PREDIÇÃO DA EVASÃO ESCOLAR NOS CURSOS SUPERIORES DO IFMG – *CAMPUS* BAMBUÍ COM O APOIO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Eduardo Cardoso Melo
IFMG - Instituto Federal de Minas Gerais
eduardo.melo@ifmg.edu.br

Fernanda Sumika Hojo de Souza
UFOP - Universidade Federal de Ouro Preto
fsumika@ufop.edu.br

Edimilson Batista dos Santos
UFSJ – Universidade Federal de São João
edimilson.santos@ufsj.edu.br

Resumo
A evasão escolar é um fenômeno complexo e caracterizado por sofrer influência de diversas variáveis. Os efeitos da evasão são muito danosos aos envolvidos: instituições de ensino perdem ou deixam de receber recursos financeiros, estudantes geram dívidas difíceis de serem quitadas, o mercado de trabalho continua com dificuldades para receber profissionais com maior qualificação e a sociedade perde a oportunidade de desenvolver mais amplamente seus indivíduos. Este artigo buscou aplicar técnicas de Aprendizado de Máquina utilizando dados de estudantes de cursos superiores do IFMG – Campus Bambuí com o intuito de compreender melhor as características da evasão nesta instituição e tentar predizer a probabilidade de os atuais matriculados evadirem ou conseguirem concluir o curso. Tendo as atividades baseadas na metodologia CRISP-DM, foi possível construir um modelo de classificação com cerca de 86% de acurácia para predizer o status final de cada estudante.
Palavras-chave: Aprendizado de Máquina. Inteligência Artificial. Evasão Escolar.

Abstract
Dropping out of school is a complex phenomenon characterized by being influenced by several variables. The effects of school dropout are very harmful to those involved: educational institutions lose or stop receiving financial resources, students generate debts that are difficult to pay off, the labor market continues to struggle to receive professionals with higher qualifications and society loses the opportunity to develop more widely its individuals. This article aimed to apply Machine Learning techniques using data of students from higher education courses at IFMG – Campus Bambuí in order to better understand the characteristics of dropout in this institution and try to predict the probability of current enrolled dropping out or being able to complete the course. With activities based on the CRISP-DM methodology, it was possible to build a classification model with about 86% accuracy to predict the final status of each student.
Keywords: *Machine Learning. Artificial Intelligence. School Dropout.*

1 Introdução

O ensino superior no Brasil passou por diversas e profundas mudanças nas últimas décadas, fazendo com que mais pessoas tivessem acesso a este nível de estudo a partir do crescimento no número de instituições e na quantidade de cursos ofertados. Em 2020, o país contava com 2.153 unidades de instituições privadas (com 29.713 cursos) e 304 unidades de instituições públicas (com 8.996 cursos). Mesmo com essas possibilidades, o número de ingressantes (3.703.320) ainda era inferior à quantidade total de vagas ofertadas pelas instituições, mais de 19 milhões (INEP, 2020).

Entretanto, apenas esta ampliação do acesso ao ensino superior no Brasil não garante a permanência dos estudantes nas instituições. A massificação do ensino superior trouxe consigo o aumento da evasão escolar, ocorrida por variados motivos, desde aqueles ligados essencialmente com aspectos particulares dos estudantes, até outros relacionados com questões acadêmicas e das próprias instituições (ARAÚJO, SILVA e PEDERNEIRAS, 2021).

O Instituto Federal de Minas Gerais (IFMG) foi criado em 2008 pelo Governo Federal e oferece cursos técnicos, graduações e pós-graduações em 18 *campi* localizados, em sua maioria, na região central do estado. A situação da evasão no IFMG não difere muito da encontrada no país, sendo que o último estudo mais abrangente feito no âmbito da instituição data de 2016, com dados de 2015. Esta análise mostrou que a evasão média entre os cursos superiores era de 28,2% nos cursos de Bacharelado, 36,6% nos cursos Tecnólogos e 44,8% nos cursos de Licenciatura (IFMG, 2017).

As técnicas de Aprendizado de Máquina (AM) podem ser utilizadas para analisar os dados de evasão de instituições de ensino, seja com a proposta de compreender as variáveis associadas a este fenômeno ou visando construir modelos capazes de predizer a ocorrência de novas evasões. Os trabalhos que envolvem a análise de dados de estudantes estão relacionados com uma área de pesquisa conhecida como Mineração de Dados Educacionais, a qual busca aplicar variadas técnicas de Inteligência Artificial e Estatística para gerar conhecimentos a partir do estudo em consideráveis volumes de dados. No caso dos trabalhos que visam compreender o fenômeno da evasão, na última década foram elaborados e publicados diversos estudos, tais como Carrano et al. (2019), Teodoro e Kappel (2020) e Soares et al. (2020), que se valem de técnicas de AM para analisar estatisticamente dados demográficos e acadêmicos de estudantes, a fim de que novos conhecimentos sejam gerados e possam auxiliar os gestores educacionais no planejamento e execução de programas voltados à permanência dos estudantes nas instituições.

