Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Preditiva em Relação a Evasão do Curso de Ciência da Computação na Universidade de Brasília

Giovana Pinho Garcia

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientadora
Prof.a Dr.a Maristela Terto de Holanda

Brasília
2024

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Preditiva em Relação a Evasão do Curso de Ciência da Computação na Universidade de Brasília

Giovana Pinho Garcia

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof.a Dr.a Maristela Terto de Holanda (Orientadora)
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
CIC/UnB

Prof. Dr. Jonathan Rosa Moreira
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 20 de setembro de 2024

Dedicatória

À minha família, que sempre esteve ao meu lado com apoio e paciência. Aos meus pais, Sylmara e Geraldo por acreditarem em mim e sempre me incentivarem. À minha irmã e meu cunhado Marina e Luís Gustavo, por me darem suporte nos momentos mais difíceis desta trajetória.

Aos meus amigos, Rafael Mascarenhas, Vinícius Gomes, Caroline Queiroz, por toda a parceria e companheirismo nos momentos em que eu mais precisava. Vocês tornaram essa trajetória mais leve e significativa.

Aos meus professores, que compartilharam seus conhecimentos e me ajudaram a crescer tanto profissional quanto pessoalmente. Agradeço especialmente à professora Maristela Terto de Holanda pelo incentivo, orientação e pelas valiosas contribuições ao longo do desenvolvimento deste trabalho. Suas orientações foram essenciais para a conclusão deste projeto.

Análise Preditiva em Relação a Evasão do Curso de Ciência da Computação na Universidade de Brasília

Giovana Pinho Garcia
Depart. de Ciência da Computação
Universidade de Brasília
Brasília, Brasil

Maristela Holanda
Depart. de Ciência da Computação
Universidade de Brasília
Brasília, Brasil

***Resumo*—A evasão no ensino superior é um problema crítico no Brasil, particularmente nos cursos de Ciência da Computação, e esta alta taxa se torna mais preocupante devido a grande necessidade de profissionais qualificados nesta área. Este estudo utiliza dados dos alunos da Universidade de Brasília para aplicar técnicas de mineração de dados educacionais, afim de entender os principais indicadores de evasão. Foram utilizados modelos preditivos com as técnicas de Gradient Boosting Machine (GBM), Support Vector Machine (SVM) e Random Forest (RF). A partir da modelagem dos dados, foram identificados fatores acadêmicos determinantes na probabilidade de evasão dos alunos. Analisando os dados de ingressantes entre o primeiro semestre de 2013 e o primeiro semestre de 2020, foram obtidas acurácias satisfatórias, com valores maiores de 90%. Os principais indicadores de evasão identificados foram o Índice de Rendimento Acadêmico (IRA), a quantidade de vezes que cursou a disciplina introdutória de programação e a forma de ingresso na Universidade.**

***Index Terms*—Mineração de dados educacionais; Taxa de evasão, Modelos preditivos; Evasão na Computação**

***Abstract*—Dropout in higher education is a critical issue in Brazil, particularly in Computer Science courses, and this high rate becomes even more concerning due to the great demand for qualified professionals in this area. This study uses data from students at the University of Brasília to apply educational data mining techniques in order to understand the main indicators of dropout. Predictive models were used with Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Random Forest (RF) techniques. Through data modeling, key academic factors were identified that determine the probability of student dropout. Analyzing data from students who enrolled between the first semester of 2013 and the first semester of 2020, satisfactory accuracy rates were obtained, with values higher than 90%. The main dropout indicators identified were the Academic Performance Index (IRA), the number of times the introductory programming course was taken, and the form of admission to the University.**

***Index Terms*—Educational data mining; dropout rates, predictive models; dropout in Computer Science**

I. Introdução

As desistências de estudantes em instituições de ensino superior brasileiras são um grande problema[1], que afeta o nível de educação de todo o país. Essas altas taxas de evasão prejudicam o processo acadêmico dos alunos e comprometem todo o ecossistema educacional. O fenômeno também implica em perdas de recursos e tempo para os envolvidos no processo de educação[2]. Dessa forma, é essencial identificar e entender as principais causas da desistência de alunos em instituições de ensino superior.

Em especial, o curso de Ciência da Computação enfrenta taxas de evasão significativamente altas no contexto global e nacional[3–6]. Estratégias para mitigar esse problema têm sido cada vez mais discutidas e implementadas, buscando melhorar o sucesso acadêmico dos estudantes [7–10].

Neste contexto, este trabalho é um estudo quantitativo que visa entender quais são as principais características acadêmicas que afetam a conclusão ou desistência no curso de Ciência da Computação da Universidade de Brasília (UnB). O estudo tem como objetivo construir modelos de predição e analisar a importância das variáveis, permitindo identificar quais fatores têm maior impacto na evasão dos estudantes.

Este artigo foi dividido da seguinte forma: a Seção II fornece uma explicação sobre Mineração de Dados Educacionais; a Seção III aponta para alguns trabalhos similares; a Seção IV descreve a metodologia utilizada; a Seção V apresenta e discute os resultados obtidos; a Seção VI relata as limitações deste estudo e a Seção VII traz as conclusões e possíveis trabalhos futuros.

II. Mineração de Dados Educacionais

Mineração de Dados Educacionais ou *Educational Data Mining* (EDM) consiste na aplicação de técnicas de mineração para o contexto de Dados Educacionais[11–13]. O principal objetivo da EDM é identificar padrões e *insights* que possam apoiar a tomada de decisões informadas por parte de educadores, administradores e pesquisadores[14].

