Combining Data Mining Techniques to Analyse Factors Associated with Allocation of Socioeconomic Resources at IFMG

Eduardo Melo[2], Elisa Tuler[1], Leonardo Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ)
etuler@ufsj.edu.br, lcrocha@ufsj.edu.br
[2] Instituto Federal de Minas Gerais (IFMG)
eduardo.melo@ifmg.edu.br

Abstract. The granting of socioeconomic assistance to students from Federal Education Institutions is one of the ways found to provide finantial support during their studies, focusing primarily on those who are more socially vulnerable. Institutions carry out selection processes to identify students with a profile of demand and appropriately distribute the grants according to the budget available for this purpose. This article applied Data Mining techniques to a set of information from students who applied to receive scholarships at IFMG - Campus Bambuí, seeking to identify the attributes associated with the distribution of benefits and analyzing the adequacy of the current indicator used by the institution to classify the level of social vulnerability of students. The proposed methodology involved combining different machine learning algorithms, such as data classification and feature selection techniques. In addition to identifying the degree of importance of each attribute in the constructed model, the differential of this article is to present well-founded suggestions for new attributes that could be able to improve the index used by the institution and, consequently, optimize the workload of those involved with the analysis of selective processes. The composition of the institution's index with five new attributes resulted in a gain of around 10% in rating performance.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: Data Mining, Automatic Classification, Feature Selection

1. INTRODUÇÃO

As Instituições Federais de Ensino, especificamente as Universidades e Institutos Federais de Ciência e Tecnologia, estão inseridas, atualmente, em um ambiente de mudanças tecnológicas no qual a ampliação dos serviços eletrônicos oferecidos vem ocorrendo na medida em que os gestores percebem as vantagens do emprego dos recursos de TIC nos processos administrativos. Um exemplo real deste contexto pode ser obtido a partir da experiência da Diretoria de Assistência Estudantil (DIRAE) do Instituto Federal de Ciência e Tecnologia de Minas Gerais (IFMG), abordada por [Melo et al. 2021]. Em 2011, a instituição implementou o Programa de Assistência Estudantil (PAE), que pode ser entendido como um conjunto de princípios e diretrizes para a orientar o desenvolvimento de ações que buscam democratizar o acesso e a permanência dos estudantes. Entretanto, até 2018, todo o processo relacionado com a concessão de auxílios socioeconômicos para alunos em situação de vulnerabilidade social era feito por meio da análise de documentações em papel entregues em cada campus. Apenas a título de ilustração, a documentação comprobatória de cada aluno era, em média, de 55 páginas. Ao considerar que cerca de 70% dos 10.000 alunos da instituição participam de tais processos, é possível notar o volume de documentos a serem analisados pelos Assistentes Sociais. A partir de 2019, o IFMG passou a contar com um sistema de informação gerencial para controlar todo o processo de concessão

Esse trabalho foi parcialmente financiado por CNPq, CAPES e Fapemig.

de auxílios socioeconômicos, desde a entrega da documentação digital pelos alunos até a publicação dos resultados e gerenciamento das liberações de auxílios para os contemplados.

Ainda nesse contexto, o IFMG dispõe de um índice (chamado "Índice de Vulnerabilidade Social-IVS) calculado automaticamente pelo sistema de gestão mencionado que busca indicar o nível de vulnerabilidade de cada estudante. O próprio sistema já realiza uma pré-seleção, com base exclusivamente no IVS, daqueles estudantes que possuem maior demanda e, portanto, estão classificados à próxima etapa do processo seletivo, que é a conferência da documentação por parte dos Assistentes Sociais. O IVS é um índice composto por cálculos que envolvem 15 variáveis cujos valores são coletados do cadastro do estudante e dos membros do seu grupo familiar. É importante ressaltar que o IVS, por si só, não indica se o estudante terá o seu pedido deferido ou indeferido. Ele é utilizado para pré-selecionar os estudantes que possuem maior demanda em função dos recursos financeiros disponíveis pelo Campus, para que sua documentação seja analisada pelos Assistentes Sociais em uma etapa posterior. A importância do IVS está, justamente, no fato de filtrar (e reduzir) a quantidade de análises a serem feitas pelos profissionais da instituição, além de contribuir para a identificação daqueles estudantes que tendem a necessitar de apoio socioeconômico ao longo de todo o seu curso.