Este artigo objetivou aplicar técnicas selecionadas de aprendizado de máquina supervisionado nos dados demográficos e acadêmicos de estudantes de cursos superiores presenciais do IFMG – Campus Bambuí visando construir um modelo capaz de classificá-los e predizer a probabilidade de evasão daqueles atualmente matriculados, bem como identificar os atributos mais associados com este fenômeno no Campus. É importante ressaltar que o intuito não era propor novas técnicas relacionadas com o contexto da Mineração de Dados Educacionais, mas sim aplicar experimentadas técnicas em uma instituição cujos dados ainda não foram trabalhados em modelos baseados em Aprendizado de Máquina como suporte para geração de novas informações que, eventualmente, podem ser úteis aos gestores públicos atuantes no Campus em questão. Entende-se que esta é a principal contribuição do artigo (permitindo, inclusive, sua replicação em outros *campi* da instituição), em conjunto com a identificação dos atributos associados com a evasão. Merece menção, ainda, o fato de terem sido utilizados dados mais recentes do que aqueles vinculados ao último estudo formal do IFMG sobre evasão.

O texto está estruturado em seções para melhor organização dos conteúdos. A Seção 2 apresenta brevemente definições relacionadas com o Aprendizado de Máquina, bem como ilustra o panorama do tema com base em alguns trabalhos correlatos. A Seção 3 descreve os

procedimentos metodológicos adotados, cujos resultados são apresentados e explorados na Seção 4. Por fim, na Seção 5 as conclusões encerram o entendimento acerca do artigo.

2 Referencial Teórico

Esta seção apresenta a definição básica relacionada ao tema de Aprendizado de Máquina e discute objetivamente os conteúdos e resultados de alguns trabalhos que aplicaram técnicas de AM no entendimento da evasão em cursos superiores.

2.1 Aprendizado de Máquina

Monard e Baranauskas (2003) definem o Aprendizado de Máquina como uma área da Inteligência Artificial que visa construir técnicas computacionais sobre o aprendizado e elaborar sistemas que apresentem capacidade de obter automaticamente novos conhecimentos. Esta linha de entendimento é a mesma adotada por Mitchell (1997), ao afirmar que o AM também possibilita o desenvolvimento de novas habilidades e suporta de variadas formas a organização do conhecimento existente. Batista (2003) complementa este entendimento inicial indicando que a capacidade de aprendizado das máquinas se mostra fundamental para a existência do seu comportamento inteligente, o que Sousa (2020) compreende como sendo a capacidade da máquina aprender com um conjunto de dados com a menor quantidade possível de intervenção humana.

Para Cerri e Carvalho (2019), o Aprendizado de Máquina pode ser entendido como uma subárea da Inteligência Artificial que atua na construção de algoritmos capazes de, a partir da experiência adquirida com a análise de dados de treinamento, aprender a executar completamente diversos tipos de tarefas, tais como o agrupamento e classificação de dados. Essas tarefas podem ter objetivos tanto preditivos, isto é, quando objetiva-se prever a ocorrência de determinado fenômeno, quanto descritivos, onde a proposta é analisar os dados para descrever o fenômeno em questão.

Uma categorização mais ampla dos sistemas de aprendizado é feita por Michalski, Bratko e Kubat (1998), dividindo-os em sistemas do tipo caixa preta e sistemas orientados a conhecimento. Os primeiros possuem a capacidade de elaboração da sua representação particular sobre determinado conceito; em outras palavras, seus aspectos internos não podem ser interpretados com facilidade por quem recebe o resultado do processamento, pois o foco não está na explicação das atividades conduzidas ao longo do processo. Já os sistemas da segunda categoria visam justamente promover compreensão, pelos seres humanos, por meio de estruturas simbólicas que representam os resultados obtidos no processamento.

As técnicas de Aprendizado de Máquina mais populares estão vinculadas com o conceito de indução e podem ser compreendidas a partir do contexto do tipo de aprendizagem esperado, sendo divididas em supervisionadas e não supervisionadas. Quando se trata de aprendizado supervisionado, os dados a serem trabalhados na etapa de treinamento já conhecem a classe alvo associada, isto é, eles já sabem os resultados finais a serem encontrados. Isso permite que o algoritmo entenda e, principalmente, aprenda como aquele resultado foi obtido, para que, em seguida, tente prever a saída em outro conjunto de dados cujo valor da classe alvo ainda não foi definido para cada instância. Quando o valor desta classe alvo é discreto, denomina-se que a tarefa é de classificação; quando a classe alvo permite valores contínuos, então a tarefa é conhecida como regressão. O aprendizado não supervisionado busca encontrar e formar possíveis agrupamentos (também chamados de *clusters*) a partir dos dados analisados pelo algoritmo executado, permitindo que seja conduzida uma análise sobre o significado de cada agrupamento no contexto do problema e auxiliando na identificação de classe alvo para as instâncias envolvidas. (MONARD e BARANAUSKAS, 2003)

Uma proposta de metodologia que pode auxiliar na condução de estudos e projetos na área de Aprendizado de Máquina é conhecida como CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Seu objetivo principal é servir como uma espécie de arquitetura reutilizável em projetos de análise de dados em empresas dos mais variados segmentos, oferecendo um modelo de processo padrão e baseado em ciclos iterativos. O CRISP-DM propõe a estruturação do projeto em seis fases, cada uma com um escopo bem definido: **Compreensão do negócio**, visa identificar os problemas organizacionais que demandaram a iniciação do projeto, entender as demandas dos envolvidos e definir os objetivos a serem atingidos com a análise dos dados; **Compreensão dos dados**, consiste em analisar e entender os dados atualmente disponíveis na organização, bem como aqueles necessários para a consecução dos objetivos propostos, sendo que esta etapa pode ser realizada em conjunto com a primeira por serem complementares; **Preparação dos dados**, busca adequar os dados obtidos ao padrão mínimo requerido para a construção dos modelos, utilizando técnicas de seleção, limpeza, adequação, derivação de atributos, dentre outras; **Modelagem**, procura selecionar e aplicar diversas técnicas de Aprendizado de Máquina tendo como base os objetivos traçados anteriormente, realizando testes e calibragem nos parâmetros para a obtenção de melhores resultados; **Avaliação**, objetiva analisar a construção dos modelos e avaliar se os resultados obtidos estão dentro do esperado ou se alguma questão adicional relacionada ao domínio ainda precisa ser considerada; **Implantação**, busca definir o formato mais adequado para que os resultados do modelo escolhido sejam interpretáveis pelos usuários e disponibilizar a solução em um ambiente de produção (WIRTH e HIPP, 2000).

