrita. Além disso, nenhum trabalho propôs uma abordagem mais específica para o cenário educacional, limitando-se apenas a métodos já consolidados de SA e/ou a combinação deles.

Essa dissertação amplia o escopo das aplicações de diferentes metodologias de SA para evasão em uma base de dados pública e nacional de alta relevância para o ensino superior brasileiro com o objetivo de encontrar os principais fatores que impactam na classificação da evasão no ensino superior brasileiro por meio de técnicas de seleção de atributos. Dentre as metodologias usadas está o método proposto FlexAG, que leva em consideração o agrupamento de atributos educacionais, com a intenção de descobrir se priorizar determinados grupos impacta positivamente na busca pelo conjunto otimizado.

4- Metodologia

Este capítulo apresenta a metodologia de pesquisa exploratória desenvolvida para a análise comparativa das técnicas de Seleção de Atributos (SA) nos dados do Censo da Educação Superior (CES) do Brasil [BRASIL, 2020] com o objetivo encontrar os principais fatores que impactam na classificação de evasão no ensino superior brasileiro por meio de técnicas de seleção de atributos. Foram usados métodos clássicos de SA e uma proposta de flexibilização da abordagem de *wrapper* com AG (FlexAG) que possibilita o especialista em educação adicionar informações específicas ao processo.

A Figura 9 ilustra o diagrama com as etapas da metodologia adotada neste trabalho. O esquema começa a partir da entrada do conjunto de dados do CES utilizado nos experimentos, e é seguida pelo mecanismo de pré-processamento. Nele a primeira etapa consiste no *Tratamento dos Dados* (1) com uso de metodologias de integração, limpeza, transformação e normalização dos dados. Essa etapa é necessária para converter os dados brutos em dados prontos para serem usados na SA e processados na classificação.

Esses dados foram divididos em dois conjuntos, cada um com 50% das amostras. Um conjunto foi utilizado na etapa de *Seleção de Atributos* (2), que também faz parte do pré-processamento. Ela consiste em testar diferentes abordagens para selecionar o melhor conjunto de atributos para ser usado na classificação de forma que otimize recursos computacionais de tempo e memória, aumente a precisão do modelo e facilite a análise do problema. A outra metade das amostras da base de dados foi utilizada na etapa de classificação que recebe como entrada o conjunto de atributos selecionados na etapa de SA.

A etapa de *Classificação* (3) utiliza os mesmos classificadores usados na abordagem embutida e teve como saída as métricas de desempenho de classificação para cada conjunto de atributos testado. Essas informações foram comparadas e avaliadas na etapa de *Avaliação Experimental* (4), permitindo concluir a cerca de quais combinações de técnicas e atributos tiveram maior destaque.

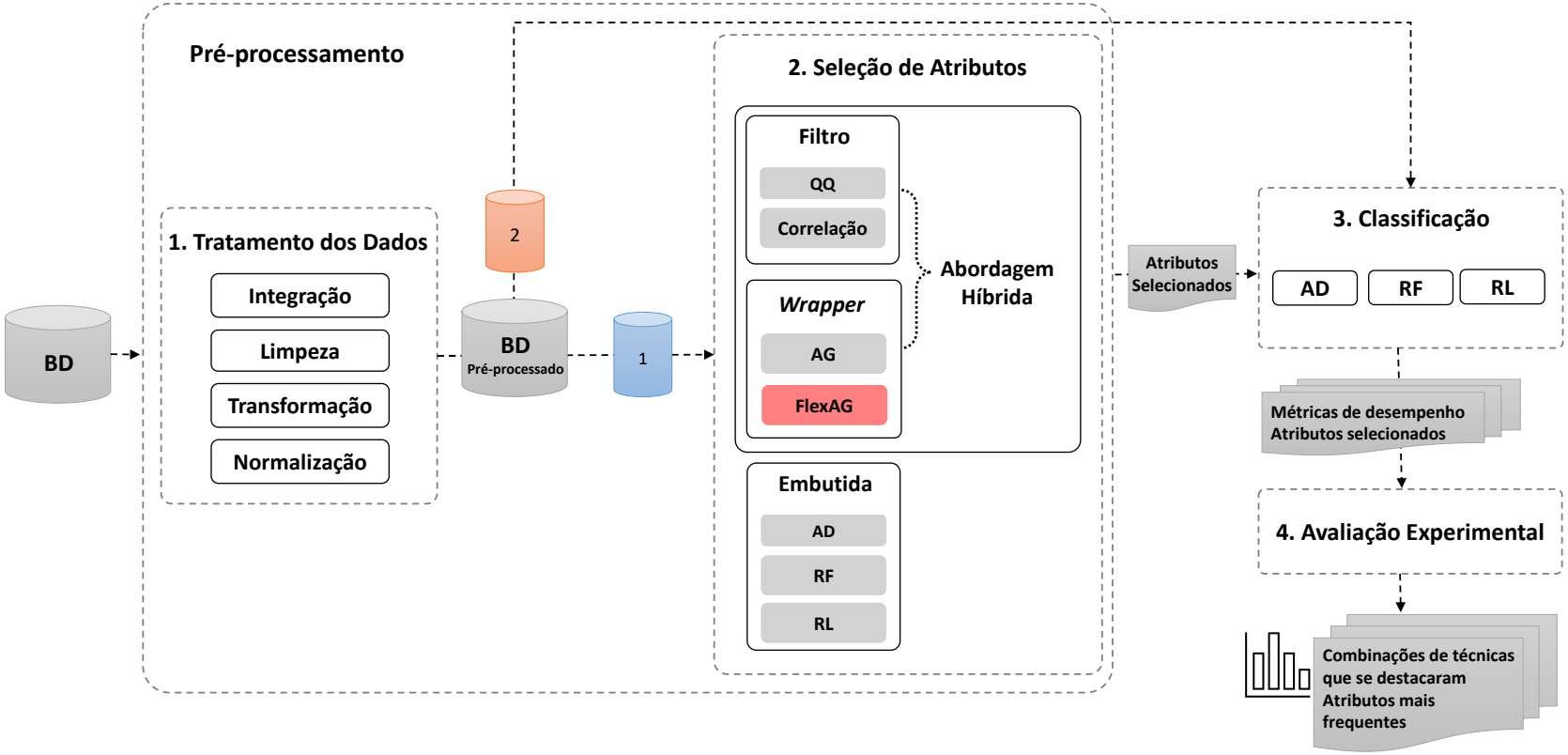

Figura 9 – Diagrama da Metodologia Experimental adotada neste trabalho.

As etapas descritas estão melhor detalhadas nas próximas seções. Na Seção 4.1 está a descrição da base de dados e na Seção 4.2 encontra-se as informações referentes as técnicas de pré-processamento. Embora a SA possa ser considerada uma etapa do pré-processamento, ela está apresentada de forma separada na Seção 4.3 pois tem protagonismo no objetivo desse estudo. Por fim, a Seção 4.5 apresenta os detalhes da classificação e a avaliação experimental realizada.

4.1- Base de dados

O CES é um relatório nacional e anual, que foi criado com a finalidade de coletar informações do ensino superior no Brasil. Desde 2000, os dados são enviados por todas as IES por meio de questionário eletrônico no site do INEP [IBGE, 2021]. As informações são separadas em tabelas referentes aos alunos, instituições, docentes, cursos e locais de oferta, juntamente com notas estatísticas que destacam determinados indicadores (percentual de docentes por modalidade de ensino (presencial/EaD), quantidade de matrículas por categoria administrativa (público/privado), número de vagas por cursos de graduação, entre outros).

O objetivo é oferecer informações confiáveis para conhecer e acompanhar os cursos de educação superior, disponibilizando dados que possam contribuir para programas de melhoria e acompanhamento de tendências do setor educacional tanto no âmbito público como privado [INEP, 2021]. Diversos artigos publicados [Teodoro and Kappel, 2020; Barros, 2015; Campos et al., 2018, 2019; de CAMPOS et al., 2016; da Silva et al., 2019; Canedo et al., 2019] usaram o CES como fonte de pesquisas com intuito de tentar responder perguntas sobre a qualidade da educação brasileira e seus problemas e, então, indicar estratégias de planejamento para o setor. Esses trabalhos são usados no Capítulo 5 para fundamentar e comparar os resultados do experimento.

As informações são referentes à infraestrutura das IES, vagas oferecidas, candidatos, matrículas, ingressantes, concluintes e docentes em diferentes categorias administrativas, sendo que alguns indicadores estão preenchidos apenas para alguns tipos de cursos, como por exemplo, informações de financiamento para IES privadas e de local de oferta para cursos presenciais [IBGE, 2021; INEP, 2021]. Até o presente momento,

estão disponíveis dados de 1995 a 2020[1]. Os resultados apresentados no Capítulo 5 são referentes ao ano de 2017, que contém $11.589.194$ registros relacionados às matrículas ativas naquele ano.

A Figura 10 mostra as relações e chaves existentes entre as tabelas disponíveis. A base é formada por seis tabelas: DM CURSO (114 atributos), DM IES (47 atributos), TABELA AUXILIAR OCDE (8 atributos), DM DOCENTES (41 atributos), DM LOCAL OFERTA (50 atributos) e DM ALUNO (108 atributos) que armazenam informações dos cursos, das instituições de ensino, da nomenclatura internacional da área de cada curso, dos docentes, do local de oferta do curso e dos alunos, respectivamente. Cada aluno na tabela DM ALUNOS está associado a um curso (DM CURSO), a uma instituição de ensino (DM IES) e a um código internacional da área do curso (TABELA AUXILIAR OCDE). Cada docente na tabela DM DOCENTE está vinculado a uma instituição de ensino (DM IES) e cada local de oferta (DM LOCAL OFERTA) possui vários cursos associados (DM CURSO).

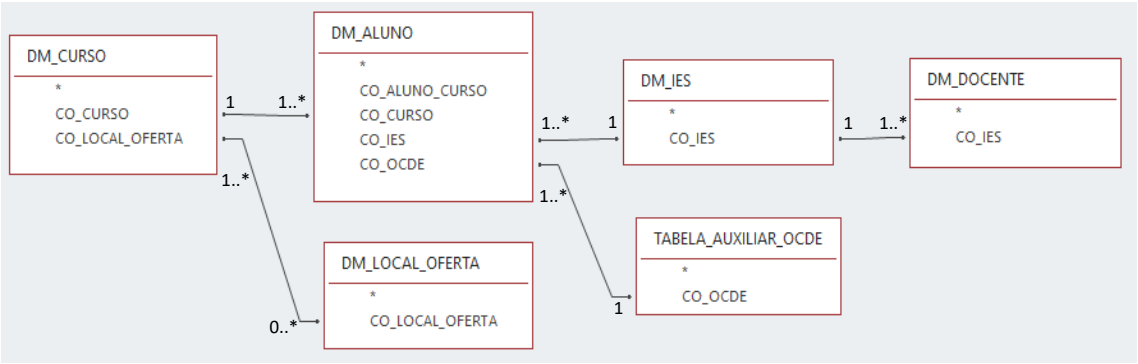

Figura 10 – Esquema de Relacionamentos entre as Tabelas do CES de 2017.

No Anexo A estão descritos os metadados de todos os atributos utilizados nos experimentos desse trabalho, os demais foram retirados ou modificados na etapa de pré-processamento, detalhada na Seção 4.2.

[1]Bases disponíveis em: `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior`. Em 20/04/2022 o formato das bases foram alterados para se adequar a LGPD (Lei Geral de Proteção de Dados), sendo assim uma pequena amostra das bases no formato original foi disponibilizado link `https://github.com/daniellefalbuquerque/selecao_de_atributos_CES.git` devido ao extenso tamanho das bases

4.2- Pré-processamento

No mundo real, as bases de dados podem apresentar diversos problemas que reduzem a qualidade da informação que será extraída dela. Alguns aspectos gerais dos dados devem ser avaliados previamente como precisão, completude, consistência, credibilidade, interpretabilidade e periodicidade de atualização [Han et al., 2012]. Esses aspectos garantem que a qualidade dos dados esteja apropriada para o estudo. Neste trabalho foram realizadas as principais etapas de pré-processamento de dados: integração, limpeza, transformação de dados, normalização e redução [Han et al., 2012], com o objetivo de torná-los mais adequados e confiáveis para o uso das metodologias de mineração de dados.

A integração dos dados entre as tabelas disponíveis foi realizada com base nas chaves descritas na Figura 10. Nos resultados apresentados no Capítulo 5 não são contemplados os atributos da tabela com as informações do local de oferta das vagas (DM LOCAL OFERTA). Esses atributos não foram usados pois só estão disponíveis para cursos presenciais, de forma que haveria muitos dados faltantes para a modalidade EaD.

As informações da tabela de docentes (DM DOCENTE) foram agrupadas para obtenção de novos atributos por instituição de ensino (campo CO IES) expressos em proporção, como % de professores com doutorado, % de professores com mestrado, % de professores com dedicação exclusiva, % que trabalham em pesquisa, extensão e EaD, % do sexo masculino, % de professores negros e pardos e % de professores substitutos. Da tabela de nomenclaturas internacionais (TABELA AUXILIAR OCDE), foram usados apenas os campos de nomes da área geral e da área específica. Alguns critérios foram adotados para descartar determinados atributos das tabelas de IES, CURSO e ALUNO, são eles:

- Os atributos que estavam presentes em várias tabelas foram trazidos apenas uma vez, como por exemplo TP_CATEGORIA_ADMINISTRATIVA, que estava presente na tabela aluno e de IES.
- Foram mantidos os atributos que representavam a totalidade de atributos específicos, enquanto os específicos foram retirados. Por exemplo, a quantidade total de inscritos foi mantida enquanto os campos de detalhamento como QT_INSC_ANUAL_EAD e

QT_INSC_ANUAL_MATUTINO foram excluídos.

- Foram descartados atributos numéricos que tinham mais de 30% de valores ausentes, por não poder serem encaixados em categorias e pela dificuldade de imputar novos dados.
- Os atributos que naturalmente traziam informação direta sobre a classe não foram considerados, como IN CONCLUINTE e CARGA HORARIA CURSADA.
- Foram desconsiderados atributos que pudessem trazer algum tipo de ambiguidade após o tratamento de dados nulos. Como por exemplo o campo IN_MOBILIDADE_ACADEMICA, onde de acordo com a tabela de metadados fornecida não há informação disponível para os alunos de universidade pública federal.
- Foram desconsiderados atributos que incorporaram funções de novos atributos criados, como o detalhamento de receitas e despesas.
- Foram desconsiderados atributos referentes apenas a coleta dos dados, como o semestre no qual o dado foi coletado.
- Foram desconsiderados atributos que representam a mesma informação, como por exemplo, CO_IES e NO_IES, no qual o código representa a mesma identificação que o nome.

Além da junção das tabelas do CES, foram incorporadas as informações de região do curso para as amostras referentes a cursos presenciais, podendo ser Norte, Nordeste, Sul, Sudeste e Centro-Oeste e também as informações de latitude e longitude dos municípios referentes ao local do curso, com o objetivo de transformar o atributo de localização de categórico para numérico. No caso dos cursos EaD, foi criada uma categoria EaD para as regiões, enquanto a latitude e a longitude foram consideradas como zero. Por último, foi realizado um filtro para retirar da base os alunos que estavam em situação de “cursando”, “falecido” e “transferido”, para que a base pudesse conter apenas aqueles indivíduos que abandonaram, categorizados como “matrícula trancada” e “desvinculado”, além daqueles que obtiveram sucesso, nomeados como “formados”.

Na limpeza dos dados, o tratamento de dados ausentes foi um outro dificultador a ser superado. Para isso, foram adotadas as seguintes estratégias:

- Alguns atributos foram excluídos por conterem mais de 30% de dados ausentes, a lista desses atributos está na Tabela 4.

- Algumas amostras foram excluídas por conter dados ausentes em atributos que não foram eliminados pois tinham menos de 30% de dados ausentes. A lista dos atributos que não foram excluídos mas que tiveram amostras excluídas quando o valor estava ausente pode ser vista na Tabela 5.

- Em alguns atributos os dados ausentes se transformaram em uma nova categoria. Por exemplo, no caso do atributo TP_TURNO que já continha quatro categorias (integral, matutino, vespertino e noturno), os dados ausentes foram atribuídos a uma nova categoria chamada EaD, conforme mencionado no dicionário de dados.

- Em alguns atributos os dados ausentes foram agrupados em categorias já existentes, de acordo com o contexto explicado na tabela de metadados fornecida pelo INEP. Por exemplo, para o campo IN_RESERVA_ETNICO, o valor 0 significa não ter reserva de vagas do tipo étnica, enquanto o valor 1 representa a presença desse modelo de reserva. Neste caso, os dados ausentes significam não ter nenhum modelo de reserva de vagas. Com isso, os dados ausentes foram preenchidos com o valor 0.

Colunas excluídas	
Atributos	% DE AUSENTES
CO_UF_NASCIMENTO	30%
CO_MUNICIPIO_NASCIMENTO	30%
IN_INGRESSO_OUTRO_TIPO_SELECAO	100%
IN_INGRESSO_OUTRA_FORMA	100%
IN_INGRESSO_PROCESSO_SELETIVO	100%
CO_CURSO_POLO	74%

Tabela 4 – Atributos excluídos devido à quantidade elevada de dados ausentes.

A etapa de transformação de dados foi realizada para que os dados pudessem estar em formatos apropriados para o uso das técnicas de SA e de classificação. A criação de novos atributos foi feita por meio de combinações de outros atributos com operações de soma e porcentagem, por exemplo, o percentual de despesa com pesquisa pela IES, foi calculado a partir do somatório de despesas totais frente ao gasto apenas com pesquisa. Já o campo de data de ingresso foi transformado em binário, de forma que 0 representa que o aluno entrou no primeiro semestre e 1 que ele entrou no segundo semestre.

Linhas excluídas	
Atributo	% de ausentes
IN_GRATUITO	0,13%
NO_OCDE_AREA_GERAL	0,44%
QT_MATRICULA_TOTAL	0,13%
QT_CONCLUINTE_TOTAL	0,13%
QT_INGRESSO_TOTAL	0,13%
QT_INGRESSO_VAGA_NOVA	0,13%
QT_VAGA_TOTAL	0,13%

Tabela 5 – Atributos que tinham poucos dados ausentes e por isso apenas as linhas foram excluídas.

A codificação é uma forma de transformar atributos categóricos em numéricos, para que possam ser processados pelos métodos de MDE. Na base do CES, a grande maioria dos campos já estavam codificados, por isso ela foi realizada de forma complementar de duas maneiras: transformação binária e transformação simples. A primeira transformação cria uma nova coluna binária para cada categoria do atributo e foi usada principalmente na tabela de docentes para facilitar a criação de novos atributos. Já a segunda transformação foi usada em campos que estavam sendo representados por um código de muitos caracteres, passando a ser codificados de 1 a n, onde n é a quantidade de categorias. Esse processo aconteceu com os atributos referentes à unidade federativa, à área geral do curso e à IES.