Mesmo considerando a utilização de sistemas informatizados como apoio ao processo de análise das informações de candidatos a auxílios socioeconômicos, como ocorre no IFMG, a definição final dos beneficiários é feita individual e manualmente por Assistentes Sociais com base, na maioria das vezes, em conhecimentos subjetivos sobre os candidatos, além de informações não necessariamente integradas em função do grande volume de dados coletados em cada edital. Esta situação pode tornar o processo falho, pois depende exclusivamente dos conhecimentos prévios de seres humanos que, por sua vez, podem não estar disponíveis em todos os editais oferecidos pela instituição. Há de se considerar que o volume de recursos públicos envolvidos apenas no IFMG com auxílio financeiro a estudantes é vultoso, mais de 8 milhões de reais em 2020 conforme [BRASIL 2021], e que há a necessidade de se avaliar um grande volume de documentos de cada candidato, mantendo o processo ainda bastante complexo.

A partir da caracterização dos problemas relacionados com o tema em questão, o presente artigo busca combinar e aplicar técnicas de mineração de dados (MD) em um conjunto de informações de alunos que se candidataram ao recebimento de auxílios no IFMG - Campus Bambuí, buscando identificar quais são os atributos associados com a distribuição dos benefícios e analisando a adequação do atual indicador utilizado pela instituição para classificar o nível de vulnerabilidade social dos alunos. A metodologia proposta envolveu a combinação de algoritmos de aprendizado de máquina, tais como classificação automática e seleção de características. Além da identificação do grau de importância de cada atributo no modelo construído, o diferencial deste artigo consiste em apresentar embasadas sugestões de novos atributos que poderiam ser capazes de melhorar o índice utilizado pela instituição e, consequentemente, otimizar a rotina de trabalho dos envolvidos com as análises dos processos seletivos.

2. TRABALHOS RELACIONADOS

Nas pesquisas realizadas para este artigo foi encontrado apenas um estudo que relacionou diretamente a aplicação de técnicas de MD em informações de candidatos a auxílios socioeconômicos em instituições federais de ensino superior. [Soares 2020] construiu uma abordagem para analisar os dados socioeconômicos de estudantes de um campus do Instituto Federal do Amazonas que pleitearam auxílios. A proposta era encontrar um algoritmo de MD eficiente para identificar aqueles alunos que, por meio de suas características, estavam aptos ao recebimento do auxílio. Foram avaliados 12 classificadores, sendo que os melhores resultados foram obtidos pelo *Random Forest*, utilizado em nossas análises.

O trabalho de [Carrano et al. 2019] aplicou técnicas de MD para analisar quais eram os atributos mais importantes na identificação da evasão em uma instituição de ensino superior pública. Apesar do contexto diferente, o conjunto de técnicas utilizadas é coerente com o objetivo proposto neste artigo. Os autores criaram uma base com dados pessoais, acadêmicos e socioeconômicos dos estudantes e

trabalharam com subconjuntos de dados com períodos diferentes para o treinamento e validação. Os algoritmos de classificação foram combinados com a métrica de seleção de características *Information Gain*. O trabalho não se ateve apenas à análise da predição da evasão, mas foi além ao qualificar os atributos que mostraram maior importância para o processo de classificação.

[Pereira et al. 2015] afirmam que existe uma relação direta entre a qualidade dos dados de treinamento e o desempenho das técnicas de classificação. Neste aspecto, os atributos identificados como irrelevantes ou redundantes podem ocasionar perda de acurácia na classificação. Os autores trabalharam com uma adaptação da técnica Information Gain para seleção de atributos em um contexto de classificação, denominando-a de MLInfoGain. Foram executadas simulações em diversos conjuntos de dados com o objetivo de comparar o desempenho da técnica proposta com outras como BR+InfoGain, Copy+InfoGain e LP+InfoGain. Os autores constataram que a técnica adaptada conseguiu obter desempenho e resultados próximos aos das outras, sendo melhor em bases de dados mais volumosas.