Em função de estudos acadêmicos que aplicaram técnicas de Aprendizado de Máquina na análise e predição da evasão escolar indicarem bons resultados com os algoritmos *Random Forest* e *Support Vector Machine* (SVM), como mostrado na seção 2.2 deste artigo, ambos foram escolhidos para a construção do modelo de predição da evasão escolar no IFMG – Campus Bambuí e são brevemente abordados a seguir.

Para Guenther e Schonlau (2016), *Support Vector Machine* é uma técnica de Aprendizado de Máquina supervisionado que atua, prioritariamente, em tarefas de classificação de dados em diferentes grupos, mas também encontra bons resultados em tarefas de regressão. Seu funcionamento é baseado na análise e criação de um espaço entre duas classes que permita a previsão de outros vetores de atributos, conhecido como hiperplano. A proposta é que ele possua a maior distância entre duas classes. Desta forma, a técnica permite a identificação da classe que uma nova instância ocupará quando a mesma for analisada pelo algoritmo, tendo como base o lado da reta onde ela está.

De acordo com Tan, Steinbach e Kumar (2014), *Random Forest* é uma técnica de classificação baseada no conceito de árvores de decisão, sendo que o seu grande diferencial é a capacidade de construir e combinar a predição feita por diversas árvores, dentro de um conjunto denominado como floresta, para tentar encontrar melhores resultados em termos de acurácia. Os autores ressaltam o fato de que as árvores de decisão criadas na floresta trabalham com vetores independentes e definidos aleatoriamente a partir da fonte inicial dos dados, além da possibilidade de também utilizar o algoritmo em tarefas de regressão. Uma possível desvantagem deste algoritmo ocorre quando o modelo demanda uma grande quantidade de árvores para realizar predições com maior acurácia.

2.2 Trabalhos relacionados

Em um trabalho recente e abrangente em termos de objeto de estudo, Teodoro e Kappel (2020) estudaram o fenômeno da evasão escolar no contexto das Instituições de Ensino Superior (IES) públicas no Brasil. A proposta era identificar as características mais determinantes para a

evasão e, assim, tentar predizer o possível abandono de outros estudantes com características semelhantes. Cinco algoritmos de Aprendizado de Máquina (*Naive Bayes*, *K-Nearest Neighbors*, Árvores de Decisão, *Random Forest* e Redes Neurais) foram utilizados na construção de modelos com um conjunto de dados obtido junto ao INEP. Como principal resultado, o estudo indicou que a evasão nas IES públicas está mais relacionada com a idade, com a carga horária total do curso escolhido e com eventuais participações em atividades extracurriculares. Os algoritmos *Random Forest* e Redes Neurais apresentaram o melhor desempenho em termos de acurácia como modelos preditores, alcançando quase 80% na predição de casos de evasão.

A evasão de estudantes de cursos de graduação do Instituto Federal do Maranhão foi analisada por Gonçalvez, Silva e Cortes (2018) com o objetivo de identificar aqueles com tendência de abandonar a instituição. Baseados na metodologia proposta pelo processo KDD (*Knowledge Discovery in Databases*), os autores empregaram os algoritmos *Naive Bayes*, *Support Vector Machine* e J48 para a construção dos modelos de aprendizagem. Como principais resultados, o estudo apontou que o algoritmo J48 obteve melhor acurácia, seguido do SVM e, por fim, do *Naive Bayes*. Ressalta-se, entretanto, que a média final dos três algoritmos foi superior a 94%. É interessante notar que não houve diferenças significativas entre as classificações realizadas pelos três algoritmos, nem mesmo quando técnicas de seleção de atributos (*Information Gain* e *Correlation Based Feature Selection*) foram aplicadas no conjunto de dados. Tal situação pode ser explicada pelo amplo trabalho de pré-processamento dos dados realizado pelos autores antes da aplicação das técnicas de Aprendizado de Máquina.

Carrano et al. (2019) analisaram os dados acadêmicos, demográficos e socioeconômicos dos estudantes da Universidade Federal de São João del-Rei (UFSJ) e construíram um modelo preditivo capaz de indicar indivíduos em risco de evasão. Além disso, o estudo identificou os atributos que naquele momento apresentavam maior relevância para a ocorrência da evasão, sendo estes ligados a questões acadêmicas, tais como o desempenho e assiduidade dos estudantes. Um diferencial deste trabalho é que, além da análise do histórico dos dados da instituição, os pesquisadores também aplicaram o modelo concebido nos dados dos atuais estudantes, indicando aqueles que demandam acompanhamento por parte da gestão organizacional. Os autores sugerem o uso de algoritmos baseados em árvores de decisão para replicação da metodologia proposta, em função de apresentarem desempenho satisfatório e simplicidade de aplicação.