Outro processo importante no pré-processamento é a normalização dos dados. Ela é capaz de facilitar e até mesmo viabilizar o uso de determinadas técnicas de mineração de dados quando os atributos tiverem intervalos de valores muito diferentes. Os dados foram normalizados com o método MinMax, apresentado na Equação 7 [Han et al., 2012], para manter todos os valores da base entre 0 e 1. Na Equação vemos que v'_i é o novo valor, v_i é o valor antigo, min_A é o valor mínimo do atributo, max_A é o valor máximo, e $newmax_A$ e $newmin_A$ são os valores máximo e mínimo, respectivamente, do novo intervalo normalizado que vai de 0 a 1.

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A \tag{7}$$

A base de dados tratada foi dividida em 50% para ser usada na SA e 50% para validação dos conjuntos selecionados através da classificação com validação cruzada. Apesar da etapa de SA fazer parte do pré-processamento, ela possui maior destaque no objetivo desse trabalho e será discutida com mais detalhes na próxima seção.

4.3- Seleção de atributos

A SA é o foco desse trabalho e representa uma importante técnica de redução dos dados. Como visto no Capítulo 2, ela tem papel fundamental em bases de dados extensas com muitos atributos, pois consegue filtrar apenas a informação que será relevante para a descoberta de conhecimento. Isso reduz o custo computacional no contexto do uso de algoritmos de aprendizado de máquina e de armazenamento de dados, tornando o modelo mais simples.

Sendo assim, as seguintes abordagens de SA são usadas: (i) filtro com os métodos de QQ e correlação, (ii) *wrapper* com o uso do AG e a variação proposta neste trabalho (FlexAG), (iii) embutida com os classificadores AD, RF e RL, e (iv) híbrida com combinações de filtro e *wrapper*. A abordagem FlexAG está detalhada na Seção 4.4. Vale mencionar que os classificadores AD, RF e RL são empregados por possuírem capacidade de ranquear os melhores atributos durante o processo de classificação.

Os atributos selecionados serão discutidos e comparados a fim gerar conclusões a cerca das principais causas de evasão. Essa análise também será realizada para recortes contextualizados da base de dados, ou seja, para entender a diferença dos atributos selecionados nas instituições públicas ou privadas, e em cursos presenciais ou à distância.

4.4- Abordagem FlexAG

A abordagem FlexAG é uma variação da abordagem *wrapper* com o uso do AG. Com ela é possível investigar, tentando inferir, se grupos de atributos são mais importantes que outros na classificação de evasão dos alunos. Para isso, pode-se categorizar os atributos em grupos de acordo com a natureza deles, como por exemplo, atributos referentes a dados pessoais, do curso, da instituição, financeiros e de comportamento do aluno. Esses agrupamentos são incorporados na função objetivo do AG, fazendo com que o método tenha uma parte flexível e adaptada ao contexto educacional.

Essa abordagem utiliza o AG, conforme já mencionado, para selecionar os atribu-

tos criando conjuntos que serão avaliados pela função objetivo FO proposta na Equação 8. Cada conjunto é um indivíduo representado por um vetor binário b, onde cada elemento b_i indica a presença (1) ou não (0) de um atributo i, e N é o número total de atributos do conjunto de dados ($|b| = N$). A cada geração acontecem as operações de cruzamento e mutação, que fazem variação evolutiva de cada indivíduo.

As prioridades dos grupos de atributos são atribuídas por meio de ponderamentos (pesos) na FO a ser maximizada. Os pesos são representados pelo vetor p, onde cada elemento p_i indica o peso da prioridade atribuída ao grupo no qual o atributo i pertence. Dessa forma, é possível dar prioridade para um determinado grupo, como por exemplo, para o grupo referente às informações financeiras, e entender se aqueles atributos tem um impacto maior no processo de SA e de otimização da classificação.

A Equação 8 é composta por três componentes, sendo que os dois primeiros representam a função objetivo padrão no uso do AG para SA. São eles: 1) a métrica de avaliação da classificação com os atributos selecionados (f_1); 2) uma penalização referente à quantidade de atributos selecionados (num_select); e 3) uma parcela com os ponderamentos das prioridades dos grupos de atributos, *i.e.*, o somatório da multiplicação de p_i por b_i, onde $\gamma = \sum_{i=1}^{N} p_i$ é um fator de normalização. Portanto, as duas primeiras componentes referem-se à aplicação padrão do AG (*i.e.* que usualmente é realizada na SA) e a terceira representa a parte flexível para priorização de grupos de atributos.

$$FO = f_1 - \frac{num\_select}{N} + \frac{1}{\gamma} \times \sum_{i=1}^{N} p_i \times b_i \tag{8}$$

O FlexAG permite que especialistas em educação tenham maior flexibilidade para inserirem informações empíricas no processo de SA de acordo com as diversas necessidades e realidades que o ambiente educacional apresenta, permitindo que características intrínsecas de cada ambiente escolar sejam melhor exploradas.

Um exemplo de aplicação seria o especialista em educação pode dizer que a localidade onde o estudante reside é mais importante para a evasão do estudante do que a participação dele na aula. A função objetivo atribuiria, prioridade maior para localidade de residência (peso 0,7) e prioridade menor para atributos de participação em aula (peso 0,3). Este caso exemplificaria alunos que precisam acordar muito cedo para chegar na instituição de ensino e acabam chegando atrasados nas aulas e desistindo do curso.

Neste trabalho, o método FlexAG foi utilizado primeiramente em uma base de

dados menor e pública de desempenho escolar a fim de validar se o método conseguiria atingir resultados significantes. Após essa validação, o método foi implementado na base de dados do CES. A próxima seção aborda as etapas de classificação e avaliação experimental.

4.5- Classificação e Avaliação Experimental

Existe uma tendência de que os dados reduzidos por técnicas de SA tenham seu desempenho preditivo melhorado, visto que os atributos que poderiam confundir o modelo são retirados. Diante disto, espera-se encontrar essa melhoria nas métricas de avaliação de desempenho do modelo de classificação. No entanto, este resultado pode ser influenciado pelo tipo de classificador empregado. Para avaliar isso, os experimentos realizados buscam testar diferentes conjuntos de atributos selecionados pelas técnicas de SA adotadas com diferentes classificadores.

Na etapa de classificação, os dados foram submetidos a validação cruzada estratificada para evitar *overfitting* e reduzir o efeito do desbalanceamento de dados. O método de validação cruzada usado foi o k-*fold* [2] com $k = 10$. Os classificadores usados foram AD, RF e RL. Outros classificadores como Rede Neural (RN), NB e KNN não foram utilizados com a base do CES com o intuito de manter o uso dos mesmos classificadores em todos experimentos realizados, de forma que classificadores que não realizam a SA na abordagem embutida não puderam ser considerados. No entanto, esses classificadores foram utilizados na validação do método de FlexAG com a base menor de desempenho escolar, pois nesse caso não haveria comparação com a abordagem embutida.

Para a análise dos resultados, a média e o desvio padrão de métricas de classificação obtidas por meio da validação cruzada foram comparados. As métricas usadas estão descritas na Seção 5.2. Além das métricas de classificação, a quantidade de atributos selecionados e o tempo computacional também são utilizados como métricas de desempenho para os resultados. Os tempos avaliados estão associados ao tempo gasto pela SA e à média de tempo da classificação em cada uma das 10 divisões do k-*fold*.

O Algoritmo 1 descreve o processo de validação dos resultados para os testes

[2] Este valor foi o mais utilizado na literatura relacionada consultada.

realizados com os diferentes métodos de SA e classificadores. Os parâmetros de entrada são: a base de dados d, os conjuntos das técnicas de seleção de atributos SA, definido como $SA = \{sa_1, sa_2, \ldots, sa_n\}$ e os métodos de aprendizado de máquina AM, definido como $AM = \{am_1, am_2, \ldots, am_m\}$. A Tabela 6 descreve os parâmetros usados.

Parâmetro	**Descrição**	**Valores**
d	Base de dados do CES de 2017	BR: base completa PUB: base parcial com instituições públicas PRIV: base parcial com instituições privadas PRES: base parcial com cursos presenciais EAD: base parcial com cursos EaD
sa	Técnica de seleção de atributos	QQ: Qui-Quadrado[1] CORRE: Correlação[1] AG: Algoritmo Genético EMB: Abordagem embutida[1,2] QQ_AG: método híbrido[1] com QQ e AG CORRE_AG: método híbrido[1] com CORRE e AG FlexAG: AG proposto CORRE_FlexAG: AG proposto com filtro[1]
am	Modelo de aprendizado de máquina	AD: árvore de decisão RF: *random forest* RL: regressão logística

Tabela 6 – Parâmetros da Metodologia.

[1] No Capítulo 5 as referências aos métodos de filtro, embutido, híbrido e híbrido com FlexAG aparecem com um valor ao lado que corresponde ao corte do ranqueamento usado.

[2] O método embutido será realizado de acordo com o classificador usado no processo de classificação, que pode ser AD, RF, RL.

O algoritmo inicia-se com o pré-processamento dos dados (Linha 1), que resulta em duas novas bases dd_1 e dd_2, cada uma com 50% das amostras da base de dados de entrada d. Para cada técnica de seleção de atributos i em SA (Linha 2), a base dd_1 é usada para gerar o conjunto A_i de atributos selecionados (Linhas 3). A seguir, cada método de aprendizado de máquina j em AM (Linha 4) é aplicado para classificar os registros da base de dados dd_2 usando o conjunto de atributos A_i. Este processo tem como saída as métricas de desempenho para cada classificador (Linha 5). Os resultados obtidos para cada classificador no conjunto de atributos A_i são, então, unidos na variável R_i (Linha 6). No final, todos os conjuntos de atributos selecionados (A) e os resultados obtidos a partir deles (R) são retornados pelo algoritmo (Linha 7). No próximo capítulo estes resultados serão avaliados e comparados, a fim de compreender as combinações de técnicas e os conjuntos de atributos que se destacaram em desempenho de classificação,

redução do volume de dados e tempo de processamento.

Algoritmo 1 – Metodologia Geral (d, SA, AM)

```
1 (dd_1, dd_2) ← tratamento_dos_dados(d);
2 for i ← 1 to |SA| do
3 |   A_i ← selecao_dos_atributos(dd_1, sa_i);
4 |   for j ← 1 to |AM| do
5 |   |   resultado_j ← classificacao(dd_2, A_i, am_j);
6 |   |   R_i ← R_i ∪ resultado_j;
7 return (A, R)
```

5- Avaliação Experimental

Este capítulo apresenta os resultados obtidos nesta dissertação de mestrado. A Seção 5.1 descreve detalhes de implementação dos algoritmos adotados na metodologia e a Seção 5.2 aborda as métricas de desempenho usadas na avaliação dos resultados. A Seção 5.3 faz uma análise exploratória dos dados do CES após a etapa de tratamento de dados, expondo os principais fatores que serão avaliados separadamente nos experimentos. A partir dela, foi possível extrair algumas percepções da evasão para diferentes cenários da educação superior brasileira, especificamente em instituições públicas *versus* privadas, e em cursos nas modalidades presenciais *versus* EaD. A Seção 5.4 apresenta os resultados obtidos a partir da metodologia proposta para a análise comparativa das combinações de métodos de SA e classificadores referente ao cenário geral da base do CES e aos cenários específicos: estudantes de IES públicas e privadas, e que frequentam cursos presenciais ou de EaD. A Seção 5.5 apresenta os resultados obtidos pela abordagem FlexAG proposta. Por último, a Seção 5.6 traz o modelo de classificação da evasão para os dados do CES de 2018 e 2019, afim de compreender se os conjuntos de atributos selecionados pelas combinações de SA e classificadores que alcançaram os melhores resultados do estudo conseguem manter resultados relevantes em bases futuras.

5.1- Parâmetros

Nesta seção são descritos os parâmetros e bibliotecas usadas na implementação deste trabalho, sumarizados na Tabela 7. Os experimentos foram implementados em Python, versão 3, e realizados em uma máquina Intel(R) Xeon(R) Gold 5120 CPU 2.20GHz, com 28 núcleos e $192GB$ de memória. As bibliotecas utilizadas foram *pandas*[1], *scikit-*

[1] `https://pandas.pydata.org`

learn[2] e *deap*[3][4].

Os parâmetros usados nos algoritmos de QQ e correlação foram os padrões das bibliotecas *scikit-learn* e *pandas*, respectivamente. A seleção dos k primeiros atributos para realizar os cortes nas técnicas de ranqueamento foram fixos em: 20, 40 e 60. Os parâmetros usados no AG foram inspirados no artigo de Santos et al. [2020] e avaliados por testes de validação de parâmetros considerando a assertividade e o tempo de processamento. Dessa forma foi utilizado o valor de 15 para tamanho da população e quantidade de gerações; $0,09$ para taxa de *crossover*; e $0,01$ para taxa de mutação.

Na abordagem embutida foi usada a mesma configuração da classificação, sendo ela composta pelos parâmetros padrões da biblioteca *scikit-learn* para o classificador AD, RF e RL. As alterações feitas foram referentes ao número de estimadores da RF, definido como 15 e os parâmetros $max\_iter = 200$, $random\_state = 42$, $solver = sag$ e $C = 0,5$ definidos para RL. A validação cruzada foi realizada de modo estratificado, com permutação dos elementos dos grupos a cada experimento e 10 grupos de *k-fold*.

Método	**Configuração dos Parâmetros**
QQ	Configuração padrão Scikit-learn, k = 20, 40 e 60
Correlação	Pearson, configuração padrão Pandas. k = 20, 40 e 60
AG	População = 15, Gerações = 15, $crossover = 0,09$, mutação = $0,01$
AD	Configuração padrão *scikit-learn*
RF	Estimadores = 15, Configuração padrão *scikit-learn*
RL	$max\_iter = 200$, $random\_state = 42$, $solver = sag$, $C = 0,5$, Configuração padrão Sckit-learn
Validação Cruzada	Estratificada, k-fold= 10, $shuffle = True$

Tabela 7 – Parâmetros utilizados

5.2- Métricas

As métricas de desempenho são indicadores quantificáveis usados para avaliar e comparar a qualidade dos resultados. Nesse trabalho as métricas de classificação foram avaliadas a fim de mensurar a qualidade dos modelos em classificar uma situação de evasão. As medidas foram obtidas a partir da matriz de confusão, que mede a quantidade

[2] `https://scikit-learn.org/stable`

[3] `https://deap.readthedocs.io/en/master`

[4] Os códigos usados estão disponíveis em: `https://github.com/daniellefalbuquerque/selecao_de_atributos_CES.git`

de acertos e erros do modelo para cada classe. Na Tabela 8 está ilustrada a matriz de confusão, que compara a classe prevista do item com a real classe daquela amostra.

		Classe Prevista	
		Positivo	Negativo
Classe Real	Positivo	VP	FN
	Negativo	FP	VN

Tabela 8 – Matriz de confusão.

A partir da matriz de confusão, temos os Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). O VP é referente à quantidade de elementos positivos classificados corretamente, como por exemplo, a quantidade de alunos que foram classificados como evadidos e realmente evadiram. O VN é referente à quantidade de amostras classificadas como negativas corretamente, como o aluno que não evadiu e assim foi classificado pelo modelo. Os FP e FN são referentes às classificações erradas. No primeiro caso, o aluno foi classificado como evadido mas na realidade ele não evadiu, enquanto que no segundo caso, alunos que evadiram foram classificados como não evadidos.

Os valores de VP, VN, FP e FN são usados para calcular métricas de desempenho, como $acurácia$, $precisão$, $recall$ e f_1-Score (f_1). A $acurácia$ está apresentada na Equação 9 e representa, de modo geral, a porcentagem de amostras que foram classificadas corretamente. A $precisão$ está representada na Equação 10 e significa a porcentagem de acertos dentre todos as amostras classificadas como positivas, enquanto que o $recall$, demonstrado na Equação 11, representa a porcentagem de acertos dentre todas as amostras que realmente são positivas. A f_1 combina $precisão$ e $recall$ por meio de uma média harmônica de modo a trazer uma única medida que indica a qualidade geral do modelo, como visto na Equação 12. Essa métrica é mais vantajosa do que a acurácia porque possui sensibilidade para classes desbalanceadas.

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{9}$$

$$precisão = \frac{VP}{VP + FP} \tag{10}$$

$$recall = \frac{VP}{VP + FN} \tag{11}$$

$$f_1 = 2 \times \frac{precisão \times recall}{precisão + recall} \tag{12}$$

Neste trabalho a métrica f_1 foi utilizada na função objetivo do algoritmo genético, na avaliação final das classificações e serviu de base comparativa, pois ela consegue agregar duas medidas importantes, que são a $precisão$ e $recall$ e, assim, ter uma melhor avaliação tanto da classe minoritária quanto majoritária. Essa precisão é importante para o estudo pois as classes da base de dados são desbalanceadas.

O tempo de processamento da classificação e a quantidade de atributos selecionada também foram utilizados como medidas comparativas dos resultados. Para avaliar se um resultado é melhor ou pior do que a estratégia base (sem SA) foi utilizado o método estatístico não-paramétrico Wilcoxon-Mann-Whitney Fix and Jr [1955]; Komatsu [2017]. Ele verifica por meio do teste de hipótese se para um determinado nível de confiança duas distribuições de dados são estatisticamente diferentes. O valor de p encontrado no teste deve ser no máximo $0,05$ para rejeitar a hipótese nula e concluir que os resultados da validação cruzada sem e com o uso da SA são significativamente distintos.

Para completar a avaliação experimental, uma análise qualitativa foi realizada a respeito dos atributos selecionados. Nela foram discutidas as possíveis relações entre eles, abordagens realizadas na literatura e análise da distribuição dos dados por meio da frequência dos valores possíveis.

5.3- Análise exploratória dos dados do CES

A base de dados obtida após a etapa de tratamento de dados descrita na Seção 4.2 contém $4.387.008$ linhas e 105 colunas. As porcentagens de evadidos e formados ficaram em 73% e 27% das amostras, respectivamente. Isso significa que os dados da classe estão desbalanceados. Dados desbalanceados podem gerar vieses nos resultados de classificação devido à má distribuição de amostras para treino e teste. Para resolver esse problema é usado o método de validação cruzada estratificada na validação dos conjuntos e nos métodos de SA que demandam classificação.

Na Figura 11 está a proporção de evadidos e formados dentre as instituições públicas e privadas. Cerca de 17% dos alunos estão vinculados a instituições públicas, sendo que 11% estão como evadido e 6% como formados. Dos 83% que pertencem à rede privada, 62% são referentes aos evadidos e 21% aos formados.