A redução da dimensionalidade por meio da seleção de atributos foi abordada por [Viegas et al. 2012] em um contexto com dados desbalanceados. Os autores utilizaram técnicas tradicionais de seleção de atributos (Information Gain, X2, Odds Ratio e Coeficiente de Correlação) com Programação Genética para gerar um conjunto de dados reduzido. Foi identificado, ainda, ganho de 34% na performance da classificação utilizando a medida MacroF1. Outros estudos que aplicam tais técnicas podem ser encontrados em [Omuya et al. 2021], [Zebari et al. 2020] e [El-Hasnony et al. 2020].

3. METODOLOGIA

Este artigo utiliza a técnica de estudo de caso para delimitar a sua abrangência, sendo que o objeto de estudo é composto pelos estudantes do IFMG - Campus Bambuí que participaram de editais promovidos pela DIRAE para concessão de auxílios socioeconômicos a partir de 2019. Optou-se por este Campus pela atual quantidade de estudantes e por ter oferecido seis editais, o que ocasionou maior volume de dados do que outros *campi*. As etapas de trabalho e atividades do presente artigo foram estruturadas para que fosse possível o alcance dos objetivos propostos, conforme apresentado na Figura 1.

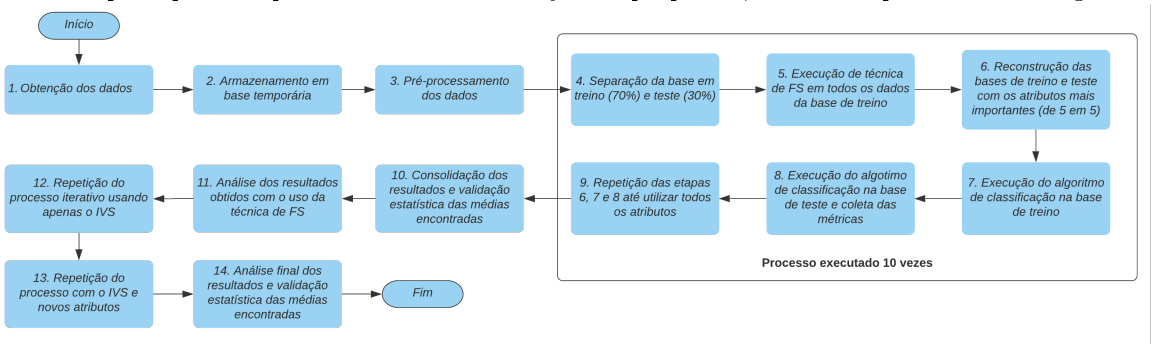

Fig. 1: Etapas propostas para o estudo

A **etapa 1** consistiu na obtenção dos dados por meio de uma consulta SQL (*Structured Query Language*) executada na base de dados de produção do sistema do IFMG. Os resultados foram exportados em um arquivo CSV (*Comma Separated Value*), o qual continha 1.367 linhas (representando o número de inscrições nos editais publicados desde 2019 no *Campus* Bambuí) e 23 colunas (representando os atributos que caracterizam o conjunto de dados). Os atributos ficaram assim nomeados: identidade_genero, estado_civil, cor, deficiencia, mae_desconhecida, nivel_curso, instituicao_ensino_fund, instituicao_ensino_med, forma_ingresso, escolaridade, situacao_trabalho, num_pessoas_grupo_familiar, num_veiculos_gf, orgf_problemas_saude, situacao_moradia, situacao_imovel, possui_imoveis_alem_moradia, meio_transporte_campus, distancia_para_campus, renda_per_capita, recebeu_auxilio_ano_anterior, idade_na_inscricao e situacao_final. O último atributo indica se a solicitação de auxílio foi deferida ou indeferida, sendo a classe alvo deste estudo.

Em seguida, para cumprir a **etapa 2**, foi criado um repositório para armazenar os dados advindos do arquivo CSV gerado na primeira etapa. As atividades relacionadas com o pré-processamento dos dados compõem a **etapa 3**, na qual foi necessário apenas preencher valores faltantes com o valor mais frequente ou com a média no caso de atributos não categóricos. Ao final desta etapa, as 1.367 instâncias continuaram disponíveis para o estudo, sendo que 643 delas (47%) se referiam a solicitações indeferidas e 724 (53%) a auxílios deferidos, indicando certo balanceamento do conjunto de dados no que se refere à classe alvo.