Soares et al. (2020) focaram no contexto da Universidade Estadual do Tocantins em sua pesquisa para compreender o padrão dos estudantes evadidos e não evadidos da instituição com o apoio de algoritmos de Aprendizado de Máquina. O algoritmo *Random Forest* foi utilizado como base para a classificação pretendida. Um ponto interessante do trabalho é a apresentação dos resultados em relação à diversas métricas, não apenas a acurácia. Destaca-se a baixa variação nos resultados entre aqueles obtidos com os dados de treinamento e aqueles com a parametrização aprimorada pelo processo de *Grid Search*, possivelmente por já estarem altos (próximos de 100%).

Um exemplo de estudo internacional sobre a aplicação de técnicas de Aprendizado de Máquina foi conduzido por Berka e Marek (2021) em uma universidade da República Tcheca. Os autores utilizaram tanto dados coletados dos estudantes no momento da admissão à instituição quanto dados acadêmicos relativos ao primeiro semestre letivo. A proposta era identificar/predizer quais estudantes teriam maior tendência de terminar os estudos com sucesso. A metodologia CRISP-DM embasou a organização das etapas do estudo. Os algoritmos de classificação utilizados foram baseados em árvores de decisão, os quais facilitam o entendimento dos resultados propostos pelos modelos construídos, segundo os autores. Dentre os resultados apresentados, destaca-se o baixo impacto das características dos estudantes no momento do seu ingresso na instituição na classificação final como concluinte ou desistente. Além disso, as análises mostraram que quanto maior o tempo entre o término dos estudos de nível secundário e o ingresso no ensino superior,

maiores são as chances de desistência. Apesar da publicação do trabalho ser recente, os dados das análises são de 2013 e 2014, fazendo com que seja complexo construir eventuais contextualizações com a situação atual em função de possível mudança no perfil dos estudantes neste período.

O diferencial deste artigo quanto aos trabalhos abordados nesta seção é que, até o momento, a aplicação de técnicas de Aprendizado de Máquina utilizando dados do IFMG – Campus Bambuí com o intuito de aumentar a compreensão sobre o fenômeno da evasão nos seus cursos superiores não foi descrita em outros trabalhos científicos, motivando, portanto, a realização deste estudo.

3 Metodologia

Este artigo foi delineado e conduzido como um estudo de caso, tendo como objeto de pesquisa os estudantes de cursos superiores de graduação presencial oferecidos pelo IFMG – Campus Bambuí, com o objetivo de analisar o fenômeno da evasão escolar a partir da aplicação de técnicas de Aprendizado de Máquina. A metodologia CRISP-DM serviu de base para estruturar e organizar as etapas de trabalho, bem como as atividades necessárias para o alcance do objetivo proposto. É importante ressaltar que a etapa de implantação proposta por esta metodologia não fez parte do escopo deste artigo.

3.1 Compreensão do negócio

A primeira etapa consistiu no entendimento das questões relacionadas com a evasão escolar, em especial sobre como ela vem sendo conduzida na instituição especificada. Foi possível observar que o último estudo formal feito pelo IFMG sobre este assunto ocorreu em 2016, motivo pelo qual as análises deste artigo podem contribuir para que os gestores tomem conhecimento do contexto após este período. Conforme apresentado na seção introdutória, o percentual de evasão no IFMG era alto, porém tratava-se de um indicador geral da instituição (composta atualmente por 18 *campi*). A realização deste estudo possibilitou compreender e detalhar a realidade específica do Campus Bambuí quanto à evasão dos estudantes de cursos presenciais de graduação.

3.2 Compreensão dos dados

Nesta etapa foram analisados os dados estruturados disponíveis dentro do ambiente institucional, isto é, aqueles que atualmente são mantidos e tratados com o apoio de sistemas de informação. O IFMG utiliza um sistema de gestão acadêmica fornecido pela empresa TOTVS, o qual armazena tanto dados demográficos quanto acadêmicos dos estudantes, bem como dos cursos e disciplinas oferecidos pelos *campi*. No caso do Campus Bambuí, o sistema armazena dados desde quando foi implantado oficialmente em 2013. A Diretoria de Assuntos Estudantis forneceu, ainda, planilhas eletrônicas contendo dados de beneficiários de bolsas de auxílio socioeconômico. Como o conjunto de dados disponível possibilitava a identificação de estudantes formados, evadidos e matriculados, a continuidade da pesquisa considerando a aplicação das tarefas de classificação se mostrou viável.

3.3 Preparação dos dados

O acesso aos dados utilizados neste trabalho foi autorizado pela diretoria do IFMG - Campus Bambuí e disponibilizado pela Coordenadoria de Gestão da Tecnologia de Informação (CGTI) após a formalização do pedido para fins de pesquisa acadêmica. Os dados foram anonimizados de forma a evitar identificações. Ao final, foram gerados dois arquivos .CSV, um com dados demográficos dos estudantes vinculados a cursos superiores presenciais (turmas de ingressantes entre 2013 e 2019) e outro com dados das disciplinas cursadas por esses estudantes.

O arquivo com dados demográficos possuía 21 colunas caracterizadoras com 1.891 linhas que representavam os estudantes. O arquivo com dados acadêmicos possuía 12 colunas e 81.518 linhas representando as disciplinas cursadas. Os dados foram importados para uma base de dados *MySQL* composta de duas tabelas construídas com os mesmos atributos vindos dos arquivos. As manipulações foram feitas utilizando o software *phpMyAdmin* versão 5.0.3 em conjunto com scripts *Python* versão 3.8.