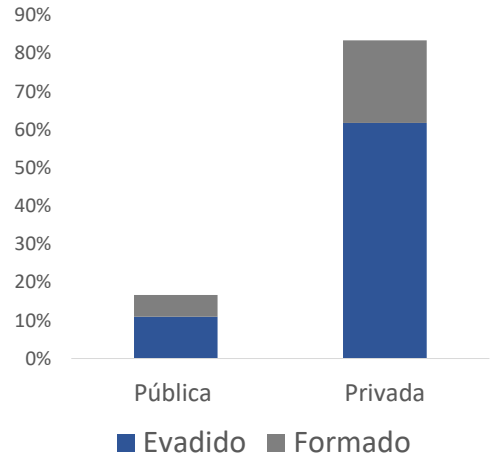

Figura 11 – Participação de Evadidos *versus* Formados nas IES públicas e privadas.

Na Figura 12 está representada a participação de estudantes que evadiram de cursos presenciais por região do curso e do EaD. Cerca de 73% dos alunos são de cursos presenciais e estão divididos na base de dados em 22% para formados e 52% para evadidos. Da mesma forma, 27% dos estudantes da base de dados são de cursos EaD, nos quais aproximadamente 6% aparecem como formados e 21% como evadidos. Observe que na distribuição de evadidos e formados pelas regiões brasileiras para o ensino presencial, as porcentagens ficaram semelhantes entre as regiões na faixa de 70% e 30%, respectivamente.

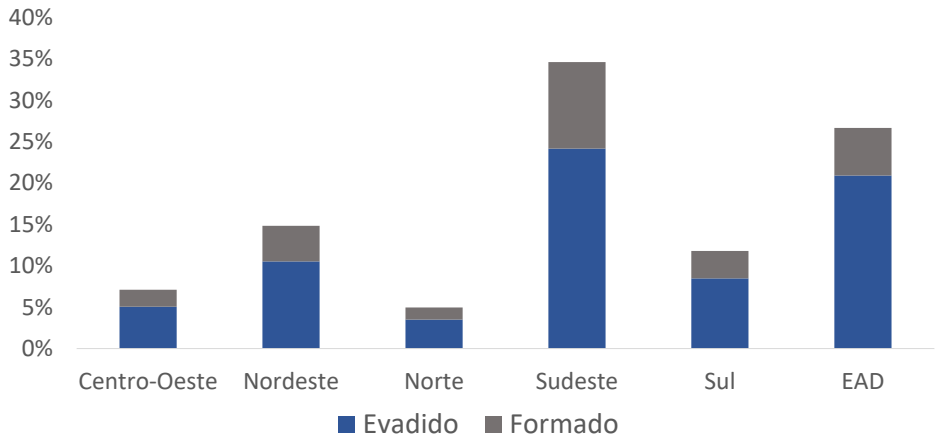

Figura 12 – Participação de Evadidos *versus* Formados nos cursos presenciais por região brasileira, e no EaD.

Ao inferir as possíveis causas da evasão, é comum pensar que os motivos que levaram um estudante a desistir do curso sejam parecidos para grupos semelhantes

(como por exemplo, entre estudantes de instituições privadas) e sejam diferentes para grupos que vivenciam a experiência acadêmica de formas distintas (como por exemplo, estudantes que cursam de forma presencial *versus* aqueles que estudam no modelo EaD). Os atributos que mais impactam cada grupo de forma particular são importantes, pois se analisados apenas no contexto geral podem mascarar problemas específicos de determinados perfis de estudantes.

Por esse motivo, as categorias administrativas (pública e privada) e as modalidades de ensino (presencial e EaD) foram avaliadas separadamente nas Seções 5.4.2 e 5.4.3, respectivamente. Essa divisão visa auxiliar na identificação da existência de indícios que sejam distintos para a evasão de cada grupo de estudantes.

5.4- Análise comparativa dos métodos

Esta seção apresenta os resultados referentes à comparação das combinações de algoritmos de classificação e técnicas de SA que melhor auxiliam na identificação dos principais atributos associados à evasão no contexto do ensino superior brasileiro. A Seção 5.4.1 analisa por meio das métricas de desempenho quais métodos de SA são mais adequados para cada classificador considerando o cenário geral, ou seja, a base inteira do CES. Além disso, a seção discute quais atributos foram considerados mais importantes na evasão, trazendo informações sobre como eles foram citados na literatura da MDE.

As seções seguintes avaliam se há diferença nos atributos selecionados e nas combinações mais assertivas de métodos de SA e classificadores para perfis específicos de estudantes. Na Seção 5.4.2 são avaliados separadamente estudantes vindos de IES públicas e privadas, enquanto a Seção 5.4.3 analisa os resultados para estudantes vindos de cursos presenciais ou do EaD.

5.4.1 Cenário geral

Para avaliar a evasão no contexto geral da base do Censo, seguindo a metodologia proposta, as técnicas de SA usadas foram: (i) abordagens de filtro com correlação e qui-quadrado, (ii) abordagem embutida com os três classificadores utilizados nesse estudo, (iii) abordagem *wrapper* com a função objetivo básica e (iv) abordagem híbrida com filtro e *wrapper*. Nas abordagens de filtro e embutida, foram usados cortes dos primeiros 20, 40 e 60 atributos a partir do ranqueamento encontrado para montar os conjuntos de atributos selecionados.

Dos resultados obtidos, são apresentados os 10 que tiveram o maior valor de f_1 para cada classificador usado. Nas Figuras 13, 14 e 15 são apresentados a média de f_1, a quantidade de atributos selecionados e o tempo de classificação (em segundos) resultantes da validação cruzada dos conjuntos selecionados para os classificadores AD, RF e RL, respectivamente. A linha vermelha indica os valores das métricas para o modelo padrão sem seleção de atributos para fins comparativos.

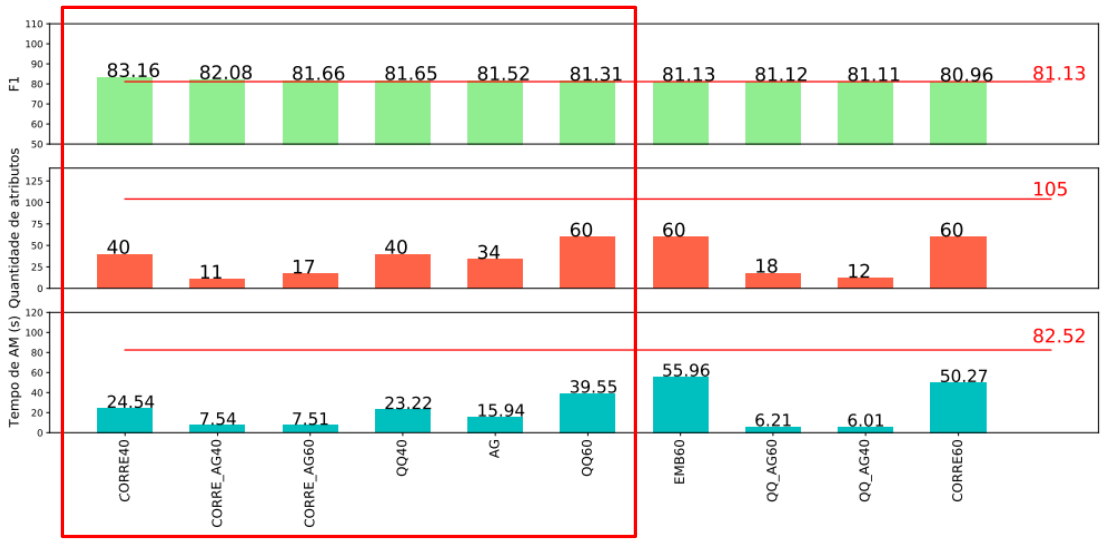

Figura 13 – Resultados da seleção de atributos com classificador Árvore de Decisão

Para a AD, os métodos em destaque obtiveram um f_1 estatisticamente maior do que o cenário sem seleção de atributos. Foram eles: CORRE40, CORRE_AG40, CORRE_AG60, QQ40, AG e QQ60. Estes métodos selecionaram, respectivamente, 40, 11, 17, 40, 34 e 60 atributos, o que possibilitou a redução do volume de dados de 10 a 60% sem perder a qualidade da classificação. Neste classificador a abordagem híbrida teve

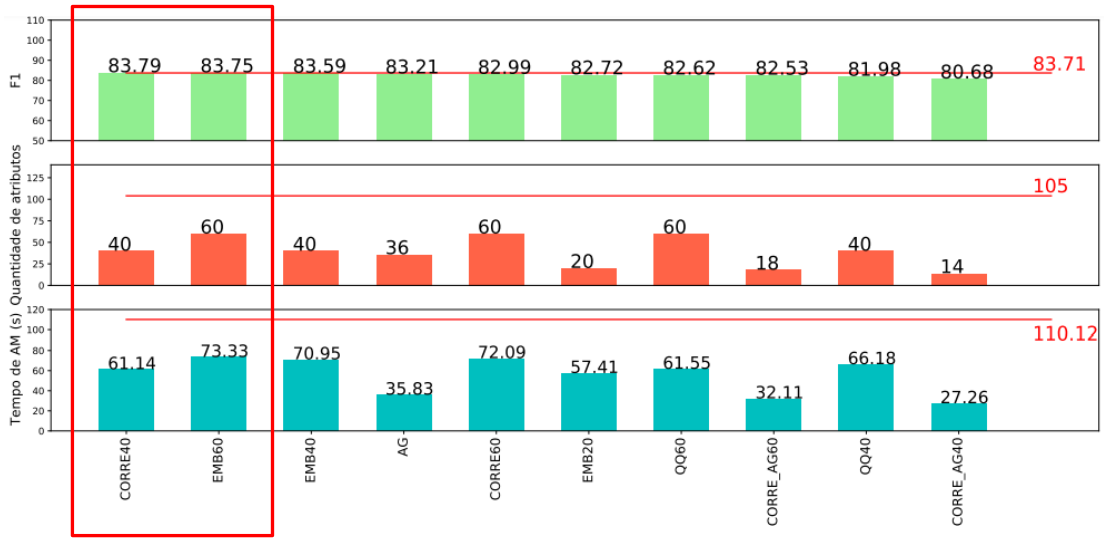

Figura 14 – Resultados da seleção de atributos com classificador *Random Forest*

destaque pois além de deixar a classificação mais assertiva, reduziu o número de atributos para apenas 10% do valor inicial e o tempo de classificação para 9% do necessário com a base completa. Esse resultado é interessante pois simplifica a análise das possíveis causas que levam a evasão e facilita a proposta de medidas de prevenção, visto que é possível determinar um pequeno conjunto de atributos que devem ser analisados.

O classificador RF é do tipo modelo múltiplo e utilizou a combinação de 15 AD's. Devido a sua complexidade, ele parte de um f_1 sem SA mais alto do que os demais classificadores. Sendo assim, apenas dois métodos conseguiram superar o f_1 do modelo sem SA, que foram: CORRE40 e EMB60. No entanto, esses resultados não apresentaram significância estatística. Com outros três métodos de seleção, a RF teve uma f_1 em torno de 83%, foram eles: EMB40, AG (com 36 atributos) e CORRE60. Uma desvantagem deste classificador é o tempo de processamento que costuma ser maior em relação aos outros classificadores. Com isso, o uso da SA pode ser ainda mais interessante quando aplicada em conjunto com a RF para reduzir o tempo computacional. A redução do tempo ficou em torno de 55% para os resultados que superaram o f_1 da base completa. Com o uso do AG, a redução do tempo foi 32% do total mantendo um f_1 de $83,2\%$.

O classificador RL foi o que apresentou menor f_1 no modelo sem SA, mas mesmo assim nenhum dos métodos de SA aplicados contribuiu para melhorá-lo. No entanto, os métodos de ranqueamento embutido e de filtro apresentaram f_1 semelhante ao da base completa e conseguiram reduzir o volume de dados necessários, assim como o tempo de classificação.

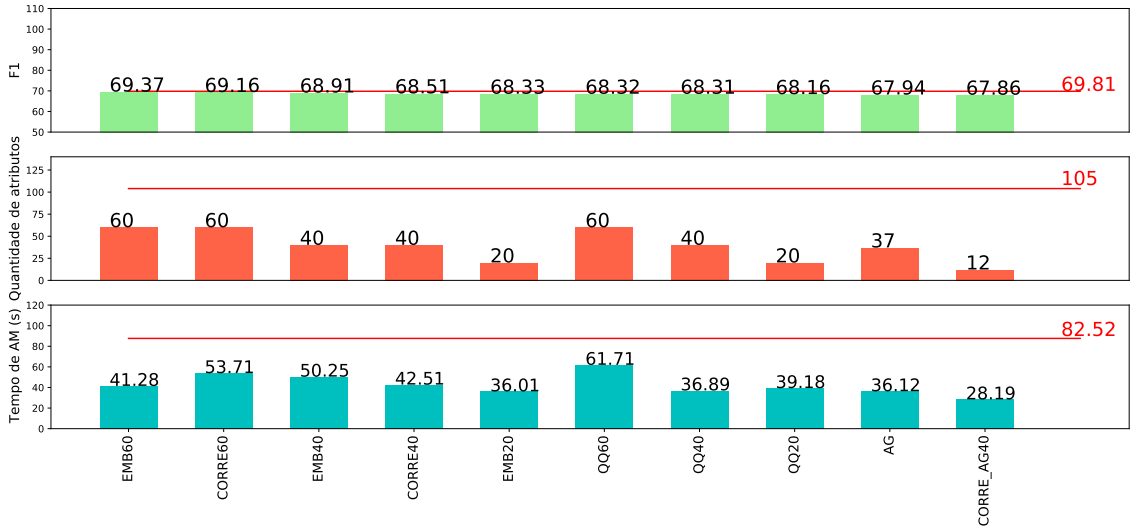

Figura 15 – Resultados da seleção de atributos com classificador Regressão Logística

A Tabela 9 apresenta uma comparação dos métodos de seleção de atributos que tiveram o valor médio de f_1 maior do que o modelo sem SA (*baseline*), por isso apenas os resultados para os classificadores AD e RF são apresentados. Essa comparação é feita por meio do p-valor encontrado no teste estatístico de Wilcoxon. Pode-se notar que os seis primeiros resultados de AD foram estatisticamente melhores do que o modelo *baseline*, enquanto que para RF nenhum dos resultados com ganho de f_1 apresentou p-valor menor do que $0,05$ e por isso não passaram no teste estatístico.

Classif.	**Método**	f_1 **sem SA (%)**	f_1 **Método (%)**	**p-valor**
AD	CORRE40	81,13 $\pm$ 0,04	83,16 $\pm$ 0,11	0,005
	AG_CORRE40		82,08 $\pm$ 0,10	0,005
	AG_CORRE60		81,66 $\pm$ 0,07	0,005
	QUI40		81,65 $\pm$ 0,09	0,005
	AG		81,52 $\pm$ 0,09	0,005
	QUI60		81,30 $\pm$ 0,09	0,012
	EMB60		81,13 $\pm$ 0,06	0,798
RF	CORRE40	83,71 $\pm$ 0,05	83,79 $\pm$ 0,09	0,074
	EMB60		83,75 $\pm$ 0,07	0,241

Tabela 9 – Média e desvio padrão de f_1 e p-valor dos métodos de SA que apresentaram maior f_1 em relação ao modelo sem SA.

A Tabela 10 mostra os tempos (em segundos) necessários para o processo de SA em cada método cujo resultado foi apresentado nos gráficos anteriores. O método mais rápido foi o QQ, seguido pela correlação. Os métodos de filtro são conhecidos na literatura por serem mais rápidos, no entanto, é possível observar que o QQ obteve mais destaque por levar apenas cerca de 3% do tempo necessário para a correlação. Apesar

disso, conforme visto anteriormente, a correlação possibilitou resultados mais assertivos.

Os métodos de ranqueamento embutido precisaram de um pouco mais de tempo do que a correlação, enquanto que o uso do AG elevou bastante o consumo de tempo, chegando em torno de $11,6$ mil segundos com o classificador RF. As abordagens híbridas tiveram sucesso em reduzir esse tempo gasto pelo AG pois levaram uma base menor para a otimização. Essa informação é interessante para entender o custo computacional dos métodos. No entanto, na prática, o processo de seleção de atributos pode ser feito poucas vezes, de tempos em tempos, enquanto que a classificação é um processo de atualização contínua, o que torna o tempo de classificação mais relevante que o tempo de SA.

SA	**Classificador**	**Tempo de seleção (s)**
QQ	-	1,73
CORRE	-	65,87
EMB	AD	78,73
EMB	RL	77,89
EMB	RF	96,75
AG	AD	7.692,09
AG	RL	7.630,38
AG	RF	11.569,41
CORRE_AG60	AD	3.904,43
CORRE_AG40	AD	2.522,13
CORRE_AG40	RL	3.904,43
CORRE_AG60	RF	8.815,69
CORRE_AG40	RF	5.157,31
QQ_AG60	AD	3.558,83
QQ_AG40	AD	2.804,46

Tabela 10 – Tempo de seleção de atributos para os métodos de SA analisados.

Para facilitar a análise das possíveis causas que levam a evasão dos estudantes, vamos considerar apenas os cinco primeiros métodos de SA que apresentaram maior média de f_1 para cada um dos três classificadores, eles serão chamados de *top* cinco. Com isso, temos um total 15 conjuntos selecionados que serão considerados para medir os k atributos mais frequentes. A Tabela 11 apresenta os atributos que tiveram frequência igual ou acima de 80% nos conjuntos selecionados por esses métodos (ou seja, apareceram em pelo menos 12 dos 15 conjuntos possíveis. Esta quantidade foi escolhida para simplificar a análise visto que muitos atributos aparecem com frequência alta nesses conjuntos, acima de 70%. Na Figura 16 encontra-se o detalhamento do conteúdo dos dados desses atributos para as classes de evadidos e formados com o propósito de facilitar a análise.

Atributos/Classificadores	Freq (%)
NU_ANO_INGRESSO	100
IN_ATIVIDADE_EXTRACURRICULAR	93
IN_FINANCIAMENTO_ESTUDANTIL	93
NU_CARGA_HORARIA	86
IN_INGRESSO_TOTAL	86
IN_FIN_NAOREEMB_PROUNI_INTEGR	80

Tabela 11 – Atributos mais frequentes nos conjuntos selecionados com a base inteira.

Os gráficos com atributos binários, como a participação em atividades extracurriculares (IN_ATIVIDADE_EXTRACURRICULAR), presença de financiamento estudantil (IN_FINANCIAMENTO_ESTUDANTIL), percentual de alunos ingressantes (IN_INGRESSO_TOTAL) e presença de financiamento Programa Universidade para Todos (PROUNI) (IN_FIN_ NAOREEMB_PROUNI_INTEGR), são representados como o percentual no qual a resposta foi positiva para os evadidos e para os formados. Por exemplo, para o atributo binário que diz se o aluno é ingressante ou não, podemos identificar que 23% dos que evadiram eram ingressantes, enquanto 2% dos que se formaram eram ingressantes (provavelmente vindos de reingresso). Os dados numéricos contínuos (por exemplo, NU_CARGA_HORARIA) são representados pela frequência nos intervalos de valores, sendo uma curva para evadidos e outra para formados. Por último, os atributos categóricos ou numéricos discretos com poucas opções (como por exemplo, NU_ANO_INGRESSO) são mostrados com a frequência de cada categoria ou valor.