As etapas 4 a 9 estão vinculadas à execução da técnica de seleção de atributos e com a construção e aplicação de um modelo de classificação. Essas etapas foram executadas 10 vezes, possibilitando a coleta de resultados em bases de treino e teste compostas de diferentes instâncias a cada iteração do processo. Na **etapa 4** a base de dados era separada em dois subconjuntos denominados "treino"(com 70% das instâncias) e "teste"(com 30% das instâncias), mantendo-se a proporção entre as classes nas duas bases. Passando à **etapa 5**, os dados da base de treino eram utilizados na execução da técnica de *Feature Selection* conhecida como *ExtraTreesClassifier* para identificar a importância de cada atributo. Trata-se de um método *ensemble* que visa ajustar a quantidade de árvores de decisão aleatórias, isto é, as árvores extras como indicadas em seu nome, em amostras menores de todo o conjunto de dados para otimizar os resultados da predição e controlar o sobreajuste. Por padrão, o cálculo da importância dos atributos é baseado na importância de *Gini*, o qual geralmente isola em um ramo da árvore a classe mais frequente e considera a normalização do critério proporcionado pelo atributo. Foi criada uma lista com os atributos e sua respectiva importância, ordenada de forma decrescente. Esta lista é a entrada necessária para a **etapa 6**, na qual as bases de treino e teste são reconstruídas em termos de colunas para que fiquem apenas com os atributos mais importantes em cada execução, sempre incrementando de cinco em cinco de acordo com cada iteração. Foram criados 4 grupos: 5 atributos mais importantes, 10 atributos mais importantes, 15 atributos mais importantes e todos os atributos. A **etapa 7** consistia na execução do algoritmo *Random Forest* na base de treino para construção do modelo de classificação. Trata-se de um algoritmo capaz de construir e combinar o resultado predito por várias árvores de decisão localizadas em um mesmo contexto, isto é, uma floresta. Em seguida ao treinamento do modelo de classificação, o mesmo era aplicado na base de teste para predizer a classe alvo definida e permitir a coleta da métrica *MacroF1* que indica o resultado do processo com cada conjunto de atributos (**etapa 8**). O diferencial desta métrica consiste em encontrar a qualidade do modelo se baseando na média harmônica entre os resultados das métricas *precision* e *recall*, produzindo apenas um indicador a ser considerado na análise. Finalizando este ciclo, as tarefas das etapas 6, 7 e 8 eram repetidas até que todos os conjuntos de atributos fossem testados (**etapa 9**).

Na **etapa 10** realizamos uma validação para compreender se as diferenças entre as médias das métricas coletadas eram estatisticamente significativas, indicando eventualmente que determinada quantidade de atributos está associada com melhores resultados de classificação. Optou-se pela aplicação do Teste de Tukey, em especial pelo fato de ser uma abordagem que permite comparações múltiplas entre todos os pares indicados [Abdi and Williams 2010]. As médias foram comparadas entre 4 grupos (5, 10, e 15 atributos mais importantes e todos os atributos). Na **etapa 11** os resultados obtidos com a técnica de *Feature Selection* (coletados na etapa 5) são analisados, dando origem a uma listagem contendo a ordenação dos atributos conforme sua medida de importância. A **etapa 12** repetiu o processo da etapas 4, 7 e 8 (por 10 vezes), utilizando o mesmo conjunto de instâncias, porém apenas com um atributo representando o IVS e a classe alvo. A proposta era verificar qual a representatividade do IVS na classificação das instâncias. Dentro do escopo da **etapa 13**, visando compreender se os resultados do processo de classificação poderiam ser melhorados, foi criado um conjunto de dados com as mesmas instâncias utilizadas nas etapas anteriores, porém contendo apenas o atributo IVS e os cinco atributos mais importantes dentre aqueles que não fazem parte atualmente do cálculo do IVS. Todo o processo das etapas 4, 7 e 8 foi repetido com este conjunto de dados e os resultados comparados com aqueles encontrados nas etapas anteriores. A **etapa 14** consistiu na análise final dos resultados obtidos a partir da realização de todas as atividades mencionadas, os quais são descritos e discutidos na próxima seção.