Inicialmente, 17 registros de estudantes foram excluídos da base por não conterem valor no atributo que indica o status atual no curso, dado essencial para a realização da pesquisa proposta. Na sequência, os valores de cada atributo foram analisados para identificar problemas ou incoerências. Outros 22 registros de estudantes foram excluídos por estarem duplicados na base de dados. É importante ressaltar que todos os registros de dados acadêmicos dos estudantes excluídos também foram eliminados da base de dados. Quanto aos dados faltantes, como a maioria dos campos no sistema acadêmico é de preenchimento obrigatório, a análise de dados faltantes foi pequena, se restringindo a apenas um atributo (forma de ingresso no IFMG). Optou-se por preencher os 3 registros que tinham esse dado faltante com o valor "Vestibular", pois este era o valor encontrado em mais de 50% dos registros. A correção de dados não foi necessária, pois o preenchimento da maioria dos dados no sistema acadêmico é feito por meio de seleção em caixas de combinação, e não em campos de preenchimento livre pelo usuário, o que evita a entrada de dados inconsistentes ou redundantes.

Os dados que indicam se o estudante recebeu auxílio socioeconômico foram obtidos a partir de outra fonte, a Diretoria de Assuntos Estudantis do IFMG. Na verdade, apenas os CPFs dos beneficiários e o ano da concessão foram disponibilizados, sendo os mesmos importados para a base de dados. Em seguida, foi criado um atributo na tabela de estudantes para indicar se o mesmo recebeu auxílio socioeconômico em algum momento de seu vínculo com o IFMG.

Alguns atributos foram transformados e agregados para melhor representar os dados. O atributo "turno" foi alterado para que registros com valor "Noturno 50 min" fossem alterados para apenas "Noturno", pois esse valor com a indicação de "50 min" foi utilizado em função da mudança na mensuração do tempo das aulas do Campus e não é mais válido. O atributo "forma_ingresso" possuía valores utilizados em registros antigos denominados "Exame de seleção", os quais foram agrupados com o valor "Vestibular". O atributo "status_aluno_curso" teve os valores "Trancado", "Desligado" e "Transferência" alterados para "Evadido".

Atributos foram criados (e tiveram seus valores calculados) para representar novas informações relacionadas com cada estudante. O primeiro deles armazena o tempo em anos decorrido entre o término do Ensino Médio e o ingresso no curso. A modalidade de cada curso e a classificação de sua área na CAPES também foram armazenados em novos atributos. A partir dos dados acadêmicos dos estudantes, foram obtidas (e armazenadas) as seguintes informações: quantidade de aprovações e reprovações, percentual de aprovação e reprovação nas disciplinas do primeiro semestre cursado, coeficiente de rendimento nas disciplinas do primeiro semestre cursado e a quantidade de períodos cursados. A opção por algumas destas informações refletirem os resultados do primeiro semestre cursado por cada estudante foi tomada em função do último estudo do IFMG sobre evasão indicar que, em média, 68% dos alunos da instituição abandonam o curso no primeiro ano, notadamente no primeiro semestre letivo.

Ao final das atividades de preparação, o conjunto de dados ficou com 11 atributos caracterizadores assim nomeados: *sexo*, *cor_raca*, *forma_ingresso*, *tempo_em_ingresso*, *modalidade_curso*, *area_curso_capes*, *turno*, *carga_horaria*, *cr_primeiro_semestre*, *perc_repr_sem_1*, *recebeu_auxilio*. Além disso, cada instância possui a indicação da classe alvo

com três valores possíveis: evadido, formado ou matriculado. De um total de 1.852 instâncias de exemplos, 801 possuíam a classe evadido, 307 formados e 744 matriculados.

3.4 Modelagem

A proposta desta etapa era (1) construir modelos capazes de classificar as instâncias disponíveis de estudantes evadidos e formados, isto é, aqueles que desistiram do curso ou conseguiram finalizá-lo, e (2) aplicar o modelo com melhores resultados para predizer a evasão dos estudantes que constam atualmente como matriculados. Foram aplicadas duas técnicas de Aprendizado de Máquina com abordagens diferentes de atuação para compreender o fenômeno da evasão nos dados do IFMG – Campus Bambuí, ambas selecionadas em função de sua utilização em estudos desta natureza: *Random Forest* e SVM. O artigo não teve como meta comparar o desempenho geral dessas técnicas, mas sim utilizar os resultados de cada uma delas como auxílio no entendimento das classificações realizadas, bem como dos atributos que mais impactaram na evasão. As atividades desta etapa foram executadas com a linguagem de programação *Python* versão 3.8 no ambiente *Google Colab*, contando com apoio das bibliotecas *pandas*, *scikit-learn*, *matplotlib*, *shap* e *pycaret*.

Para a avaliação dos atributos em maior grau de associação com a classe alvo deste estudo, optou-se pelo emprego do processo conhecido como *Feature Importance* (GUYON e ELISSEEFF, 2003) implementado pela classe *RandomForestClassifier* da biblioteca *scikit-learn*. Quanto à construção dos modelos de classificação com cada um dos algoritmos testados, optou-se pela utilização da abordagem de validação cruzada com vinte *folds* e manutenção da estratificação. As métricas coletadas para análise do desempenho dos modelos foram acurácia, precisão, revocação e Macro F1.

3.5 Avaliação

A avaliação dos modelos construídos foi sendo feita à medida que as classificações eram executadas, permitindo a realização de ajustes e validações. Os resultados finais apresentados se referem àqueles obtidos pela melhor configuração de cada modelo.

4 Resultados

Nesta seção são apresentados os principais resultados obtidos a partir da análise dos dados tratados e de acordo com a metodologia exposta anteriormente.