O atributo que apareceu em todos os conjuntos selecionados foi o ano de ingresso no curso (observe que essa informação fornece indiretamente o tempo que o aluno está no curso), citado em alguns trabalhos Cabello et al. [2021]; Mussliner et al. [2021]; Belletati [2011] como fator relevante no contexto de evasão por diversos motivos, como: dificuldade de adaptação à vida acadêmica, falta de programas de integração de ingressantes, dúvidas relacionadas a vocação profissional, entre outros. Na Figura 16 vemos que a maior concentração de evadidos ingressaram em 2016, um ano antes do Censo, enquanto que a maioria dos formados entraram em 2013. Isso representa uma média de permanência no curso de 4 anos para aqueles que concluíram.

O engajamento dos alunos foi um fator considerado importante na probabilidade dele evadir, pois cerca de 19% dos alunos formados no ano de 2017 participavam de alguma atividade extracurricular (IN_ATIVIDADE_EXTRACURRICULAR) como: estágio não obrigatório, monitoria, extensão ou pesquisa científica, contra 6% dos que abandonaram

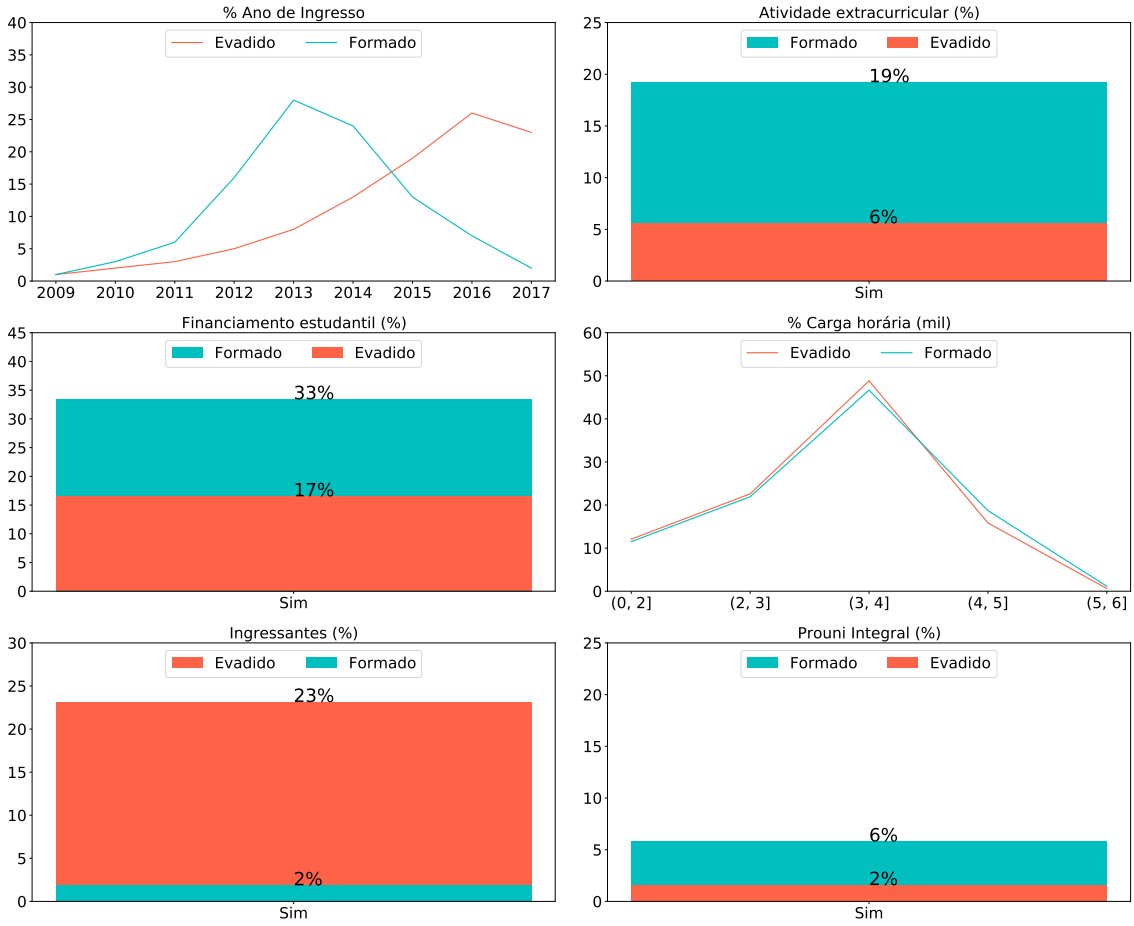

Figura 16 – Detalhamento dos atributos mais frequentes nos conjuntos selecionados.

o curso. Isso corrobora com o estudo de Teodoro and Kappel [2020] sobre o risco de evasão no ensino superior público do Brasil. Os autores encontraram alta relevância de atividades extracurriculares no risco de evasão do aluno, com o uso da SA pela correlação e pela abordagem embutida do classificador RF.

Outro fator com grande destaque é o uso do financiamento estudantil, que é um benefício utilizado por 33% dos alunos formados no ano de 2017, sendo 6% apenas no programa PROUNI com bolsa de estudos integral. O fator dessa participação em programas de financiamento ser menor para o grupo dos evadidos representa a importância de recursos financeiros na decisão de abandono de curso. O trabalho de Santos [2021] corrobora com essa ideia ao fazer uma revisão bibliográfica das principais causas da evasão universitária na literatura e encontrar a dificuldade financeira como segunda principal causa de evasão no contexto geral e a mais importante para os alunos de faculdades particulares.

Por último, a carga horária do curso foi outro atributo que obteve alta relevância,

assim como encontrado por Teodoro and Kappel [2020]. Vemos que os evadidos estão ligeiramente mais concentrados em uma carga horária intermediária de 3000 horas, enquanto que em cursos mais curtos e mais longos não há diferença significativa.

Portanto, os resultados para o cenário geral mostraram que a SA conseguiu alcançar ganhos de redução do tempo de classificação e aumento da assertividade de classificação. O classificador RF apresentou as melhores valores de f_1 e o classificador AD obteve os melhores ganhos de qualidade de classificação (f_1) em relação ao modelo sem SA. A abordagem híbrida teve destaque pois conseguiu simplificar a análise das causas de evasão, visto que reduziu a quantidade de atributos a ser analisada. A maioria dos atributos mais frequentes nos conjuntos selecionados já foram citados na literatura como relevantes no contexto da evasão de estudantes, com destaque para questões referentes ao engajamento nas atividades da IES, planejamento financeiro, tempo de duração do curso e se o estudante se encontra como ingressante ou não.

5.4.2 Instituições públicas e privadas

A análise da base de dados completa pode ocultar fatores que são importantes para um grupo de alunos com características semelhantes. Por isso, esta seção aborda a replicação da metodologia utilizada anteriormente para bases de dados parciais referentes aos alunos de IES públicas ou privadas com o objetivo de compreender se há diferença nas causas de evasão ou na assertividade da classificação. Para simplificar a replicação, apenas os cinco métodos de SA que apresentaram melhor f_1 para cada classificador (os *top* cinco) nos resultados da base completa foram aplicados.

A Figura 17 mostra as quantidades de atributos selecionados e as médias de f_1 obtidos por cada classificador com os métodos de SA aplicados para a base de estudantes de IES públicas. Observe que o classificador AD para a abordagem híbrida (CORRE_AG40) obteve um f_1 de $81,58\%$ com apenas 12 atributos selecionados. Novamente a abordagem híbrida se destaca pois essa baixa quantidade de atributos selecionados simplifica a análise do problema.

A Figura 18 mostra as métricas de desempenho para cada classificador com os métodos de SA aplicados para a base de IES privadas. Para esta base, o destaque

também aconteceu com a combinação da abordagem híbrida (CORRE_AG40) com o classificador AD, que obteve f_1 de $80,37\%$ e apenas 10 atributos selecionados. Outro resultado interessante foi a abordagem embutida com RF, que alcançou o valor de f_1 mais alto entre os encontrados até o momento, de $83,82\%$, selecionando 60 atributos.

De forma geral, os classificadores de AD e RF apresentaram maior f_1 com a base de IES privadas, enquanto o classificador RL teve seu melhor desempenho na base das IES públicas. Os maiores valores de f_1 foram encontrados na aplicação dos métodos na base de IES privadas, o que demonstra que ela possui maior facilidade na classificação de evasão do que a base de IES públicas. Algumas das possíveis causas para isso poderiam ser o fato do tamanho da amostra de dados ser maior para os alunos da rede privada ou devido ao perfil dos estudantes. Apesar das combinações de abordagem embutida com o classificador RF apresentarem os maiores f_1 para ambas as bases de dados, a abordagem híbrida com o classificador AD apresentou maior vantagem pela considerável redução da quantidade de atributos para ambas as bases.

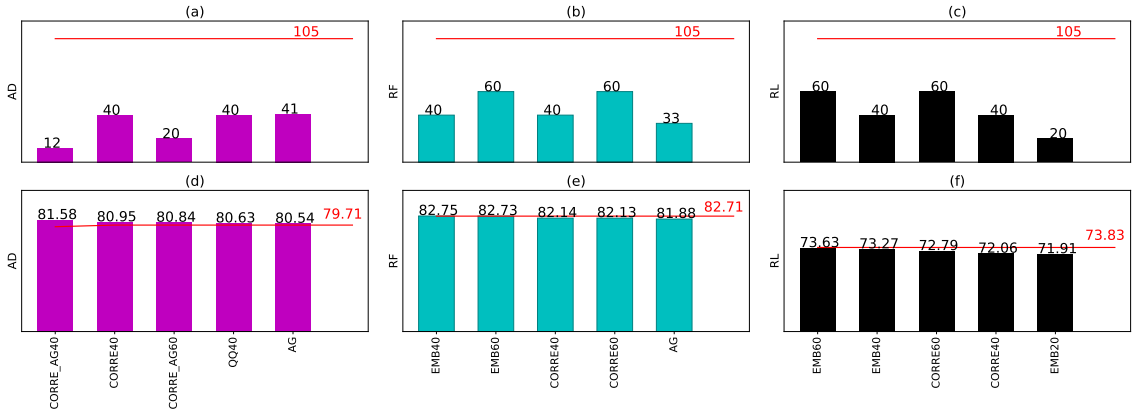

Figura 17 – Os gráficos (a), (b) e (c) são referentes às quantidades de atributos selecionados e os gráficos (d), (e) e (f) são referentes aos valores de f_1 obtidos por cada classificador usando os *top* cinco métodos de SA para a base de dados parcial referente às instituições públicas de ensino.

Na Tabela 12 estão os atributos mais frequentes nos conjuntos selecionados pelos métodos de SA com as bases públicas e privadas. É possível notar que mais atributos foram selecionados para as IES públicas do que para as IES privadas. Nas privadas foram considerados mais os aspectos financeiros e de porte da instituição, como participação no PROUNI (IN_FIN_NAOREEMB_PROUNI_INTEGR), receitas da IES (receitas) e o total de docentes (Total docente). Já nas públicas esses atributos não apareceram na lista, dando espaço para a participação em atividades extracurriculares

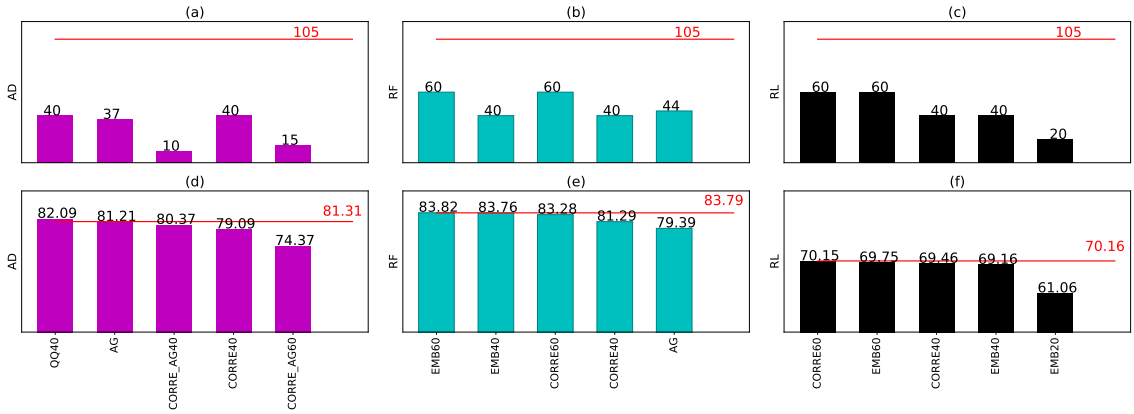

Figura 18 – Os gráficos (a), (b) e (c) são referentes às quantidades de atributos selecionados e os gráficos (d), (e) e (f) são referentes aos valores f_1 obtidos por cada classificador usando os *top* cinco métodos de SA para a base de dados parcial referente às instituições privadas de ensino.

(IN_ATIVIDADE_EXTRACURRICULAR), carga horária do curso (QT_CARGA_HORARIA_TOTAL), quantidade de vagas no curso (QT_VAGA_TOTAL), percentual de professores doutores (PER_DOC), ingresso por meio de vestibular (IN_INGRESSO_VESTIBULAR) e quantidade total de ingressantes no curso (QT_INGRESSO_TOTAL). Fatores como a quantidade de concluintes (QT_CONCLUINTE_TOTAL), percentual de ingressantes (IN_INGRESSO_TOTAL) e ano de ingresso (NU_ANO_INGRESSO) foram considerados importantes para ambos os cenários.

Públicas	Privadas
NU_ANO_INGRESSO	**NU_ANO_INGRESSO**
IN_COMPLEMENTAR_MONITORIA	Total_docente
IN_COMPLEMENTAR_ESTAGIO	**QT_CONCLUINTE_TOTAL**
IN_INGRESSO_TOTAL	IN_FIN_NAOREEMB_PROUNI_INTEGR
QT_CONCLUINTE_TOTAL	**IN_INGRESSO_TOTAL**
IN_ATIVIDADE_EXTRACURRICULAR	receitas
IN_COMPLEMENTAR_PESQUISA	
QT_CARGA_HORARIA_TOTAL	
PER_DOC	
IN_INGRESSO_VESTIBULAR	
QT_INGRESSO_TOTAL	
QT_VAGA_TOTAL	

Tabela 12 – Atributos mais frequentes nos conjuntos selecionados com a base de instituições públicas e privadas.

As Figuras 19 e 20 mostram a distribuição dos dados para esses atributos. Alguns dos fatores que apareceram como relevantes nos resultados do ensino público são a participação de atividades extracurriculares mais específicas como estágio (IN_COMPLE-

MENTAR_ESTAGIO), pesquisa (IN_COMPLEMENTAR_PESQUISA) e monitoria (IN_COMPLEMENTAR_MONITORIA). Cerca de 17% dos alunos formados participaram de atividades extracurriculares, contra uma participação quase nula dos evadidos. O perfil do aluno engajado em atividades extraclasses no ensino superior público é bastante conhecido e já foi relatado em Teodoro and Kappel [2020] ao mostrar o destaque das atividades extracurriculares no índice de conclusão do curso.

No gráfico do percentual de professores com nível de doutorado é possível perceber que há uma concentração de alunos formados onde o percentual de professores doutores é maior. Esse atributo do CES já foi analisado por Campos et al. [2019], assim como a quantidade de docentes e a participação deles em pesquisa como fatores importantes no agrupamento de perfis semelhantes de IES.

A quantidade de concluintes, de ingressantes e de vagas dos cursos também entraram nesse ranqueamento de atributos relevantes no contexto do entendimento dos riscos de evasão. A alta frequência de baixos valores pode estar associado aos cursos que foram fechados ou abertos ao longo da jornada do aluno. Estes atributos também foram considerados importantes na abordagem de Teodoro and Kappel [2020] de IES públicas. A partir do gráficos, é possível notar que a frequência dos formados é maior em cursos com muitos concluintes e que a frequência de evadidos é maior em cursos com poucos ingressantes.

O atributo IN_INGRESSO_VESTIBULAR indica se o aluno ingressou na faculdade por meio de um vestibular. 61% dos formados estavam nessa condição, contra 45% dos evadidos. Como o ano de ingresso dos formados em média foi de 4 anos e dos evadidos de 1 ano, é coerente pensar que ao longo desses três anos houve uma expansão do modelo ENEM na qual mais IES aderiram como forma de ingresso em detrimento do vestibular. Outros atributos presentes, como o ano de ingresso, a carga horária e o percentual de ingressantes, também foram encontrados na análise do cenário geral.

Apesar de não atingirem o critério de ter frequência igual ou acima de 80% nos conjuntos selecionados na base das IES públicas, os atributos IN_RESERVA_RENDA_FAMILIAR, IN_RESERVA_ETNICO e IN_APOIO_SOCIAL apareceram em $73,3\%$ dos conjuntos possíveis e se mostraram relevantes para a base de dados do ensino superior público. A reserva de vagas foi um fator destacado por Teodoro and Kappel [2020], como ligado ao maior risco de abandono de curso no ensino superior público por ter correlação negativa com a evasão. No entanto, o impacto da política de cotas em universidades públicas já foi

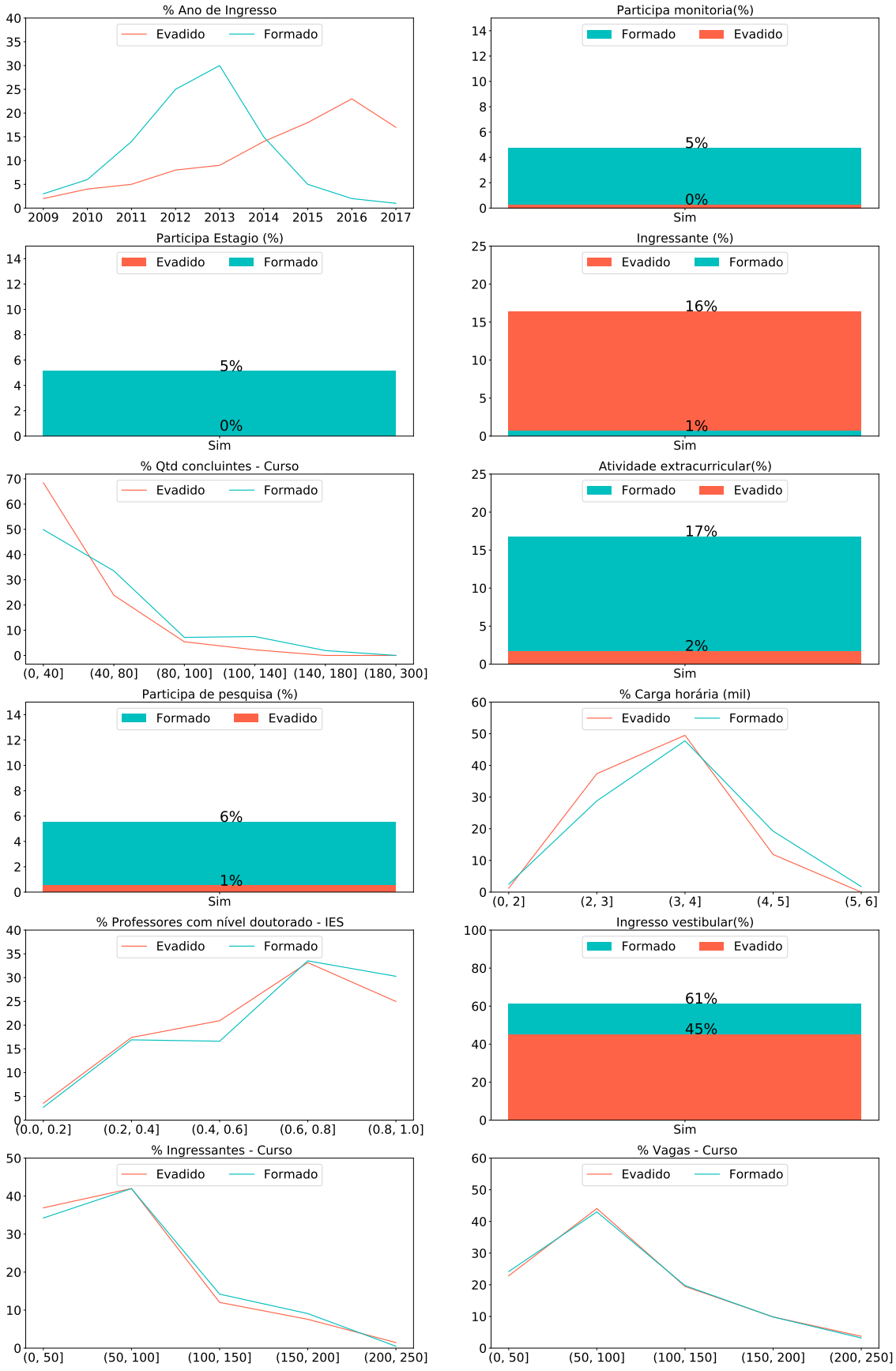

Figura 19 – Detalhamento dos atributos que mais apareceram nos conjuntos selecionados para a base de dados de instituições públicas.