4. ANÁLISE DOS RESULTADOS

A caracterização da amostra utilizada neste experimento contribui para melhor compreensão do conjunto de dados trabalhado. Neste sentido, alguns atributos apresentaram certo equilíbrio entre as 1.367 instâncias disponíveis. A identidade de gênero está relativamente balanceada, apresentando 57% de estudantes do sexo feminino e 43% do sexo masculino, assim como ocorreu com a forma de ingresso (51% via cotas e 49% por ampla concorrência) e se recebeu algum auxílio no ano anterior à participação no edital (55% não receberam e 45% receberam). Por outro lado, outros atributos se mostraram com pouca variação em seus valores. 97% dos estudantes indicaram ser solteiros, 98% não possuem deficiência, a mãe de 96% deles não é desconhecida, 90% cursaram o Ensino Fundamental e 95% o Ensino Médio em instituições públicas, 88% não possuem problemas de saúde, 90% utilizam transporte público e 92% não trabalham formalmente. Alguns atributos apresentaram variação intermediária entre as categorias. Neste grupo é possível destacar que 67% dos estudantes estavam matriculados em cursos de graduação e 33% em cursos técnicos; 53% vivem com familiares, 39% vivem em repúblicas ou moradia estudantil e 8% vivem sozinhos; 59% moram em imóvel próprio, 22% em imóvel alugado e 19% em outras situações; 44,5% se identificaram como brancos, 45% como pardos e 10,5% como pretos.

Primeiramente apresentamos a medida de importância de cada atributo para o conjunto de dados, coletada em todas as 10 execuções do processo descrito na seção 3. A Tabela 2 apresenta a média final para cada atributo, permitindo que os gestores da Diretoria de Assistência Estudantil do IFMG tenham uma noção de quais características estão mais associadas com a concessão dos auxílios socioeconômicos até o momento, especificamente no Campus Bambuí. A última coluna desta tabela indica se o atributo, atualmente, é utilizado na fórmula de cálculo do IVS do IFMG (mais da metade deles (doze) são utilizados pelo IVS, porém a maioria apresenta pouca importância na comparação geral).

Table I: Média final da importância dos atributos

Atributo	**Média final**	**Está no IVS**
renda_per_capita	0,1851	Sim
idade_na_inscricao	0,1308	Não
num_pessoas_grupo_familiar	0,0803	Não
distancia_para_campus	0,0762	Sim
cor	0,0582	Não
situacao_imovel	0,0562	Sim
recebeu_auxilio_ano_anterior	0,0501	Não
situacao_moradia	0,0441	Não
escolaridade	0,0428	Sim
identidade_genero	0,0383	Não
forma_ingresso	0,0345	Não
possui_veiculo	0,0315	Sim
orgf_problemas_saude	0,0259	Sim
instituicao_ensino_fund	0,0246	Sim
nivel_curso	0,0225	Não
meio_transporte_para_campus	0,0219	Não
situacao_trabalho	0,0216	Sim
possui_imovel_alem_moradia	0,0189	Sim
instituicao_ensino_med	0,0121	Sim
mae_desconhecida	0,0094	Não
estado_civil	0,0084	Sim
deficiencia	0,0065	Sim

Conforme exposto na seção 3, para cada um dos quatro grupos de atributos, realizamos as 10 execuções do processo calculando o valor médio da métrica *MacroF1*. O objetivo é identificar qual dos grupos apresenta melhor desempenho na classificação das instâncias (deferidas ou indeferidas), comparando par a par os resultados. Em cada comparação, para verificar que as diferenças era estatisticamente significativas, aplicamos o Teste de Tukey (quando o resultado da comparação é menor que 0,05, tem-se uma diferença média significativa). A Tabela 2 apresenta tanto os resultados do teste quanto aqueles da métrica MacroF1 de cada conjunto de atributos. Quando o resultado para

determinado grupo de atributos é estatisticamente superior, representamos com o símbolo ▲, estatisticamente inferior com o símbolo ▼ e para aqueles sem significância é utilizado o símbolo ●. Podemos notar que em todas as comparações os resultados obtidos utilizando os 10 atributos mais relevantes foram os que apresentaram os melhores resultados.