4.1 Caracterização da amostra

Como o foco do estudo era estudar a evasão nos cursos de graduação presencial do Campus Bambuí, foram analisados dados de seis cursos da modalidade Bacharelado (Administração, Agronomia, Engenharia de Alimentos, Engenharia de Computação, Engenharia de Produção e Zootecnia) e dois de Licenciatura (Ciências Biológicas e Física). A Tabela 1 apresenta a porcentagem da evasão de cada curso e turma pesquisados, considerando o período entre 2013 e 2019, confirmando a necessidade de estudos mais aprofundados sobre as variáveis associadas a este fenômeno.

Tabela 1 – Evasão por curso e turma no IFMG – *Campus* Bambuí

Curso	2013	2014	2015	2016	2017	2018	2019	**Média**
Administração	26%	49%	26%	21%	42%	22%	18%	**29%**
Agronomia	20%	15%	37%	31%	8%	37%	18%	**24%**
Ciências Biológicas	59%	ST	63%	57%	50%	46%	43%	**53%**
Engenharia de Alimentos	ST	53%	54%	63%	67%	56%	56%	**58%**
Engenharia de Computação	50%	48%	33%	35%	36%	52%	19%	**39%**
Engenharia de Produção	50%	55%	50%	35%	47%	25%	58%	**46%**
Física	79%	75%	60%	80%	42%	78%	61%	**68%**
Zootecnia	56%	64%	56%	39%	41%	31%	35%	**46%**
Média	**49%**	**51%**	**47%**	**45%**	**42%**	**43%**	**39%**	**45%**

ST: Sem turma
Fonte: Os autores (2022)

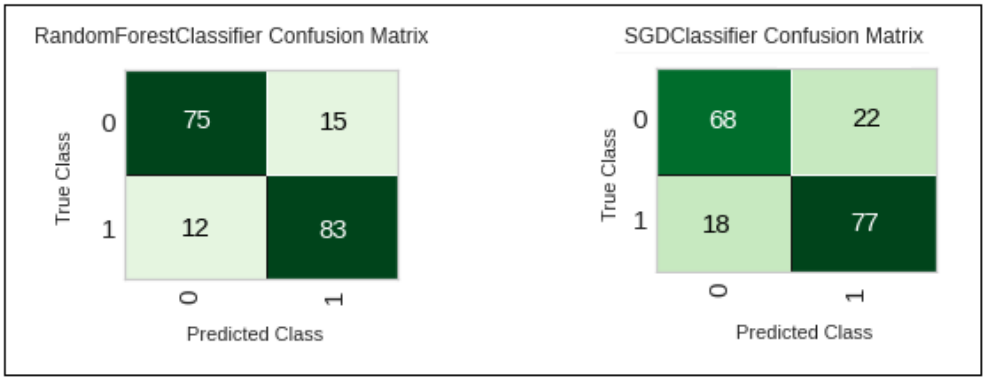
No que se refere ao sexo dos estudantes existentes no conjunto de dados, foi verificado um equilíbrio, pois 50,4% eram do sexo feminino e 49,6% do sexo masculino. Quanto à cor, 48,9% se declararam brancos, 39,1% pardos e 12% negros. A forma de ingresso na instituição ficou concentrada nas opções SISU (39,1%) e Vestibular (51,8%). A maioria estudava integralmente (64,2%), isto é, com atividades letivas pela manhã e tarde, enquanto 35,8% estudavam apenas no período noturno. Por fim, destaca-se que 31,4% dos estudantes receberam auxílio socioeconômico.

4.2 Criação e avaliação dos modelos de classificação

Conforme exposto na seção de Metodologia, o processo de criação dos modelos de classificação foi elaborado considerando os algoritmos *Random Forest* e SVM. Nesta etapa o conjunto de dados foi filtrado para utilizar apenas instâncias de estudantes evadidos ou formados. Como havia um desbalanceamento quanto à quantidade de instâncias para cada uma dessas classes (307 formados e 801 evadidos), optou-se pela aplicação da técnica de *over-sampling* SMOTE (*Synthetic Minority Over-sampling Technique*) com o suporte da biblioteca *Imbalanced-Learn*, fazendo com que a quantidade final de instâncias de evadidos fosse a mesma de formados.

A Figura 1 apresenta a Matriz de Confusão obtida para cada um dos algoritmos aplicados. O resultado do *Random Forest* na classificação correta dos evadidos (classe 0) é superior ao do SVM, assim como na classificação correta de formados (classe 1).

Figura 1: Matriz de Confusão após o processo de predição nos dados de teste
Fonte: Os autores (2022)



Ao analisar as métricas obtidas por cada modelo durante o processo de predição com os dados de teste, conforme apresentado na Tabela 2, é possível notar que o *Random Forest* obteve resultados mais altos em comparação com o SVM. Foi encontrada diferença estatisticamente significativa entre eles ao aplicar o Teste *ANOVA* (F=8,29 e F crítico=3,91) e o Teste de *Tukey* (p=0,027) nos resultados de classificação dos modelos.