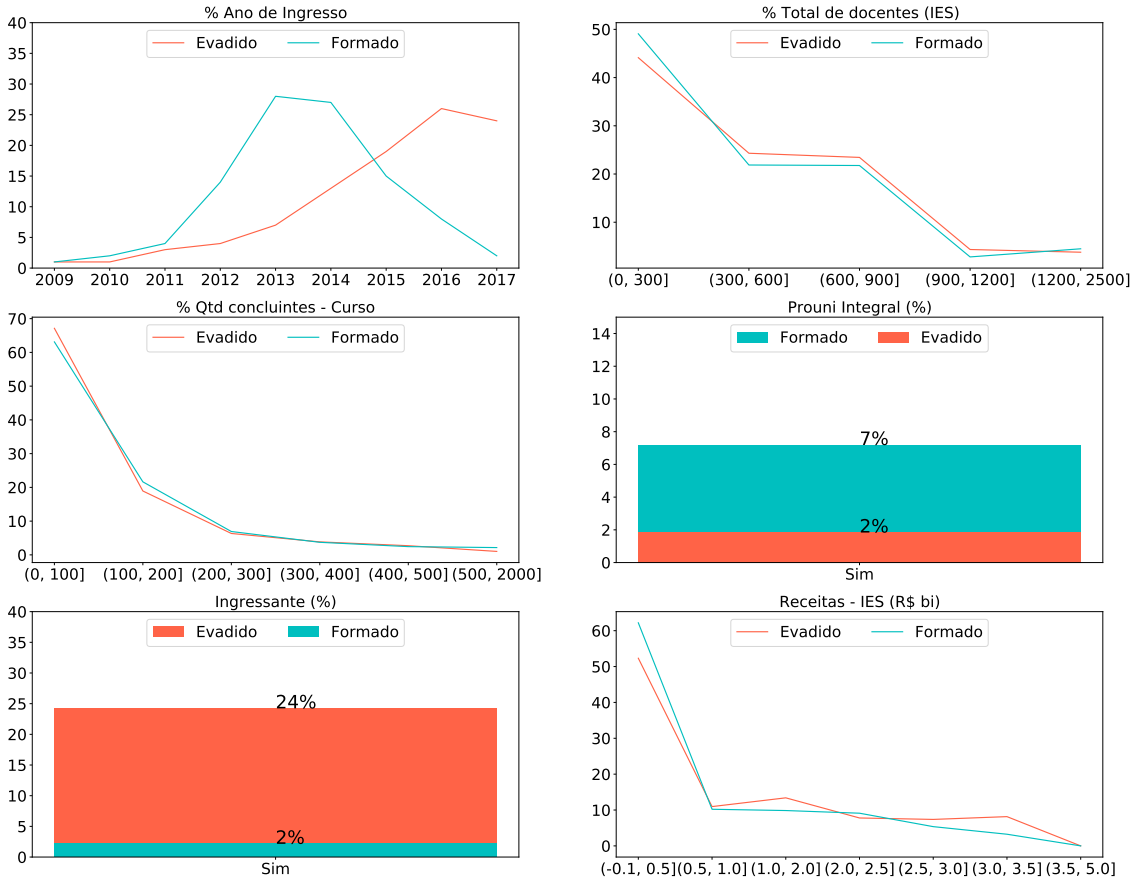

Figura 20 – Detalhamento dos atributos que mais apareceram nos conjuntos selecionados para a base de dados de instituições privadas.

discutido em trabalhos anteriores Farias [2019]; Andrade et al. [2021] como forma eficaz de redução das desigualdades sociais na educação.

Os atributos que tiveram maior destaque no ensino privado estão relacionados ao financiamento estudantil PROUNI de bolsa integral, total de docentes e receitas da IES. Enquanto o primeiro atributo demonstra a importância da questão financeira na permanência do estudante Santos [2021]; Mussliner et al. [2021], os dois últimos estão intimamente relacionados ao porte da instituição. No gráfico percebe-se que a frequência dos valores de receitas, total de docentes e quantidade de concluintes apresentam comportamento semelhante para as curvas de formados e evadidos. O ano de ingresso e o percentual de ingressantes também apareceram no gráfico, mas já foram citados anteriormente no cenário geral e nas IES públicas. Para o ano de ingresso percebe-se uma variação em relação ao gráfico das IES públicas, pois há mais formados que entraram no ano de 2014. Isso possivelmente pode estar relacionado a maior variedade de cursos de menor duração nas IES privadas.

Portanto, pode-se notar que há diferença na seleção de atributos mais importantes do cenário geral comparada a dos cenários de IES públicas e privadas, bem como no nível de assertividade de classificação e na escolha das melhores combinações de técnicas de SA e classificadores. A abordagem híbrida com a AD manteve sua posição de vantagem nesses experimentos pela redução significativa da quantidade de atributos para serem analisados. Os atributos selecionados para as IES privadas foram em menor quantidade e referentes às questões financeiras e ao porte da IES, enquanto as IES públicas apresentaram mais fatores importantes, principalmente referente ao engajamento do estudante.

5.4.3 Cursos presenciais e EaD

De forma similar, o experimento foi replicado para as bases de dados referentes aos estudantes de cursos presenciais e de EaD, afim de compreender melhor as diferenças dos riscos da evasão. A Figura 21 mostra a quantidade de atributos selecionados e as médias de f_1 das classificações, usando os métodos de SA selecionados (*top* cinco) para os cursos presenciais. O destaque ficou para a combinação da abordagem embutida com o classificador RF, que apresentou o melhor f_1, de $83,9\%$ com o uso de 60 atributos.

Na Figura 22 estão apresentados os resultados com a base de cursos EaD, tendo como destaque o uso do AG com os classificadores RF e AD. O primeiro selecionou 40 atributos e alcançou um f_1 de $83,61\%$, e segundo selecionou 37 atributos com f_1 de $83,68\%$. De forma comparativa, o nível de assertividade foi muito próximo para os cenários presencial e EaD com o classificador RF. Todavia, para a AD e a RL, a base de EaD apresentou maior assertividade, ou seja, teve maior facilidade de previsão do que a de cursos presenciais.

A Tabela 13 mostra a lista de atributos mais frequentes nos conjuntos selecionados para ambas as modalidades.

Observe que os fatores de financiamento estudantil (IN_FINAN-CIAMENTO_ESTUDANTIL), quantidade de ingresso no curso (QT_INGRESSO_TOTAL) e ano de ingresso (NU_ANO_INGRESSO) aparecem para as duas modalidades de curso. Para o presencial, o percentual de mes-

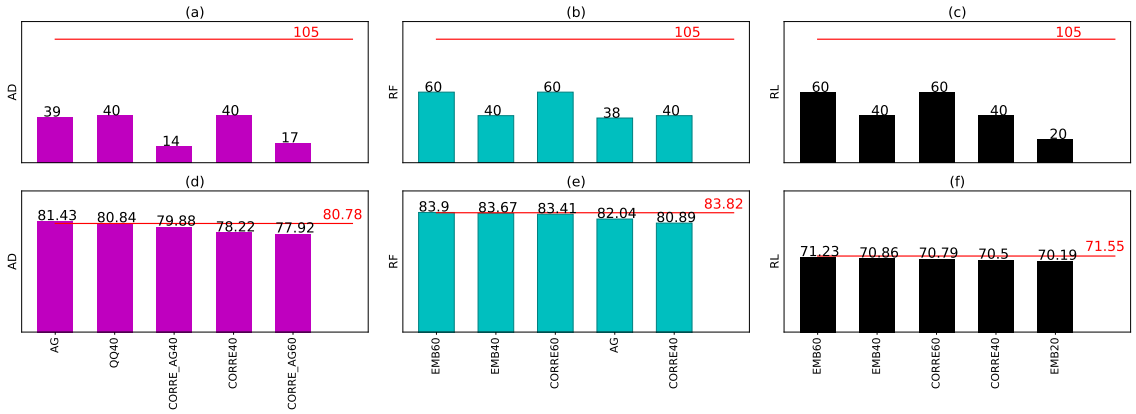

Figura 21 – Os gráficos (a), (b) e (c) são referentes às quantidades de atributos selecionados e os gráficos (d), (e) e (f) são referentes aos valores de f_1 obtidos por cada classificador usando os *top* cinco métodos de SA para a base de dados parcial referente a cursos na modalidade presencial.

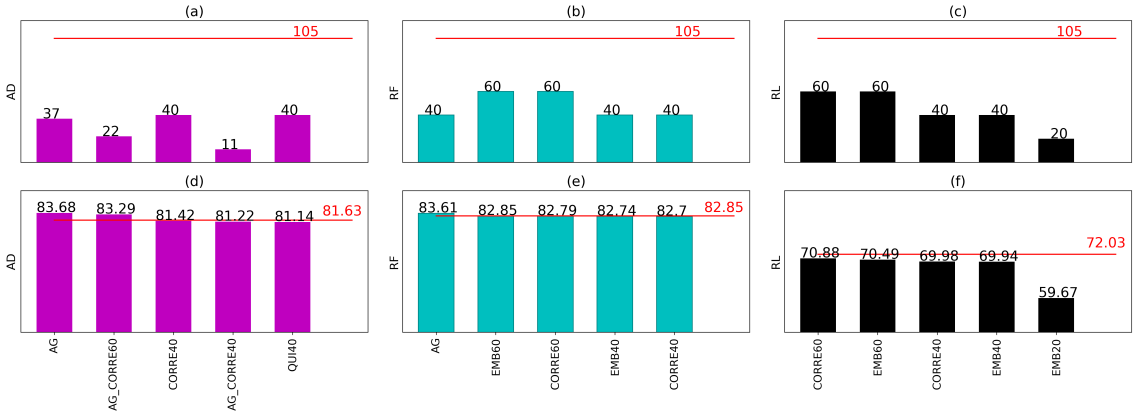

Figura 22 – Os gráficos (a), (b) e (c) são referentes às quantidades de atributos selecionados e os gráficos (d), (e) e (f) são referentes aos valores de f_1 obtidos por cada classificador usando os *top* cinco métodos de SA para a base de dados parcial referente a cursos na modalidade EaD.

tres (PER_MESTRE), a participação em extensão (IN_COMPLEMENTAR_EXTENSAO) e a quantidade de concluintes (QT_CONCLU-INTE_TOTAL) foram mais relevantes, enquanto que nos cursos EaD fatores como a carga horária (QT_CARGA_HORARIA_TOTAL), percentual de professores que trabalham em EaD (PER_TRAB_EAD), quantidade de funcionários técnicos-administrativos (QT_TEC_TOTAL), receitas e uso do benefício da bolsa de trabalho (IN_APOIO_BOLSA_TRA-BALHO) completaram a lista.

A Figura 23 mostra os gráficos dos atributos mais frequentes nos conjuntos selecionados para o ensino presencial e a Figura 24 para o ensino EaD. No ensino presencial a questão financeira foi novamente relevante, com destaque para o programa PROUNI integral [Santos, 2021]. Essa questão é impactante pois os alunos de cursos

Presencial	EaD
NU_ANO_INGRESSO	**NU_ANO_INGRESSO**
IN_INGRESSO_TOTAL	IN_APOIO_BOLSA_TRABALHO
QT_INGRESSO_TOTAL	**QT_INGRESSO_TOTAL**
IN_COMPLEMENTAR_EXTENSAO	**IN_FINANCIAMENTO_ESTUDANTIL**
PER_MESTRE	receitas
IN_FIN_NAOREEMB_PROUNI_INTEGR	PER_TRAB_EAD
QT_CONCLUINTE_TOTAL	QT_INGRESSO_VAGA_NOVA
IN_FINANCIAMENTO_ESTUDANTIL	QT_CARGA_HORARIA_TOTAL
	QT_TEC_TOTAL

Tabela 13 – Atributos mais frequentes nos conjuntos selecionados com a base de cursos nas modalidades presenciais e EaD.

presenciais possuem gastos relevantes com transporte, alimentação e moradia. Dessa forma, ter um auxilio financeiro na mensalidade do curso pode facilitar a manutenção dessas despesas básicas. A participação em extensão apareceu com 15% nos formados e 6% nos evadidos, isso mostra que o estudo presencial influencia nas oportunidades de engajamento dos alunos. Além disso, as curvas do percentual de professores com nível de mestrado, quantidade de concluintes, quantidade de ingressantes tiveram comportamento semelhante para formados e evadidos.

Na Figura 24 vemos que para os cursos EaD a porcentagem de professores que trabalham em EaD é maior no grupo dos formados. Isso leva a inferir que as IES especializadas nessa modalidade de ensino tendem a ter maior sucesso. Um atributo que surpreendeu ao aparecer na lista dos mais importantes é o auxílio bolsa trabalho, que é um tipo de atividade de extracurricular onde o aluno recebe um valor financeiro para trabalhar para a IES, mantendo um vínculo com a IES.

Ainda na base de EaD, fatores referentes ao tamanho da IES, como receita e quantidade de funcionários técnicos, tiveram destaque. Essas informações administrativas foram debatidas no estudo sobre as IES feito por Campos et al. [2019], onde elas foram agrupadas de acordo com suas características em comum. A ideia foi permitir que as IES pudessem compartilhar ou combinar recursos para conquistar uma posição mais sustentável e competitiva no mercado. Além disso, é possível notar que os cursos com menor quantidade de vagas novas tiveram maior concentração de evadidos. A distribuição dos dados dos atributos de quantidade de vagas novas e quantidade de ingressantes se mantiveram semelhantes para as curvas de formados e evadidos.

Na base EaD, o ano de ingresso que teve maior frequência para os formados foi

Figura 23 – Detalhamento dos atributos que mais apareceram nos conjuntos selecionados para a base de dados de cursos da modalidade presencial.

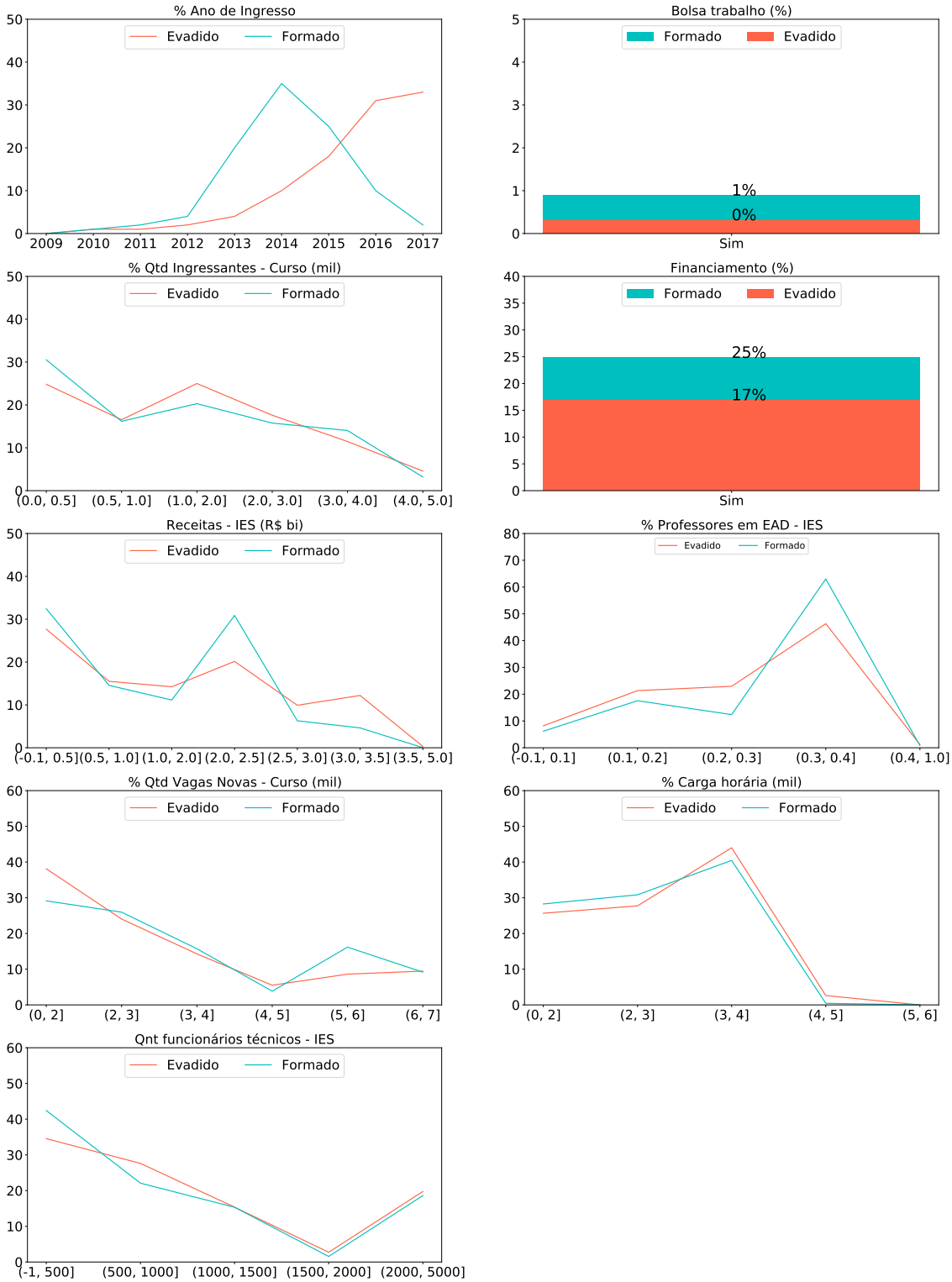

Figura 24 – Detalhamento dos atributos que mais apareceram nos conjuntos selecionados para a base de dados de cursos da modalidade EaD

2014, enquanto no cenário dos cursos presenciais, o ano de ingresso de maior frequência para os formados foi 2013. Isso mostra que os cursos no formato EaD tem menor duração. Essa informação é confirmada ao comparar o gráfico de carga horária da Figura 24 com o da Figura 16, pois existem mais cursos em EaD com carga horária baixa do que no cenário geral. Outro ponto de comparação é referente ao gráfico de financiamento estudantil para presencial (Figura 23) e EaD (Figura 24). Percebe-se que 25% dos estudantes formados do ensino EaD tem algum tipo de financiamento, contra 36% dos estudantes de cursos presenciais. Algumas causas possíveis para essa diferença poderiam ser o maior custo de se fazer um curso presencial.