Table II: Validação estatística das médias MacroF1

Grupos de atributos comparados/MacroF1		Significância
5+ (0,6139)	10+ (0,6524)	0,004 ▲
	15+ (0,6204)	0,826 ●
	Todos (0,6199)	0,894 ●
10+ (0,6524)	15+ (0,6204)	0,038 ▼
	Todos (0,6199)	0,029 ▼
15+ (0,6204)	Todos (0,6199)	0,996 ●

Buscando compreender se o IVS é representativo no processo de classificação, foi criado um conjunto de dados contendo apenas dois atributos, o IVS e a classe, com as mesmas instâncias analisadas anteriormente. O processo de classificação foi repetido 10 vezes para se obter a média final da classificação. Em seguida, tendo como propósito verificar uma possível otimização na classificação das mesmas instâncias, foi criado outro conjunto de dados denominado "IVS+"com o IVS e os cinco atributos mais importantes ainda não usados no cálculo do IVS (*idade_na_inscricao, num_pessoas_grupo_familiar, cor, recebeu_auxilio_ano_anterior* e *situacao_moradia*). A Tabela 3 apresenta o resultado final dos três grupos mencionados (10 mais importantes, apenas IVS e IVS+). Podemos observar que não houve variação significativa entre o grupo 10+ e o IVS. Por outro lado, quando a comparação desses grupos é feita com o IVS+, foi encontrada variação positivamente significativa que poderia proporcionar mais de 10% na melhoria do resultado final da classificação, passando de 0,6524 para 0,7330.

Table III: Validação estatística nos novos grupos

Grupos de atributos comparados		Significância
10+ (0,6524)	IVS (0,6517)	1,000 ●
	IVS+ (0,7330)	0,000 ▲
IVS (0,6517)	IVS+ (0,7330)	0,000 ▲

Tratando sobre a análise da eventual incorporação dos cinco atributos ao cálculo do IVS, é importante abordar o contexto de cada um deles. O ideal é que este índice utilize variáveis que permitam aos interessados a participação em iguais condições para pleitear o auxílio, sendo diferenciados no momento do cálculo com base nas suas características próprias. Neste sentido, há de se considerar que os atributos *cor*, *idade_na_inscricao* e *recebeu_auxilio_ano_anterior* não seriam os mais oportunos para inclusão no cálculo do índice, uma vez que indicam características que não justificariam pontuações variadas em função de seus valores. Por exemplo, a realidade socioeconômica de um estudante com 20 anos de idade que pleitea o auxílio pode ser a mesma de outro estudante com 50 anos, não se justificando a variação nos pesos a partir da análise deste atributo. O mesmo raciocínio pode ser aplicado no caso de pontuar de forma diferente aqueles estudantes que já receberam auxílio no ano anterior ou tendo a cor como base. Sob outro prisma, os atributos *num_pessoas_grupo_familiar* e *situacao_moradia* poderiam ser incorporados ao cálculo do IVS por caracterizarem melhor a vulnerabilidade do estudante.

Outra maneira de analisar a adequação dos potenciais atributos para inclusão no cálculo do IVS é compreendendo a relação de cada um deles com a classe alvo em questão, a qual indica se a solicitação de auxílio foi deferida ou indeferida. Considerando as duas variáveis numéricas (*idade_na_inscricao* e *num_pessoas_grupo_familiar*), nota-se pelos gráficos apresentados na Figura 2 que a distribuição (ou não) dos benefícios é bastante equilibrada em relação à idade dos alunos solicitantes (A), enquanto uma variação é perceptível de acordo com a quantidade de pessoas que compõem o grupo familiar dos alunos (B). Destaca-se, ainda, o caso específico de grupos compostos por 4 ou 5 pessoas, nos quais a quantidade de solicitações indeferidas foi maior do que os deferimentos, fato que poderia ensejar uma análise diferente por parte do Assistente Social a partir do momento em que este atributo estivesse integrado ao IVS.