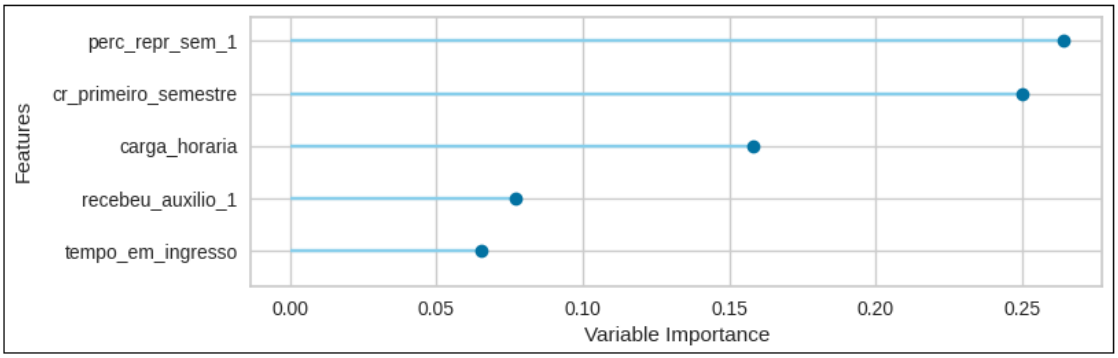
Tabela 2 – Métricas obtidas após o processo de predição nos dados de teste

Modelo	***Acurácia***	***Precisão***	***Revocação***	***F1***
Random Forest	0,8601	0,8382	0,8916	0,8634
SVM	0,7838	0,7184	0,8706	0,7872

Fonte: Os autores (2022)

Na sequência, verificou-se a importância de cada atributo do conjunto de dados para as classificações realizadas pela técnica *Random Forest Feature Importance*. Conforme apresentado na Figura 2, cinco atributos estão relacionados com cerca de 83% de importância na classificação, sendo o percentual de reprovações no primeiro semestre letivo o principal deles, com cerca de 27% de importância. O segundo atributo indicado também apresenta alta importância (25%) e refere-se ao coeficiente de rendimento do estudante durante o primeiro semestre letivo, enquanto o terceiro atributo indica a carga horária total do curso. O penúltimo atributo indica se o estudante recebeu auxílio socioeconômico no ano anterior, enquanto o último trata do tempo em anos entre o término do Ensino Médio e o ingresso na graduação.

Figura 2: Atributos mais importantes para o modelo de classificação
Fonte: Os autores (2022)



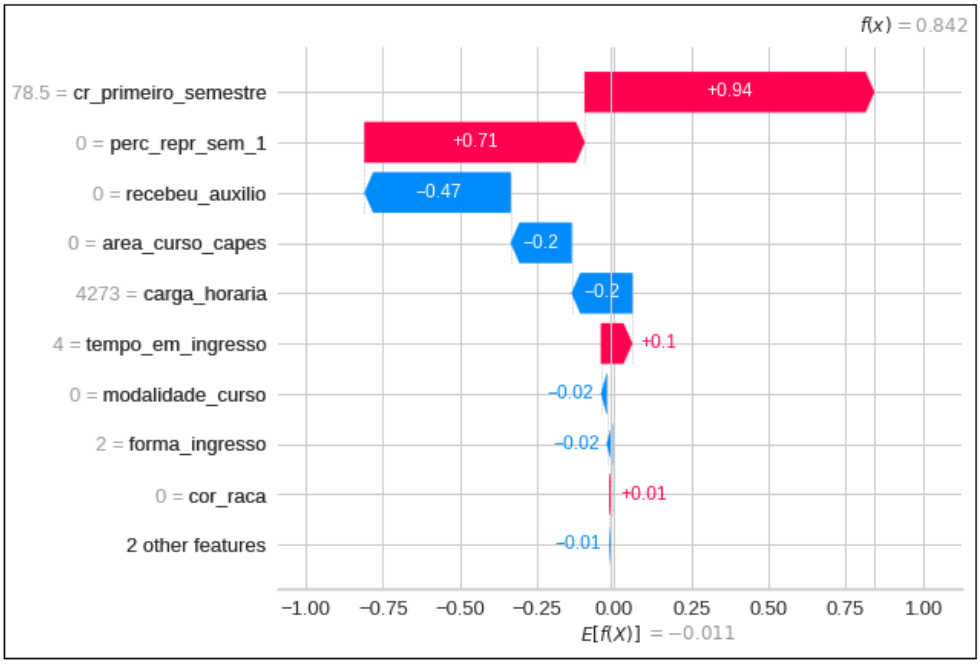
4.3 Predição entre os alunos matriculados

A última etapa do artigo consistiu na aplicação do modelo com melhor desempenho (baseado no algoritmo *Random Forest*) nos dados dos estudantes atualmente matriculados no IFMG – Campus Bambuí, objetivando predizer o nível de evasão entre esse público. Em um total de 744 instâncias de estudantes matriculados, 382 deles (51,3%) foram classificados como evadidos e 362 (48,7%) como formados. O percentual de confiabilidade da predição do modelo obteve média de 87%, com desvio padrão de 4% (valores calculados pela biblioteca *pycaret*).

Analisando a Figura 3 é possível notar o impacto de alguns atributos na predição realizada pelo modelo construído neste artigo. Trata-se de uma instância que foi classificada com alta probabilidade de concluir o seu curso, ou seja, no momento este aluno não possui características similares às daqueles que evadiram desde 2013. Nota-se que os atributos *cr_primeiro_semestre* (média de 78,5 pontos) e *perc_repr_sem_1* (sem reprovações) aumentam a probabilidade de término do curso, enquanto o fato de não ter recebido auxílio socioeconômico faz o papel inverso

em conjunto com os atributos *carga_horaria* e *area_curso_capes*, porém não são suficientes para que a instância seja predita com a classe evadida.