Os resultados mostraram que em relação a assertividade da classificação, os cursos presenciais e EaD são semelhantes. No entanto, há diferenças na escolha dos atributos mais relevantes e na distribuição deles. Fatores como a carga horária e financiamento estudantil tiveram comportamentos distintos em relação a essas duas bases, e fatores mais específicos, como o percentual de professores que trabalham em EaD e o percentual de mestres, completaram as diferenças percebidas entre os atributos selecionados nos dois conjuntos de dados.

5.5- Experimentos com FlexAG

Esta seção apresenta os resultados da abordagem FlexAG para SA, proposta na metodologia. Para avaliar sua viabilidade, experimentos foram realizados primeiro com uma base de dados menor e pública, utilizada em diversos trabalhos da literatura Zaffar et al. [2018b]; Farissi et al. [2020]; Jalota and Agrawal [2021] e disponível no repositório eletrônico Kaggle (Seção 5.5.1). Uma vez mostrado que seu uso trouxe resultados interessantes para esta base, a proposta foi, então, aplicada à base do CES (Seção 5.5.2).

5.5.1 Prova de conceito: Base de dados Kaggle

Essa seção aborda a aplicação da abordagem FlexAG em uma base de dados disponível no repositório eletrônico Kaggle [5]. Tal base consiste em dados de uma plataforma escolar de *e-learning*, possuindo 480 amostras de estudantes e 16 atributos. Esses atributos trazem informações sobre a classificação do desempenho do aluno, que pode ser médio, alto ou baixo. Foi usado o método de transformação de variáveis categóricas em binárias [Hancock and Khoshgoftaar, 2020] para que as categorias não fossem identificadas em escala numérica. Dessa forma, a quantidade de atributos passou de 16 para 70. Por último, foi aplicada a normalização dos atributos numéricos no intervalo de 0 a 1.

Os atributos foram divididos em três grupos: comportamental (C), acadêmico (A) ou demográfico (D). Cada grupo é formado por atributos que possuem características semelhantes. Os atributos receberam um valor de prioridade referente ao grupo ao qual pertencem (de forma alternada, um grupo recebe uma prioridade maior de $0,70$, enquanto os demais, uma mesma prioridade de $0,15$) e esse valor foi contabilizado na última parcela da função objetivo, conforme detalhado no método FlexAG no Capítulo 4.

As Tabelas 14 e 15 apresentam os resultados dos experimentos com a base de desempenho escolar do Kaggle[6]. O modelo sem SA foi usado como base de comparação (*baseline)* nos experimentos. As abordagens de SA aplicadas foram: AG padrão, FlexAG com dois métodos híbridos, que utilizam a técnica de correlação ou de QQ mais o método FlexAG. Os classificadores usados foram os três já mencionados nessa dissertação (AD, RL e RF) e também o KNN, NB e RN. As tabelas fornecem, então, as seguintes informações: a média e o desvio padrão dos valores de f_1 obtidos pelos classificadores, o p-valor de f_1 obtido a partir da aplicação do método de SA em relação ao *baseline* por meio do teste estatístico de Wilcoxon e a quantidade de atributos selecionados. O objetivo desses experimentos é comparar o nível de assertividade das abordagens e entender se a priorização de grupos de atributos sugere que algum deles seja mais relevante e/ou é capaz de impulsionar o valor de f_1 da classificação.

[5] Disponível no repositório Kaggle: https://www.kaggle.com/aljarah/xAPI-Edu-Data

[6] Devido a diferença no volume de dados, a classificação com a base Kaggle foi realizada com a divisão da base de dados em 80% para treinamento e 20% para teste enquanto que na base do CES a classificação foi feita por meio da validação cruzada.

Tabela 14 – Resultados do método FlexAG para a base de desempenho de alunos do Kaggle com os classificadores KNN, Naive Bayes e Rede Neural.

SA	KNN			Naive Bayes			Rede Neural		
	f_1	p-valor	Qtd.	f_1	p-valor	Qtd.	f_1	p-valor	Qtd.
Sem SA	67,1 ± 4,3	-	70	63,8 ± 4,6	-	70	72,6 ± 3,2	-	70
AG	70,8 ± 3,5	0,060	21	64,0 ± 4,3	0,920	22	72,4 ± 3,3	0,720	21
FlexAG (C)	**77,2** ± 3,1	0,003	32	**69,0** ± 2,7	0,032	31	**75,7** ± 2,4	0,003	32
FlexAG (D)	73,2 ± 5,0	0,007	37	65,5 ± 3,0	0,850	34	74,0 ± 3,1	0,320	36
FlexAG (A)	73,9 ± 3,5	0,005	35	66,0 ± 4,5	0,280	34	73,0 ± 4,2	0,920	38
QQ_AG	70,9 ± 3,8	0,060	10	71,0 ± 4,0	0,007	9	71,1 ± 5,0	0,590	9
QQ_FlexAG (C)	**76,2** ± 3,3	0,005	18	**72,1** ± 2,4	0,003	18	**76,0** ± 1,8	0,005	17
QQ_FlexAG (D)	72,7 ± 4,0	0,020	21	67,0 ± 3,9	0,004	19	73,0 ± 3,8	0,850	22
QQ_FlexAG (A)	70,0 ± 6,8	0,420	19	70,8 ± 4,8	0,016	20	72,2 ± 3,7	0,780	20
CORRE_AG	67,0 ± 5,8	0,780	10	70,7 ± 3,3	0,007	9	73,6 ± 3,0	0,533	9
CORRE_FlexAG (C)	**75,5** ± 3,0	0,004	18	**71,8** ± 2,4	0,003	17	75,4 ± 1,6	0,016	17
CORRE_FlexAG (D)	72,8 ± 3,7	0,004	22	66,5 ± 3,5	0,247	20	**76,4** ± 3,2	0,020	21
CORRE_FlexAG (A)	72,3 ± 3,7	0,040	21	65,2 ± 5,0	1,000	21	73,3 ± 2,1	0,850	21

Tabela 15 – Resultados do método FlexAG para a base de desempenho de alunos do Kaggle com os classificadores AD, RL e RF.

SA	RL			AD			RF		
	f_1	p-valor	Qtd.	f_1	p-valor	Qtd.	f_1	p-valor	Qtd.
Sem SA	75,3 ± 2,7	-	70	72,6 ± 2,7	-	70	79,3 ± 2,9	-	70
AG	73,1 ± 2,8	0,180	21	68,7 ± 1,9	0,009	21	71,5 ± 3,6	0,003	21
FlexAG (C)	75,3 ± 5,0	0,780	32	68,0 ± 2,8	0,009	33	76,4 ± 2,2	0,032	30
FlexAG (D)	71,2 ± 2,9	0,009	36	68,9 ± 4,0	0,100	35	74,0 ± 3,1	0,150	34
FlexAG (A)	74,6 ± 3,3	0,920	37	69,9 ± 4,2	0,130	37	71,2 ± 4,8	0,016	35
QQ_AG	69,9 ± 3,2	0,003	9	67,6 ± 5,5	0,060	8	71,3 ± 4,0	0,003	9
QQ_FlexAG (C)	74,5 ± 3,5	0,650	18	71,6 ± 3,2	0,420	19	75,8 ± 2,3	0,032	18
QQ_FlexAG (D)	72,6 ± 2,5	0,007	22	68,1 ± 3,8	0,010	21	72,9 ± 4,0	0,009	22
QQ_FlexAG (A)	73,8 ± 3,3	0,530	20	65,0 ± 5,3	0,004	20	74,0 ± 2,8	0,003	21
CORRE_AG	70,7 ± 5,2	0,040	9	60,7 ± 5,4	0,003	8	72,1 ± 3,5	0,003	10
CORRE_FlexAG (C)	76,1 ± 3,3	0,850	17	67,3 ± 3,3	0,007	18	75,7 ± 2,5	0,007	18
CORRE_FlexAG (D)	75,2 ± 2,6	1,000	20	69,5 ± 3,6	0,061	21	73,1 ± 3,2	0,005	22
CORRE_FlexAG (A)	73,3 ± 2,1	0,213	22	68,4 ± 2,3	0,012	21	71,9 ± 4,7	0,003	21

Para os resultados com os classificadores KNN, NB e RN pode-se inferir que o grupo do tipo comportamental apresentou destaque na maioria dos resultados, pois quando ele foi utilizado como prioritário, acabou elevando o valor de f_1 em relação ao resultado sem SA com significância estatística. Foi possível observar também que houve ganhos de f_1 em relação ao modelo sem SA para quase todas as propostas de métodos testadas com esses classificadores.

No entanto, ao usar os classificadores AD, RL e RF, o f_1 sem o uso de SA foi maior do que quando os métodos de SA foram aplicados. Ao analisar os resultados entre os grupos com prioridade, pode-se notar que tanto para RL quanto para RF, o grupo comportamental apresentou maior f_1 do que os resultados equivalentes para os outros grupos.

Dessa forma, o método FlexAG aplicado a uma pequena base de MDE, utilizada em outros trabalhos na literatura sobre SA, apresentou resultados interessantes para os classificadores KNN, NB e RN. Os resultados que deram prioridade para o grupo de atributos do tipo comportamental tiveram resultados estatisticamente significantes de ganho de assertividade na classificação, sendo que na abordagem híbrida, com os filtros de QQ e Correlação, a redução de atributos foi mais significativa. Para os classificadores AD, RL e RF, os resultados não foram tão expressivos, visto que já partem de um f_1 maior para o modelo sem SA.

Na próxima seção essa metodologia é aplicada à base de dados do CES com os classificadores usados na análise dos métodos clássicos de SA discutida nas seções anteriores. O objetivo é verificar se haverá algum grupo de atributos que terá destaque ou se o FlexAG conseguiu impulsionar o resultado da classificação para o caso da base do CES.

5.5.2 Censo do Ensino Superior

Conforme vimos nos experimentos com os conjuntos parciais da base do CES, alguns atributos se mostraram mais frequentes na seleção de atributos que outros em cada cenário avaliado (público *versus* privado e presencial *versus* EaD). Por exemplo, os fatores financeiros tiveram destaque na seleção de atributos das IES privadas, enquanto que nas públicas, fatores de engajamento do aluno se destacaram. Isso reforça a hipótese de que a SA com priorização de grupos de atributos pode trazer bons resultados.

Nessa mesma linha, esta seção apresenta a aplicação da metodologia FlexAG para a base do CES. Ela foi usada por ter melhorado a assertividade da classificação e por ter destacado um agrupamento de atributos como mais relevante no processo de SA para a base educacional apresentada na seção anterior. Assim, o objetivo deste experimento é compreender se algum grupo de atributos pode ter uma maior influência sobre a evasão universitária no Brasil. Além disso, avaliar se esse método é capaz de melhorar a qualidade da classificação na base do CES.

Para este experimento, os atributos foram divididos em cinco grupos de acordo com o seu significado. Os grupos definidos foram: acadêmico, institucional, engajamento,

financeiro, pessoal e curso. A alocação dos atributos nos grupos pode ser vista na tabela do Anexo A. Os pesos dados foram de $0,96$ para aquele que terá a maior prioridade do experimento e $0,01$ para os demais. Esses valores foram usados para dar maior prioridade a um grupo e uma prioridade menor e igual aos quatro grupos restantes.

A Tabela 16 mostra os resultados obtidos nos experimentos com o método FlexAG, utilizando cada grupo como prioridade. Os resultados de f_1 dos métodos foram comparados ao modelo sem SA pelo teste estatístico de Wilcoxon e o p-valor apareceu abaixo de $0,05$ para todos os valores da tabela, mostrando que eles são estatisticamente diferentes do modelo *baseline*. Esse fato é diferente do acontecido com o experimento do estudo de caso, pois a base do CES tem maior quantidade de amostras fazendo com que o modelo se comporte de maneira mais estável.

SA	**AD**		**RL**		**RF**	
	Qtd	f_1 **%**	**Qtd**	f_1 **%**	**Qtd**	f_1 **%**
Sem SA	105	81,13 $\pm$ 0,04	105	69,81 $\pm$ 0,11	105	83,71 $\pm$ 0,05
CORRE_60	60	80,96 $\pm$ 0,07	60	69,16 $\pm$ 0,09	60	82,99 $\pm$ 0,08
AG	34	81,52 $\pm$ 0,09	37	67,94 $\pm$ 0,09	36	83,20 $\pm$ 0,10
FlexAG (Academico)	44	77,92 $\pm$ 0,11	50	68,20 $\pm$ 0,10	54	81,85 $\pm$ 0,06
FlexAG (Institucional)	47	83,58 $\pm$ 0,07	50	68,46 $\pm$ 0,13	46	82,36 $\pm$ 0,10
FlexAG (Engajamento)	54	82,99 $\pm$ 0,08	45	67,33 $\pm$ 0,13	49	82,55 $\pm$ 0,08
FlexAG (Financeiro)	61	73,07 $\pm$ 0,10	49	67,20 $\pm$ 0,07	48	82,93 $\pm$ 0,07
FlexAG (Pessoal)	47	80,82 $\pm$ 0,11	52	67,55 $\pm$ 0,12	47	82,32 $\pm$ 0,09
FlexAG (Curso)	42	82,49 $\pm$ 0,09	48	67,74 $\pm$ 0,08	52	82,08 $\pm$ 0,08
CORRE_FlexAG (Academico)	26	82,86 $\pm$ 0,06	30	69,00 $\pm$ 0,12	30	82,53 $\pm$ 0,10
CORRE_FlexAG (Institucional)	22	80,05 $\pm$ 0,09	33	65,86 $\pm$ 0,09	30	80,83 $\pm$ 0,10
CORRE_FlexAG (Engajamento)	26	76,59 $\pm$ 0,08	38	68,17 $\pm$ 0,13	24	82,20 $\pm$ 0,08
CORRE_FlexAG (Financeiro)	27	82,53 $\pm$ 0,06	28	67,15 $\pm$ 0,09	30	80,32 $\pm$ 0,10
CORRE_FlexAG (Pessoal)	21	78,74 $\pm$ 0,05	17	66,20 $\pm$ 0,08	20	80,22 $\pm$ 0,10
CORRE_FlexAG (Curso)	25	82,40 $\pm$ 0,70	27	60,15 $\pm$ 0,12	29	**84,26 $\pm$ 0,10**

Tabela 16 – Resultados do método FlexAG para a base do CES.

A tabela apresenta a média e o desvio padrão do f_1 e a quantidade de atributos selecionados pelos métodos de SA para cada classificador. Os métodos avaliados são o sem SA, com o uso do AG padrão, com o uso do FlexAG atribuindo prioridades para cada grupo e também o modelo híbrido que usa os 60 primeiros atributos do ranqueamento da correlação na abordagem FlexAG. A correlação foi usada pois conseguiu impulsionar mais os resultados de classificação, como visto na Seção 5.4. Para complementar a análise, também consta na tabela a classificação usando apenas a SA com correlação.

A média dos atributos selecionados variou entre 44 e 54 para o método FlexAG e de 21 a 38 usando a abordagem híbrida CORRE_FlexAg. A combinação CORRE_FlexAg(Curso) apresentou o maior f_1 usando o classificador RF, $84,25\%$, e selecionou 29 atributos com

diferença estatisticamente significativa em relação ao *baseline*.

Sendo assim, a aplicação do método FlexAG não apresentou resultados significativos de forma que seja possível afirmar que algum dos grupos de atributos avaliados da base do CES tem maior nível de importância na classificação de evasão. No entanto, essa abordagem conseguiu impulsionar a assertividade da classificação ao apresentar o melhor valor de f_1, utilizando a combinação CORRE_FlexAg(Curso) e do classificador RF. Esse conjunto de atributos pode ser visto na Tabela 18 e contém sete atributos do grupo curso (longitude, codigo_uf, Região, latitude, NU_CARGA_HORARIA, QT_VAGA_TOTAL e QT_CARGA_HORARIA_TOTAL) dentre os 29 selecionados.

5.6- Experimento com dados do CES de 2018 e 2019

Conforme descrito na metodologia, apenas a base de dados do CES do ano de 2017 foi usada para análise comparativa das técnicas de SA e de classificação nos experimentos realizados nas seções anteriores. No entanto, a escolha do ano do censo para análise pode carregar vieses a respeito da conjuntura social, cultural e econômica do país naquele ano. Observou-se, então, a oportunidade de avaliar se os conjuntos de atributos selecionados para os dados de 2017 são capazes de auxiliar a previsão do comportamento da evasão para bases de dados de anos futuros. A partir disso, esta seção apresenta os resultados para as bases de dados dos anos de 2018 e 2019 usando os conjuntos de atributos selecionados pelas combinações de SA e de classificador que apresentaram maior f_1 e menor quantidade de atributos nos dados de 2017. Este experimento tem como objetivo testar se o modelo obtido continua realizando classificações assertivas com dados novos e comprar com a o resultado da classificação sem o uso de SA (*baseline*) nos anos de 2018 e 2019.

Para isso, utilizou-se a mesma metodologia de pré-processamento de dados descrita no Capítulo 4 para a etapa de tratamento de dados. Após essa etapa, a classificação foi realizada por meio de validação cruzada com a base inteira e depois apenas com os atributos pertencentes ao conjunto analisado.