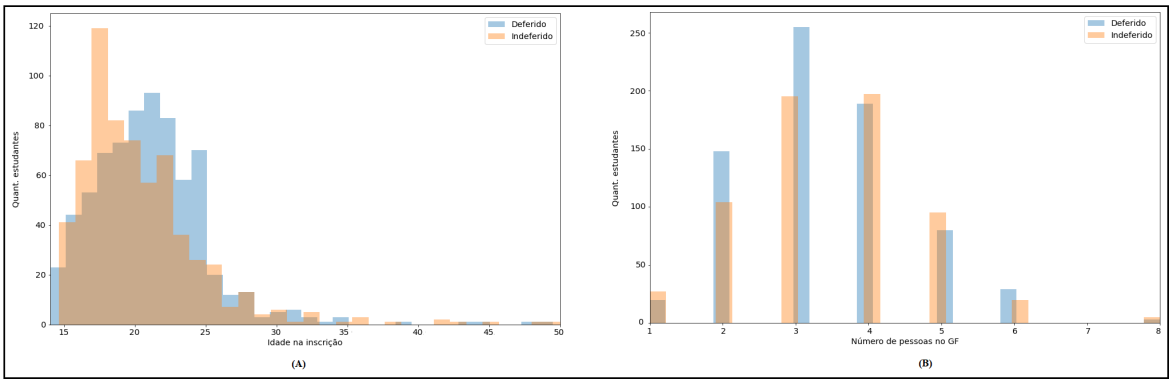

Fig. 2: Situação final por idade (A) e número de pessoas no grupo familiar (B)

De maneira análoga, é possível observar na Figura 3 que a situação final dos benefícios segue padrões diferentes quando analisados os atributos *recebeu_auxilio_ano_anterior* (A) e *situacao_moradia* (B). Enquanto praticamente o dobro das solicitações feitas por alunos que já haviam recebido o auxílio foi atendido, indicando que as condições socioeconômicas dos alunos não melhoraram no período a ponto de ainda demandarem continuidade de recebimento do benefício, a maioria das solicitações feitas por quem não havia recebido auxílio no ano anterior foram indeferidas. É interessante notar que mesmo assim um número considerável de novos alunos passou a receber o auxílio, o que pode indicar maior abrangência do programa de Assistência Estudantil no âmbito do Campus. Tal fato pode ter ocorrido por variadas razões como, por exemplo, aumento de verba orçamentária ou ampliação da divulgação dos processos seletivos junto à comunidade acadêmica. É possível conceber, portanto, que a inclusão deste atributo no cálculo do IVS provavelmente não traria mais ganhos no que se refere à encontrar alunos que apresentam maior vulnerabilidade social. Por fim, quando os dados da situação de moradia dos alunos são analisados, observa-se que a maior parte das solicitações deferidas e indeferidas foi para alunos que moram com familiares, seguido daqueles residentes na Moradia Estudantil; neste caso, faria sentido a inclusão deste indicador no cálculo do IVS para que este aspecto fosse considerado na pontuação final do aluno e auxiliasse no deferimento de seu pedido de auxílio socioeconômico, em especial pelo fato da Moradia já oferecer outros benefícios para seus residentes.

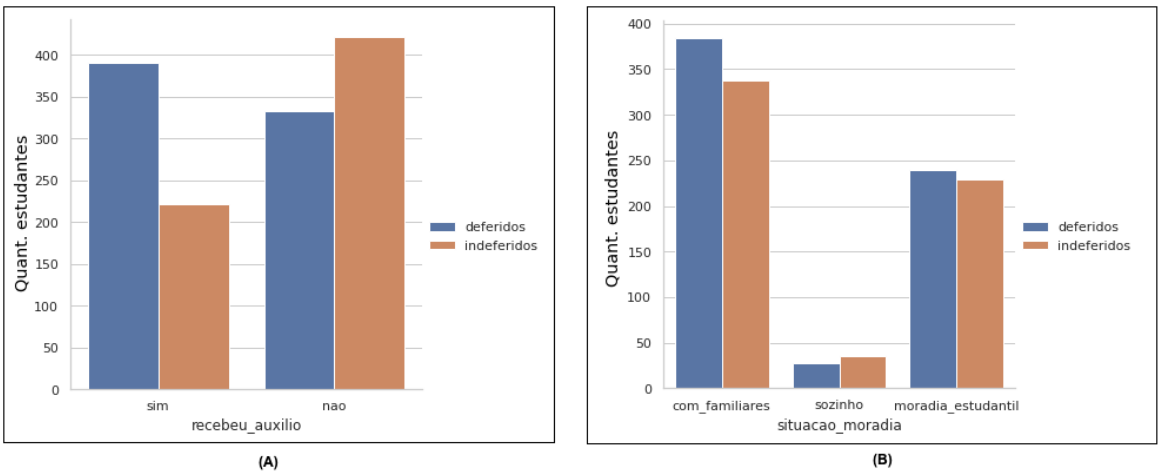

Fig. 3: Situação final por recebimento de auxílio (A) e moradia (B)