Figura 3: Peso dos atributos na predição
Fonte: Os autores (2022)



5 Considerações finais

A evasão escolar é um dos principais desafios enfrentados pelas instituições de ensino superior no Brasil, especialmente no caso daquelas de natureza pública, onde parte importante do orçamento advém justamente da quantidade de matriculados em cada ano letivo. Sendo assim, é fundamental identificar os fatores que podem estar associados com a saída do estudante e compreender como esta situação pode ser mitigada pela instituição. Uma das contribuições deste artigo consiste justamente em apresentar as variáveis mais importantes que estão associadas com o fato de o estudante evadir ou concluir o seu curso no contexto dos cursos de graduação do IFMG – Campus Bambuí, além do cálculo da possível evasão dos estudantes ainda vinculados à instituição. Neste sentido, importa ressaltar que a taxa de predição de evasão calculada pelo modelo de Aprendizado de Máquina (51,3%) é superior à taxa média de evasão encontrada nos cursos analisados entre o período de 2013 a 2019 (45%), evidenciando a necessidade de atuação organizacional para que medidas possam ser tomadas com antecedência à ocorrência deste evento.

Tendo como base os resultados apresentados neste artigo, sugere-se que a gestão do IFMG – Campus Bambuí construa um planejamento contemplando ações relacionadas com os principais atributos que vêm ocasionando evasão discente. Em relação ao percentual de reprovação e coeficiente de rendimento do primeiro semestre letivo, a instituição poderia elaborar um programa diferenciado para o acompanhamento do desempenho acadêmico de seus estudantes, isto é, que não aguardasse o final do período letivo para proceder com análises de potenciais evasões, ao mesmo tempo em que forneça meios de recuperação da aprendizagem em seu ambiente para aqueles com maior dificuldade. Como a carga horária total do curso também possui importância nesta questão, seria interessante realizar um estudo de eventuais possibilidades de otimização dos conteúdos programáticos, bem como da própria matriz curricular. A ampliação da oferta de auxílio socioeconômico também deve ser buscada pela instituição, pois foi identificada certa contribuição

para que o estudante conclua seu curso entre aqueles que receberam algum tipo de auxílio durante sua trajetória escolar.

Como trabalhos futuros, sugere-se a utilização de outros algoritmos de Aprendizado de Máquina para analisar a possibilidade de melhoria nos resultados obtidos pelo *Random Forest*, além de ampliar o período de análise para abranger o contexto da pandemia de COVID-19 e verificar o impacto na evasão. Devido à limitação dos dados serem relativos a um único campus, pretende-se ainda analisar os dados que contemplam todos os *campi* do IFMG em investigações futuras.

Referências

ARAÚJO, A. C. C.; SILVA, T. F. C.; PEDERNEIRAS, M. M. M. Reflexões sobre evasão na educação superior brasileira: possibilidades de prevenção e intervenção. **Revista Brasileira de Administração Científica**, v. 12, n. 2, 2021.

BATISTA, G. E. A. P. A. **Pré-processamento de Dados em Aprendizado de Máquina supervisionado.** 2003. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2003.

BERKA, P.; MAREK, L. Bachelor's degree student dropouts: Who tend to stay and who tend to leave?. **Studies in Educational Evaluation**, v. 70, p. 100999, 2021.

CARRANO, D.; ALBERGARIA, E. T.; INFANTE, C.; ROCHA, L. Combinando Técnicas de Mineração de Dados para Melhorar a Detecção de Indicadores de Evasão Universitária. **Brazilian Symposium on Computers in Education** (Simpósio Brasileiro de Informática na Educação-SBIE), p. 1321, 2019.

CERRI, R.; CARVALHO, A. C. P. L. F. Aprendizado de máquina: breve introdução e aplicações. **Cadernos de Ciência & Tecnologia**, v. 34, n. 3, p. 297-313, 2019.

GONÇALVES, T. C.; SILVA, J. C.; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. **Revista Brasileira de Computação Aplicada**, v. 10, n. 3, p. 11-20, 2018.

GUENTHER, N.; SCHONLAU, M. Support vector machines. **The Stata Journal**, v. 16, n. 4, p. 917-937, 2016.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157-1182, 2003.

IFMG. **A evasão escolar no IFMG:** diagnóstico e diretrizes da Política Institucional para a permanência e o êxito dos estudantes. 2017. Disponível em: https://www2.ifmg.edu.br/portal/links/relatorio-evasao-completo-rev6.pdf. Acesso em: 10 ago. 2021.

INEP. **Resumo técnico do Censo da Educação Superior 2020.** Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/estatisticaseindicadores/resumotecnicocensodaeducacaosuperior2020.pdf. Acesso em: 01 abr. 2022.

MICHALSKI, R. S.; BRATKO, I.; KUBAT, M. **Machine Learning and Data Mining:** Methods and Applications. Wiley, 1998.

MITCHELL, T. **Machine Learning.** McGraw-Hill, 1997.

MONAR, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes - Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

SOARES, L. C. C. P.; RONZANI, R. A.; CARVALHO, R. L.; SILVA, A. T. R. Aplicação de técnicas de Aprendizado de Máquina em um contexto acadêmico com foco na identificação dos alunos evadidos e não evadidos. **Humanidades & Inovação**, v. 7, n. 8, p. 223-235, 2020.

SOUSA, M. C. C. **Uma análise do algoritmo K-means como introdução ao aprendizado de máquinas.** Dissertação de Mestrado. Universidade Federal do Tocantins, Palmas, Tocantins, 2020.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining.** Pearson, 2014.

TEODORO, L. A.; KAPPEL, M. A. A. 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**, v. 28, p. 838-863, 2020.

WIRTH, R.; HIPP, J. CRISP-DM: Towards a standard process model for data mining. **International conference on the practical applications of knowledge discovery and data mining**, London, UK: Springer-Verlag, 2000.