A Tabela 17 mostra a aplicação dos dois conjuntos que obtiveram resultados mais expressivos para o cenário geral nas bases de anos futuros conforme os critérios:

maior f_1 e menor quantidade de atributos selecionados. Para o primeiro critério, foi escolhido o resultado da combinação CORRE_FlexAG (Curso) e classificador RF, que apresentou uma média de $84,26\%$ para f_1. Para o segundo, foi utilizado o resultado da combinação CORRE_AG40 e classificador AD, que obteve apenas 11 atributos. Os atributos selecionados nesses conjuntos podem ser visualizados na Tabela 18.

Os atributos comuns nos dois conjuntos foram: ano de ingresso (NU_ANO_INGRESSO), percentual de professores que trabalham em pesquisa (PER_PESQUISA), UF de localização do curso (codigo_UF), carga horária do curso (QT_CARGA_HORARIA, NU_CARGA_HORARIA) e categoria administrativa (TP_CATEGORIA_ADMINISTRATIVA). A diferença de f_1 entre as duas soluções foi de aproximadamente 2 p.p, além de uma diferença na quantidade de 18 atributos.

Os resultados mostraram que tanto para a base completa do ano de 2018 quanto para a de 2019 a seleção desses conjuntos conseguiu manter a qualidade de classificação, ficando acima de 82% para o combinação CORRE_AG40 com a AD e acima de 84% para a combinação CORRE_FlexAG (Curso) com a RF. Isso acarretou em uma redução significativa do volume de dados, sendo usado apenas 10% dos dados com a AD e 28% com a RF.

Além disso, foi possível observar que os dois conjuntos de atributos considerados melhores pelos critérios estabelecidos foram capazes, juntamente com os classificadores correspondentes, de manter o nível de assertividade de classificação quando submetidos as novas bases de dados do CES, de 2018 e 2019 se comparados a não usar nenhum método de SA. Isso mostra que no contexto deste trabalho, a seleção de atributos é uma técnica importante de pré-processamento de dados, capaz de promover vantagens de desempenho de classificação e de proporcionar resultados que se mantêm com o uso de dados novos.

Classificador	SA	Qnt	f_1 (%)	Ano Censo
RF	CORRE_FlexAG (Curso)	29	84,26 $\pm$ 0,10	2017
AD	CORRE_AG40	11	82,08 $\pm$ 0,10	2017
RF	Sem SA	105	82,90 $\pm$ 0,09	2018
RF	CORRE_FlexAG (Curso)	28	84,41 $\pm$ 0,06	2018
AD	Sem SA	105	80,62 $\pm$ 0,06	2018
AD	CORRE_AG40	11	82,05 $\pm$ 0,05	2018
RF	Sem SA	105	82,89 $\pm$ 0,06	2019
RF	CORRE_FlexAG (Curso)	28	84,40 $\pm$ 0,06	2019
AD	Sem SA	105	80,61 $\pm$ 0,05	2019
AD	CORRE_AG40	11	82,01 $\pm$ 0,06	2019

Tabela 17 – Classificação das bases de 2018 e 2019 com os atributos selecionados e *baseline*.

CORRE_FlexAG (Curso)	CORRE_AG40
NU_ANO_INGRESSO	**NU_ANO_INGRESSO**
IN_FINANCIAMENTO_ESTUDANTIL	IN_ATIVIDADE_EXTRACURRICULAR
IN_COMPLEMENTAR_EXTENSAO	IN_BOLSA_ESTAGIO
IN_COMPLEMENTAR_ESTAGIO	**PER_PESQUISA**
IN_COMPLEMENTAR_PESQUISA	PER_MESTRE
PER_PEQUISA	**codigo_uf**
IN_PARTICIPA_REDE_SOCIAL	PER_SUBSTITUTO
PER_DEDIC_EXCL	**QT_CARGA_HORARIA_TOTAL**
longitude	IN_APOIO_ALIMENTACAO
codigo_uf	**TP_CATEGORIA_ADMINISTRATIVA**
IN_BOLSA_EXTENSAO	**NU_CARGA_HORARIA**
IN_ACESSO_OUTRAS_BASES	
QT_VAGA_TOTAL	
QT_CARGA_HORARIA_TOTAL	
Regiao	
IN_BUSCA_INTEGRADA	
latitude	
IN_BOLSA_MONITORIA	
NU_CARGA_HORARIA	
TP_CATEGORIA_ADMINISTRATIVA	
DT_INGRESSO_CURSO	
IN_APOIO_TRANSPORTE	
Total_docente	
IN_APOIO_BOLSA_TRABALHO	
per_investimento	
IN_FIN_NAOREEMB_MUNICIPAL	
IN_FIN_NAOREEMB_ENT_EXTERNA	
NO_OCDE_AREA_ESPECIFICA	
IN_SERVICO_INTERNET	

Tabela 18 – Conjuntos de atributos que tiveram maior f_1 e menor quantidade de atributos selecionados dos resultados apresentados.

6- Considerações finais

Este trabalho realizou uma análise comparativa de técnicas de seleção de atributos com o uso de diferentes algoritmos de aprendizado de máquina para classificar a evasão na educação de ensino superior do Brasil. Tal análise utilizou a base de dados do CES, uma importante fonte de informação sobre o ensino superior brasileiro. O objetivo foi encontrar os principais fatores que impactam na classificação da evasão acadêmica, a fim de auxiliar profissionais de educação e gestores na análise do problema, simplificando a quantidade de dados e facilitando o processo de classificação.

De forma geral, os resultados mostraram que todas as abordagens de SA tiveram algum ganho, seja em melhora no desempenho de classificação, na redução do volume de dados ou no melhor entendimento de indícios de causas da evasão. O classificador de AD conseguiu melhorar de forma significativa a classificação de evasão e reduzir consideravelmente a quantidade de atributos. Ela foi reduzida para cerca de 10% do total de atributos por meio da abordagem híbrida CORRE_AG, mantendo uma f_1 de 82%, acima da encontrada na classificação sem SA. Outras técnicas de SA que também apresentaram resultados interessantes para AD foram os filtros de correlação e QQ, e o AG na abordagem *wrapper*.

O classificador RF em conjunto com as técnicas de SA embutida e de correlação superou o valor de f_1 alcançado sem a SA. Ainda que a quantidade de atributos selecionados por essas técnicas tenha sido relativamente alta (57% da base original), o resultado pode ser considerado positivo, pois foi possível descartar 43% dos atributos da base original sem perder informação relevante e melhorando o f_1. Por outro lado, o classificador de RL não apresentou ganhos de assertividade de classificação ao ser submetido às técnicas de seleção de atributos, mesmo partindo de um valor mais baixo de f_1 inicial. Isso mostra que a vantagem da SA pode ser diferente dependendo do algoritmo de aprendizado de máquina. No entanto, quando a RL foi executada com a base parcial de alunos de instituições públicas, a assertividade do modelo melhorou cerca de 4 p.p, o que revela que também há diferença na vantagem do classificador dependendo do recorte da base de dados.

Ao analisar os atributos que mais apareceram nos conjuntos selecionados, ob-

tivemos: ano de ingresso, atividade extracurricular, financiamento estudantil PROUNI, carga horária do curso e se o aluno é ingressante. Esses fatores já haviam sido identificados e debatidos na literatura como questões importantes na análise e prevenção de evasão acadêmica. No entanto, os atributos principais se mostraram diferentes quando as técnicas foram aplicadas a recortes da base referentes a instituições públicas ou privadas, e cursos EaD ou presenciais.

Na rede pública os atributos referentes à atividade extracurricular apareceram fortemente no grupo dos formados, enquanto que na rede privada o destaque ficou com o financiamento do tipo PROUNI e fatores referentes ao porte da IES, como receitas e número total de docentes. Para os alunos dos cursos presenciais, a questão financeira obteve destaque. Já para o grupo dos que estudam em EaD, fatores referentes à quantidade de funcionários técnico administrativos e percentual de professores em EaD foram considerados relevantes.

Um método de SA, chamado FlexAG, também foi proposto a partir da abordagem clássica de *wrapper* com AG. Essa proposta incorporou uma parcela ponderada na função objetivo do AG, permitindo que informações educacionais específicas sejam priorizadas durante o processo de classificação. Essa abordagem permite que os educadores e gestores tenham maior flexibilidade para escolher se realizarão análises focando mais em aspectos financeiros, comportamentais ou geográficos, dentre outros. Assim, o modelo desenvolvido pode ser adaptado de acordo com a realidade da instituição de ensino. Por exemplo, em regiões com Índice de Desenvolvimento Econômico (IDH) mais baixo podem focar em aspectos econômicos ou de segurança, em outras regiões, podem focar em aspectos de participação em atividades extracurriculares e etc. Dessa forma, o modelo poderá ser adequado com a realidade de cada instituição ou região do país.

Os resultados com a abordagem FlexAG se apresentaram conclusivos acerca da importância dos atributos para o tipo comportamental em uma base de dados sobre desempenho escolar em EaD, usada apenas para fins de validação da proposta. Por outro lado, quando aplicada aos dados do CES, não foi possível extrair apontamentos acerca de um grupo que tivesse resultados mais significantes. No entanto, o classificador RF em conjunto com a FlexAG, priorizando o grupo de atributos referentes às informações do curso, apresentou o maior f_1 de todo o estudo, $84,28\%$ com 29 atributos selecionados, o que indica que a abordagem é promissora e requer mais avaliações em relação aos tipos de agrupamentos a serem empregados.

Por fim, os modelos que obtiveram o menor conjunto de atributos selecionados e que apresentaram maior f_1 no estudo com os dados do CES de 2017, foram também aplicados às bases de 2018 e 2019 com intuito de verificar se esses conjuntos seriam eficazes na predição da evasão em bases futuras. O resultado demonstrou que ambos mantiveram a assertividade de classificação para os dois anos testados, o que pode ser uma importante contribuição para as análises dos programas de prevenção da evasão universitária.

Esse estudo contribui para a sociedade pois mostra quais são os fatores mais impactantes na decisão de evasão de alunos universitários no Brasil de forma quantitativa usando técnicas de mineração de dados. Com essa informação pode-se criar políticas públicas que visem a mitigação do problema e a redução do desperdício que ele causa.

Parte dos resultados obtidos neste trabalho pela abordagem FlexAG foram publicados no XXXVI Simpósio Brasileiro de Banco de Dados [de Albuquerque et al., 2021]. Como trabalhos futuros, propõe-se a realização de testes com novos classificadores e técnicas de SA, ampliando o escopo de combinações possíveis e o uso da abordagem FlexAG. Em relação à base do CES, recomenda-se a comparação dos resultados obtidos com a base dos anos de 2020 e 2021 a fim de compreender se a pandemia de COVID-19 causou algum impacto no contexto educacional estudado. Pode-se também ampliar o escopo das análises dos dados co CES e dividi-los por curso, tendo uma melhor visão de alguns aspectos importantes como a desigualdade de gênero. Além disso, seria interessante aplicar o método FlexAG em outras bases de dados referentes à educação para compreender se haverá algum grupo de atributos que demonstre destaque para a MDE.

Referências Bibliográficas

Abid, A., Kallel, I., Blanco, I., and Benayed, M. (2018). Selecting relevant educational attributes for predicting students' academic performance. *Advances in Intelligent Systems and Computing*, 736:650–660. cited By 3.

Aggarwal, C. C. (2014). *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC, 1st edition.

Ahmed, M., Tahid, S., Mitu, N., Kundu, P., and Yeasmin, S. (2020). A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. *11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*.

Ahmed, S., Al-Hamdani, R., and Croock, M. (2019). Edm preprocessing and hybrid feature selection for improving classification accuracy. *Journal of Theoretical and Applied Information Technology*, 97(1):279–289. cited By 2.

Ajibade, S.-S. M., Ahmad, N. B., and Shamsuddin, S. M. (2019). An heuristic feature selection algorithm to evaluate academic performance of students. In *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*, pages 110–114.

Alam, T. M., Mushtaq, M., Shaukat, K., Hameed, I. A., Umer Sarwar, M., and Luo, S. (2021). A novel method for performance measurement of public educational institutions using machine learning models. *Applied Sciences*, 11(19).

Almasri, A., Alkhawaldeh, R., and Çelebi, E. (2020). Clustering-based emt model for predicting student performance. *Arabian Journal for Science and Engineering*, 45(12):10067–10078.

Andrade, L. V., da Silva, R. F., and Silva, R. M. F. (2021). Sistema de cotas no ensino superior: uma análise sobre ingresso e evasão. *Revista Educação e Políticas em Debate*, 10(2):955–969.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02).

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

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

Belletati, V. C. F. (2011). *Dificuldades de alunos ingressantes na universidade pública: indicadores para reflexões sobre a docência universitária.* PhD thesis, Faculdade de educação - Universidade de São Paulo.

Bermejo, P., Gámez, J. A., and Puerta, J. M. (2011). A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. *Pattern Recognition Letters*, 32(5):701–711.

Bouaguel, W. (2016). A new approach for wrapper feature selection using genetic algorithm for big data. In Lavangnananda, K., Phon-Amnuaisuk, S., Engchuan, W., and Chan, J. H., editors, *Intelligent and Evolutionary Systems*, pages 75–83, Cham. Springer International Publishing.

BRASIL (2020). *Censo da Educação Superior 2019: notas estatísticas.* Instituto nacional de estudos e pesquisas educacionais anísio teixeira (inep).

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (2017). *Classification And Regression Trees*.

Cabello, A., Imbroisi, D., Alvarez, G., Ferreira, G. V., Arruda, J., and Freitas, S. d. (2021). Formas de ingresso em perspectiva comparada: por que o sisu aumenta a evasão? o caso da unb. 26.

Campos, S. R. M. d., Henriques, R., and Yanaze, M. H. (2018). Higher education in brazil: an exploratory study based on supply and demand conditions. *Universal Access in the Information Society*.

Campos, S. R. M. d., Henriques, R., and Yanaze, M. H. (2019). Knowledge discovery through higher education census data. *Technological Forecasting and Social Change*.

Canedo, E., Tives, H., Carvalho, R., Costa, R., Santos, G., and Okimoto, M. (2019). *Educational Data Mining: A Profile Analysis of Brazilian Students*, pages 473–488. Computational Science and Its Applications – ICCSA 2019.

Carminati, G., Augusto, R., Dallabrida, N., and Teive, R. (2020). Mineração de Dados Educacionais Visando a Identificação da Evasão no Ensino Superior. pages 461–468.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28. 40th-year commemorative issue.

Chaudhury, P. and Tripathy, H. (2020). A novel academic performance estimation model using two stage feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(3):1610–1619.

Chen, M.-S., Han, J., and Yu, P. (1997). Data mining: An overview from a database perspective. *Knowledge and Data Engineering, IEEE Transactions on*, 8:866 – 883.

Colpo, M., Primo, T., Pernas, A., and Cechinel, C. (2020). Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1102–1111, Porto Alegre, RS, Brasil. SBC.

Da Silva, D. R. (2019). *Modelo para predição de risco de evasão na educação a distância utilizando técnicas de mineração de dados.* PhD thesis, UNIVERSIDADE FEDERAL FLUMINENSE.

da Silva, P. M., Lima, M. N. C. A., Soares, W. L., Silva, I. R. R., de A. Fagundes, R. A., and de Souza, F. F. (2019). Ensemble regression models applied to dropout in higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 120–125.

Das, D., Shakir, A., Rabbani, M., Rahman, M., Shaharum, S., Khatun, S., Fadilah, N., Qaiduzzaman, K., Islam, M., and Arman, M. (2020). A comparative analysis of four classification algorithms for university students performance detection. *Lecture Notes in Electrical Engineering*, 632:415–424.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1):131 – 156.

Davies, S. and Russell, S. J. (1994). Np-completeness of searches for smallest possible feature sets.

de Albuquerque, D., Brandão, D., and Coutinho, R. (2021). Um algoritmo genético com função de aptidão flexível para seleção de atributos em dados educacionais. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 355–360, Porto Alegre, RS, Brasil. SBC.

de CAMPOS, S. R. M., HENRIQUES, R., and YANAZE, M. H. (2016). Governance of Higher Education Institutions in Brazil: an Exploratory Study Based on Supply and Demand Conditions. *Advances in Intelligent Systems and Computing*, 445:V–VI.

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

Dimic, G., Rancic, D., Macek, N., Spalevic, P., and Drasute, V. (2019). Improving the prediction accuracy in blended learning environment using synthetic minority oversampling technique. *Information Discovery and Delivery*, 47(2):76–83.

Do Couto, D. D. C. and De Santana, A. L. (2017). Mineração de dados educacionais aplicada à identificação de variáveis associadas à evasão e retenção. *CEUR Workshop Proceedings*, 1877:333–344.

dos Santos Baggi, C. A. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 16:355 – 374.

Enaro, A. and Chakraborty, S. (2020). Feature selection algorithms for predicting students academic performance using data mining techniques. *International Journal of Scientific and Technology Research*, 9(4):3622–3626. cited By 0.

Evsukoff, A. G. (2020). *Inteligência computacional: Fundamentos e aplicações*. Editora E-papers, 1st edition.

Farias, S. R. d. (2019). As cotas raciais como política de ação afirmativa para a equidade de acesso ao ensino superior. *Research, Society and Development*, 8(12):e388121762.

Farissi, A., Dahlan, H. M., and Samsuryadi (2020). Genetic Algorithm Based Feature Selection for Predicting Student's Academic Performance. *Emerging Trends in Intelligent Computing and Informatics*, pages 110–117.

Febro, J. (2019). Utilizing feature selection in identifying predicting factors of student retention. *International Journal of Advanced Computer Science and Applications*, 10.

Fix, E. and Jr, J. L. H. (1955). Significance Probabilities of the Wilcoxon Test. *The Annals of Mathematical Statistics*, 26(2):301 – 312.

Galvão, S. D. C. d. O. (2007). *A Seleção de Atributos e o Aprendizado Supervisionado de Redes Bayesianas no Contexto da Mineração de Dados*. PhD thesis, Universidade Federal de São Carlos.

Garcia Torres, M., Becerra-Alonso, D., Gómez-Vela, F., Divina, F., López-Cobo, I., and Martínez-Álvarez, F. (2020). *Analysis of Student Achievement Scores: A Machine Learning Approach*, pages 275–284.

Gitinabard, N., Khoshnevisan, F., Lynch, C., and Wang, E. (2018). Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features.

Gopalakrishnan, A., Kased, R., Yang, H., Love, M., Graterol, C., and Shada, A. (2018). A multifaceted data mining approach to understanding what factors lead college students to persist and graduate. volume 2018-January, pages 372–381.

Govindasamy, K. and Velmurugan, T. (2019). Preprocessing and feature extraction process in predicting students performance using clustering technique. *International Journal of Recent Technology and Engineering*, 8(1):2407–2413.

Grant, M. J. and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2):91–108.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection 3. *Journal of Machine Learning Research*, pages 1157–1182.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining concepts and techniques, third edition*. Morgan Kaufmann Publishers, Waltham, Mass.

Hancock, J. and Khoshgoftaar, T. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 28(7).

Hasan, M. A., Nasser, M., Ahmad, S., and Molla, M. K. (2016). Feature selection for intrusion detection using random forest. *Journal of Information Security*, 07:129–140.