5. CONCLUSÃO E TRABALHOS FUTUROS

Este artigo se baseou no processo de seleção da Assistência Estudantil do IFMG para a concessão de auxílios socioeconômicos. Técnicas de MD foram combinadas e aplicadas, permitindo identificar a importância dos atributos cujos valores caracterizam os estudantes que demandam os auxílios. Este resultado habilita os gestores da instituição a analisarem a coleta dos dados feita atualmente sob dois aspectos: quais atributos poderiam ser eliminados do questionário preenchido pelo estudante, pois não apresentam capacidade de melhorar o processo de classificação, e quais atributos deveriam passar a ter maior importância nas análises conduzidas pelos Assistentes Sociais, pois podem contribuir de forma mais significativa para a compreensão da necessidade de cada estudante. O resultado desta etapa do artigo também indicou que o grupo com os dez atributos mais importantes possui a melhor média de *MacroF1* para classificação correta das instâncias, o que sugere a necessidade de estudos mais aprofundados pelos gestores no sentido de avaliar a real necessidade de coleta e análise de alguns atributos. Outro objetivo do artigo consistiu em verificar se o índice atualmente utilizado pelo IFMG (denominado IVS) indica adequadamente aqueles estudantes com maior demanda para recebimento dos auxílios, o que foi confirmado a partir da execução do processo de classificação. Conforme apresentado, o resultado obtido com este atributo se mostrou equivalente ao gerado pelo grupo com os dez atributos mais importantes. O maior diferencial deste artigo é justamente utilizar o levantamento da importância de cada atributo para propor a incorporação de alguns deles no cálculo do IVS, aumentando a capacidade deste indicador abarcar outras características dos estudantes até então não consideradas. A partir do momento em que o IVS foi combinado com os cinco atributos mais importantes que ainda não são usados por ele em sua fórmula, o resultado da classificação foi melhorado em mais de 10%, o que poderia contribuir ainda mais para a atuação dos Assistentes Sociais no que se refere às suas atividades dentro do processo de seleção.

Como a tarefa de classificação foi baseada em apenas um algoritmo (*Random Forest*), sugere-se proceder com estudos para a adequação de outros algoritmos ao problema, tentando otimizar os resultados encontrados. Além disso, é preciso que o modelo construído também seja testado nos dados de outros *campi* do IFMG, avaliando sua capacidade de generalização em cenários com dados eventualmente diferentes daqueles envolvidos com este estudo.

REFERENCES

Abdi, H. and Williams, L. J. Newman-keuls test and tukey test. *Encyclopedia of research design*, 2010.

BRASIL. Portal da transparência, 2021.

Carrano, D., de Albergaria, E. T., Infante, C., and Rocha, L. Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. Vol. 30. pp. 1321, 2019.

El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., and Mostafa, R. R. Improved feature selection model for big data analytics. *IEEE Access* vol. 8, pp. 66989–67004, 2020.

Melo, E. C., Silva, G. d., and Silva, P. C. L. d. Computerization of the student assistance scholarships selection process: the ifmg experience. *Research, Society and Development* 10 (1), 2021.

Omuya, E. O., Okeyo, G. O., and Kimwele, M. W. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications* vol. 174, pp. 114765, 2021.

Pereira, R. B., Carvalho, A. P. d., Zadrozny, B., and Merschmann, L. H. d. C. Information gain feature selection for multi-label classification., 2015.

Soares, G. C. *SAM - uma abordagem específica de mineração de dados socioeconômicos de alunos do IF Amazonas para apoio ao processo de concessão de assistência estudantil*. M.S. thesis, Universidade Federal de Pernambuco, 2020.

Viegas, F. R., Sandin, I., Salles, T., and Rocha, L. Seleção de atributos agressiva e efetiva usando programação genética. *Revista Eletrônica de Iniciação Científica em Computação* 12 (3), 2012.

Whang, S. E. and Lee, J.-G. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment* 13 (12): 3429–3432, 2020.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., and Saeed, J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends* 1 (2): 56–70, 2020.