Hashemi, H. Z., Parvasideh, P., Larijani, Z. H., and Morad, F. (2018). Analyze students performance of a national exam using feature selection methods. In *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 7–11.

Hassan, H., Anuar, S., and Ahmad, N. (2019). Students' performance prediction model using meta-classifier approach. *Communications in Computer and Information Science*, 1000:221–231.

IBGE (2021). Censo da educação superior. https://ces.ibge.gov.br/base-de-dados/metadados/inep/censo-da-educacao-superior.html.

INEP (2021). Censo da educação superior. https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior.

Jalota, C. and Agrawal, R. (2021). Feature selection algorithms and student academic performance: A study. *Advances in Intelligent Systems and Computing*, 1165:317–328.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

Komatsu, A. (2017). Comparação dos poderes dos teste t de student e mann-whitney wilcoxon pelo método de monte carlo. *Revista de estatística da Universidade Federal de Ouro Preto*, VI:121–127.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Slovenia)*, 31:249–268.

Lal, T. N., Chapelle, O., Weston, J., and Elisseeff, A. (2006). *Embedded Methods*, pages 137–165. Springer Berlin Heidelberg, Berlin, Heidelberg.

Lei, S. (2012). A feature selection method based on information gain and genetic algorithm. *Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012*, 2:355–358.

Lilian, Z., Zheng, J., Wang, F., Li, X., Ai, B., and Qian, J. (2008). A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. *Proceedings of SPIE - The International Society for Optical Engineering*, 37.

Liu, H. and Motoda, H. (2008). *Computational methods of feature selection*. Chapman and Hall/CRC.

Liu, H. and Setiono, R. (1995). Chi2: feature selection and discretization of numeric attributes. *Proceedings of the International Conference on Tools with Artificial Intelligence*, pages 388–391.

Magalhães Hoed, R., Ladeira, M., and Leite, L. (2018). Influence of algorithmic abstraction and mathematical knowledge on rates of dropout from computing degree courses. *Journal of the Brazilian Computer Society*, 24.

Manhaes, L., Cruz, S., and Silva, G. (2015). Towards automatic prediction of student performance in stem undergraduate degree programs. *Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 247–253.

Memon, M. Q., Qu, S., Lu, Y., Memon, A., and Memon, A. R. (2021). An ensemble classification approach using improvised attribute selection. In *2021 22nd International Arab Conference on Information Technology (ACIT)*, pages 1–5.

Muchuchuti, S., Narasimhan, L., and Sidume, F. (2020). Classification model for student performance amelioration. *Lecture Notes in Networks and Systems*, 69:742–755.

Mussliner, B. O., Mussliner, M. d. S. e. S., Meza, E. B. M., and Rodríguez, G. L. (2021). O problema da evasão universitária no sistema público de ensino superior: uma proposta de ação com base na atuação de uma equipe multidisciplinar. *Brazilian Journal of Development*, 7.

Niu, Z., Li, W., Yan, X., and Wu, N. (2018). Exploring causes for the dropout on massive open online courses. In *Proceedings of ACM Turing Celebration Conference - China*, TURC '18, page 47–52, New York, NY, USA. Association for Computing Machinery.

OECD (2019). *Education at a Glance 2019*. OECD.

OECD (2020). *Education at a Glance 2020*. OECD.

Prabha, D., Siva Subramanian, R., Balakrishnan, S., and Karpagam, M. (2019). Performance evaluation of naive bayes classifier with and without filter based feature selection. *International Journal of Innovative Technology and Exploring Engineering*, 8(10):2154–2158.

Punlumjeak, W. and Rachburee, N. (2015). A comparative study of feature selection techniques for classify student performance. *Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, ICITEE 2015*, pages 425–429.

Rachburee, N. and Punlumjeak, W. (2015). A comparison of feature selection approach between greedy, ig-ratio, chi-square, and mrmr in educational mining. In *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 420–424.

Raileanu, L. and Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41:77–93.

Ramaswami, M. and Bhaskaran, R. (2009). A Study on Feature Selection Techniques in Educational Data Mining. *Journal of computing*, 1(1):7–11.

Romero, C. and Ventura, S. (2010). 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.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3.

Ruiz, R., Riquelme, J., Aguilar-Ruiz, J., and Garcia Torres, M. (2012). Fast feature selection aimed at high–dimensional data via hybrid–sequential–ranked searches. *Expert Systems with Applications*, 39:11094–11102.

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

Saeed, A., Habib, R., Zaffar, M., Quraishi, K. S., Altaf, O., Irfan, M., Głowacz, A., Tadeusiewicz, R., Huneif, M. A., Abdulwahab, A., Alduraibi, S. K., Alshehri, F. M., Alduraibi, A. K., and Almushayti, Z. (2021). Analyzing the features affecting the performance of teachers during covid-19: A multilevel feature selection. *Electronics*, 10(14).

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 313–325, Berlin, Heidelberg. Springer Berlin Heidelberg.

Santos, G. A., Belloze, K. T., Tarrataca, L., Haddad, D. B., Bordignon, A. L., and Brandao, D. N. (2020). EvolveDTree: Analyzing Student Dropout in Universities. *International Conference on Systems, Signals, and Image Processing*, 2020-July:173–178.

Santos, G. A. S., Bordignon, A. L., Oliveira, S. L. G., Haddad, D. B., Brandão, D. N., and Belloze, K. T. (2018). 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. SBC.

Santos, W. C. (2021). Evasão no ensino superior privado. *Research, Society and Development*, 10(13):e63101321034.

Silva Filho, R., Motejunas, P., Hipolito, O., and Lobo, M. (2007). A evasão no ensino superior brasileiro. *Cadernos De Pesquisa*, 37.

Sokkhey, P. and Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods. *International Journal of Advanced Computer Science and Applications*, 11(8):492–502.

Teodoro, L. d. A. and Kappel, M. A. (2020). Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, 28(0):838–863.

Thaher, T., Zaguia, A., Al Azwari, S., Mafarja, M., Chantar, H., Abuhamdah, A., Turabieh, H., Mirjalili, S., and Sheta, A. (2021). An enhanced evolutionary student performance prediction model using whale optimization algorithm boosted with sine-cosine mechanism. *Applied Sciences*, 11:1–35.

Triayudi, A. and Fitri, I. (2021). Comparison of the feature selection algorithm in educational data mining. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 19:1865.

Ullah, A., Khan, F. H., Qamar, U., and Bashir, S. (2017). Dimensionality reduction approaches and evolving challenges in high dimensional data. *ACM International Conference Proceeding Series*, pages 1–8.

UOL (2020). Na pandemia, 608 mil alunos interrompem curso no ensino superior privado. https://educacao.uol.com.br/noticias/2020/10/19/na-pandemia-inadimplencia-e-evasao-crescem-no-ensino-superior-privado.htm?cmpid=copiaecola. Outubro, 19.2020.

Urbina-Nájera, A., Camino-Hampshire, J., and Cruz Barbosa, R. (2020). University dropout: Prevention patterns through the application of educational data mining. *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa*, 26(1):1–19.

Velliangiri, S., Alagumuthukrishnan, S., and Thankumar Joseph, S. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, 165:104–111.

Venkatesh, B. and Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19(1):3–26.

Wafi, M., Faruq, U., and Supianto, A. (2019). Automatic feature selection for modified k-nearest neighbor to predict student's academic performance. pages 44–48.

Wang, D., Zhang, Z., Bai, R., and Mao, Y. (2018). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. *Journal of Computational and Applied Mathematics*, 329:307–321. The International Conference on Information and Computational Science, 2–6 August 2016, Dalian, China.

Zaffar, M., Hashmani, M., Habib, R., Quraishi, K., Irfan, M., Alqhtani, S. M., and Hamdi, M. (2021). A hybrid feature selection framework for predicting students performance. *Computers, Materials and Continua*, 70:1893–1920.

Zaffar, M., Hashmani, M., Savita, K. S., and Khan, S. (2018a). A review on feature selection methods for improving the performance of classification in educational data mining. *International Journal of Information Technology and Management*, 1.

Zaffar, M., Hashmani, M. A., and Savita, K. S. (2018b). Performance analysis of feature selection algorithm for educational data mining. *2017 IEEE Conference on Big Data and Analytics, ICBDA 2017*, 2018-January:7–12.

A- Tabela de atributos

Atributos	Descrição[1]	Grupo
despesas	Somatório das despesas de pessoal, encargos, custeio, investimento, pesquisa e outro.	Institucional
DT_INGRESSO_CURSO	Semestre de entrada do aluno.	Acadêmico
Evasao	Formado ou evadido (Matricula trancada ou desvinculado).	Classe
IN_ACESSO_OUTRAS_BASES	Bibliotecas com acesso as bases licenciadas.	Institucional
IN_ACESSO_PORTAL_CAPES	Se tem acesso ao portal da CAPES.	Institucional
IN_AJUDA_DEFICIENTE	Se tem apoio ao deficiente físico.	Institucional
IN_ALUNO_PARFOR	Se participa de programa especial para professores.	Acadêmico
IN_APOIO_ALIMENTACAO	Se recebe apoio de alimentação.	Financeiro
IN_APOIO_BOLSA_PERMANENCIA	Se recebe alguma bolsa de vulnerabilidade econômica.	Financeiro
IN_APOIO_BOLSA_TRABALHO	Se recebe remuneração por trabalho prestado a faculdade.	Financeiro
IN_APOIO_MATERIAL_DIDATICO	Recebe apoio na aquisição de material didático.	Financeiro
IN_APOIO_MORADIA	Recebe apoio na moradia.	Financeiro
IN_APOIO_SOCIAL	Recebe algum tipo de apoio social.	Financeiro
IN_APOIO_TRANSPORTE	Recebe apoio de transporte.	Financeiro
IN_ATIVIDADE_EXTRACURRICULAR	Aluno bolsista de atividade extracurricular.	Engajamento
IN_BOLSA_ESTAGIO	Aluno bolsista de estágio.	Engajamento
IN_BOLSA_EXTENSAO	Aluno bolsista de extensão.	Engajamento
IN_BOLSA_MONITORIA	Aluno bolsista de monitoria.	Engajamento
IN_BOLSA_PESQUISA	Aluno bolsista de pesquisa.	Engajamento
IN_BUSCA_INTEGRADA	Informa se as bibliotecas da IES oferecem serviços pela internet.	Institucional
IN_CAPITAL	Se o curso está localizado em uma capital.	Curso
IN_CATALOGO_ONLINE	Permite ao usuário consultar a existência e disponibilidade de itens do acervo da(s) biblioteca(s).	Institucional
IN_COMPLEMENTAR_ESTAGIO	Participa de atividade extracurricular de estágio.	Engajamento
IN_COMPLEMENTAR_EXTENSAO	Participa de atividade extracurricular de extensão.	Engajamento
IN_COMPLEMENTAR_MONITORIA	Participa de atividade extracurricular de monitoria.	Engajamento
IN_COMPLEMENTAR_PESQUISA	Participa de atividade extracurricular de pesquisa.	Engajamento
IN_FIN_NAOREEMB_ENT_EXTERNA	Financiamento estudantil reembolsável administrado por entidades externas à IES.	Financeiro
IN_FIN_NAOREEMB_ESTADUAL	Financiamento estudantil estadual não reembolsável.	Financeiro
IN_FIN_NAOREEMB_MUNICIPAL	Financiamento estudantil municipal não reembolsável.	Financeiro
IN_FIN_NAOREEMB_OUTRA	Financiamento estudantil reembolsável administrado por outras entidades.	Financeiro
IN_FIN_NAOREEMB_PROG_IES	Financiamento estudantil reembolsável administrado pela IES.	Financeiro
IN_FIN_NAOREEMB_PROUNI_INTEGR	Bolsista integral do PROUNI, tipo de financiamento estudantil não reembolsável.	Financeiro
IN_FIN_NAOREEMB_PROUNI_PARCIAL	Bolsista parcial do PROUNI, tipo de financiamento estudantil não reembolsável.	Financeiro
IN_FIN_REEMB_ENT_EXTERNA	Financiamento estudantil reembolsável administrado por entidades externas à IES.	Financeiro
IN_FIN_REEMB_ESTADUAL	Financiamento estudantil reembolsável do governo estadual.	Financeiro
IN_FIN_REEMB_FIES	Utiliza o Fundo de Financiamento Estudantil (Fies) como forma de financiamento estudantil reembolsável.	Financeiro
IN_FIN_REEMB_MUNICIPAL	Financiamento estudantil reembolsável do governo municipal.	Financeiro
IN_FIN_REEMB_OUTRA	Financiamento estudantil reembolsável administrado por outras entidades.	Financeiro
IN_FIN_REEMB_PROG_IES	Financiamento estudantil reembolsável administrado pela IES.	Financeiro
IN_FINANCIAMENTO_ESTUDANTIL	Se o aluno possui financiamento estudantil.	Financeiro
IN_GRATUITO	Gratuidade do curso.	Institucional
IN_INGRESSO_AVALIACAO_SERIADA	Aluno ingressante por meio de programa de avaliação seriada.	Acadêmico
IN_INGRESSO_CONVENIO_PECG	Aluno ingressante por meio de programa de convênio para alunos estrangeiros.	Acadêmico
IN_INGRESSO_DECISAO_JUDICIAL	Aluno ingressante por meio de decisão judicial.	Acadêmico
IN_INGRESSO_EGRESSO	Aluno ingressante por meio de ENEM.	Acadêmico
IN_INGRESSO_ENEM	Aluno ingressante por meio de ENEM.	Acadêmico
IN_INGRESSO_SELECAO_SIMPLIFICA	Aluno ingressante por meio de seleção simplificada.	Acadêmico
IN_INGRESSO_TOTAL	Se o aluno é ingressante no ano do Censo	Acadêmico
IN_INGRESSO_TRANSF_EXOFFICIO	Aluno ingressante por meio de transferência ex-officio	Acadêmico

[1]Significados retirados do anexo Dicionário de dados fornecido pelo INEP

IN_INGRESSO_VAGA_PROG_ESPECIAL	Aluno ingressante por meio de programas especiais.	Acadêmico
IN_INGRESSO_VAGA_REMANESC	Aluno ingressante por meio de vagas remanescentes.	Acadêmico
IN_INGRESSO_VESTIBULAR	Aluno ingressante por meio de vestibular.	Acadêmico
IN_PARTICIPA_REDE_SOCIAL	Biblioteca da IES participa de rede social	Institucional
IN_REPOSITORIO_INSTITUCIONAL	Possui base de dados que armazena a produção científica da IES.	Institucional
IN_RESERVA_DEFICIENCIA	Participa do programa de reserva de vagas para deficientes.	Acadêmico
IN_RESERVA_ENSINO_PUBLICO	Participa do programa de reserva de vagas para escolas públicas.	Acadêmico
IN_RESERVA_ETNICO	Participa do programa de reserva de vagas por critério étnico.	Acadêmico
IN_RESERVA_OUTRA	Participa do outro programa de reserva de vagas.	Acadêmico
IN_RESERVA_RENDA_FAMILIAR	Participa do programa de reserva de vagas por critério de renda.	Acadêmico
IN_RESERVA_VAGAS	Participa do programa de reserva de vagas.	Acadêmico
IN_SERVICO_INTERNET	Se a biblioteca tem acesso a internet.	Institucional
latitude	latitude de localização do município do curso.	Curso
longitude	longitude de localização do município do curso.	Curso
NO_IES	Nome da IES.	Institucional
NO_OCDE_AREA_ESPECIFICA	Área de estudo específica do curso.	Curso
NO_OCDE_AREA_GERAL	Área de estudo geral do curso.	Curso
codigo_uf	UF de localização do curso.	Curso
NU_ANO_INGRESSO	Ano de ingresso no curso.	Acadêmico
NU_CARGA_HORARIA	Carga horária do curso.	Curso
NU_IDADE	Idade do aluno.	Pessoal
PER_DEDIC_EXCL	% de professores com dedicação exclusiva na IES	Institucional
PER_DOC	% de professores com escolaridade de doutorado na IES.	Institucional
PER_EXTENSAO	% de professores que trabalham com extensão na IES	Institucional
PER_GRADU	% de professores com escolaridade de graduação na IES	Institucional
per_investimento	% de gastos em investimentos da IES.	Institucional
PER_MASC	% de professores homens na IES.	Institucional
PER_MESTRE	% de professores com escolaridade de mestrado na IES.	Institucional
PER_NEGROS_PARDOS	% de professores declarados negros e pardos na IES.	Institucional
PER_PESQUISA	% de professores que trabalham com pesquisa na IES.	Institucional
per_pesquisa	% gasto em pesquisa dos gastos da IES.	Institucional
PER_SUBSTITUTO	% de professores substitutos	Institucional
PER_TRAB_EAD	% de professores que trabalham no EAD	Institucional
QT_CARGA_HORARIA_TOTAL	Carga horária do curso.	Curso
QT_CONCLUINTE_TOTAL	Quantidade de concluintes no curso.	Curso
QT_INGRESSO_TOTAL	Quantidade de ingressantes total no curso.	Curso
QT_INGRESSO_VAGA_NOVA	Quantidade de ingressantes em vagas novas no curso.	Curso
QT_LIVRO_ELETRONICO	Quantidade de títulos de livros eletrônicos disponibilizados pela biblioteca.	Institucional
QT_MATRICULA_TOTAL	Quantidade de matrículas do curso.	Curso
QT_PERIODICO_ELETRONICO	Quantidade de títulos de periódicos eletrônicos adquiridos pelas bibliotecas.	Institucional
QT_TEC_TOTAL	Quantidade de técnicos administrativos da instituição.	Institucional
QT_VAGA_TOTAL	Quantidade de vagas do curso	Curso
receitas	Somatório das receitas próprias, transferências e outras.	Institucional
Região	Região do País onde o curso é localizado (Norte, Nordeste, Centro-Oeste, Sudeste e Sul)	Curso
Total_docente	Total de professores da IES.	Institucional
TP_ATRIBUTO_INGRESSO	Atributo de ingresso do aluno como área básica de ingresso do curso.	Acadêmico
TP_CATEGORIA_ADMINISTRATIVA	Se é pública, ou privada ou especial.	Institucional
TP_COR_RACA	Raça declarada do aluno	Pessoal
TP_DEFICIENCIA	Se tem deficiência fisica.	Pessoal
TP_ESCOLA_CONCLUSAO_ENS_MEDIO	Escola do ensino médio, pública ou privada.	Pessoal
TP_GRAU_ACADEMICO	Grau acadêmico diplomado.	Acadêmico
TP_MODALIDADE_ENSINO	Presencial ou EAD	Acadêmico
TP_NACIONALIDADE	Nacionalidade de alunos estrangeiros	Pessoal
TP_NIVEL_ACADEMICO	Graduação ou sequencial de graduação.	Acadêmico
TP_ORGANIZACAO_ACADEMICA	Tipo de organização acadêmica.	Institucional
TP_SEXO	Masculino ou Feminino	Pessoal
TP_TURNO	Tipo de turno	Acadêmico