Algumas aplicações da EDM estão relacionadas ao desenvolvimento de tecnologias de *e-learning*[15], à comparação no desempenho de diferentes metodologias[16], à clusterização de dados educacionais[17] e às predições de desempenho dos alunos[18].

Em resumo, EDM é uma área interdisciplinar que combina métodos de ciência de dados com questões educacionais. Alguns dos principais algoritmos de aprendizado de máquina utilizados juntamente com a EDM incluem *random forests*, *nearest neighbour*, *support vector machines*, *logistic regression*, *Naïve Bayes*, e *k-nearest neighbour* [19]. Seu potencial de transformar dados em ideias de melhorias faz da EDM uma ferramenta poderosa para promover a inovação e a eficácia no campo da educação.

III. Trabalhos Relacionados

Identificam-se diversos estudos com objetivos relacionados ao presente trabalho. Os pesquisadores buscaram, a partir de aspectos sociais e acadêmicos, encontrar padrões ou elaborar predições quanto ao desempenho dos estudantes nas primeiras disciplinas do curso ou à probabilidade de evasão dos cursos de Ciência da Computação.

Em [20], foi desenvolvido um modelo preditivo para identificar alunos propensos à evasão, utilizando dados de um modelo semestral de autoavaliação dos cursos de graduação da Universidade Federal da Paraíba (UFPB). Foram aplicados os algoritmos de Decision Trees (DT), Random Forest (RF) e Support Vector Machine (SVM). Cerca de 59% dos alunos ativos da UFPB admitidos a partir de 2017 demonstraram probabilidade de abandonar seus cursos nos testes do modelo preditivo proposto.

Alvim et al. [21] apresentam um panorama da evasão nos cursos de graduação da área de Computação no Brasil no período de 2015 e 2019. Com base em dados do Censo da Educação Superior do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), foram investigadas potenciais diferenças na evasão entre cursos e perfis demográficos específicos. A maior diferença encontrada na evasão está entre estudantes que fazem ou não uso de financiamento estudantil.

Silveira et al. [22], fizeram uma análise para prever o desempenho dos alunos na primeira matéria de programação do curso de Ciência da Computação da UnB, separando alunos cotistas e não cotistas. Foram considerados os dados socioeconômicos dos alunos, utilizando os algoritmos de Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) e RF. Os resultados do artigo apresentam uma taxa mais alta de evasão em estudantes cotistas.

Gomes et al. [23] utilizaram análise de sobrevivência para identificar os principais fatores acadêmicos e sociais que afetam o desempenho de estudantes de Ciência da Computação da Universidade de Brasília. O modelo utilizado tinha como variável resposta o tempo, em semestres, que o aluno esteve vinculado à universidade. Em concordância com demais estudos no tópico, os autores demostraram uma significativa taxa de evasão, notadamente para alunos com mais de 10 semestres na universidade.

Portanto, verifica-se que a EDM já foi aplicada com sucesso em diferentes estudos focados na evasão e desempenho de alunos de Computação. Neste trabalho, os dados educacionais da UnB serão utilizados para prever a probabilidade de evasão de cada aluno do curso de Ciência da Computação. Serão identificadas as características mais relevantes dos alunos e valores específicos desses atributos que estão mais fortemente associados à evasão.

IV. Metodologia

A metodologia utilizada para a realização desse trabalho foi a *Cross Industry Standard Process of Data Mining* (CRISP-DM) [24]. O método é composto por seis fases principais: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação. A CRISP-DM possui fases cíclicas e pode envolver iterações frequentes, permitindo refinamentos contínuos e ajustes ao longo do projeto. Assim, a metodologia fornece uma estrutura clara e sistemática para a execução de projetos de análise de dados. Ainda, a CRISP-DM é valorizada pela possibilidade de aplicação em uma ampla variedade de indústria e domínios [25]. Para este estudo, todas as fases foram utilizadas, com exceção da fase de implementação, mencionada na conclusão e trabalhos futuros.

A. Entendimento do negócio

A Universidade de Brasília foi fundada em 21 de abril de 1962, sendo uma das principais referências acadêmicas nacionais. A diversidade cultural presente em seus quatro campi - Planaltina, Gama, Ceilândia e Asa Norte (campus Darcy Ribeiro) - é uma de suas características marcantes. A pluralidade, aliada à busca permanente por soluções inovadoras, move a produção científica e o cotidiano da instituição [26].

A UnB segue atuante em todas as áreas do conhecimento, aberta às principais demandas do Brasil e do mundo. A Universidade se consolida como organismo indispensável para o desenvolvimento de uma sociedade mais íntegra e democrática [26].

O curso de Ciência da Computação da UnB foi estabelecido para atender à crescente demanda por profissionais qualificados na área de tecnologia da informação. Desde sua criação, o curso tem evoluído para acompanhar as rápidas mudanças e avanços na área de computação. O curso é diurno, presencial e possui uma habilitação que confere o grau de Bacharel. Para se graduar neste curso, o aluno pode permanecer de 8 a 14 semestres na Universidade a fim de cursar 214 créditos (3.210 horas-aula) [27].

A alta taxa de evasão de estudantes do curso de Ciência da Computação da UnB, e esta taxa é especialmente preocupante devido à maior necessidade de especialistas qualificados. Entender as razões por trás da evasão e prever quais alunos estão em risco são passos essenciais para garantir que os estudantes completem o curso com sucesso.

B. Entendimento dos dados

A fim de reconhecer as variáveis mais relevantes para a evasão dos alunos do curso de Ciência da Computação, a base de dados utilizada foi disponibilizada pela Secretaria de Tecnologia da Informação (STI) da UnB. A base possui informações de estudantes de todos os cursos da universidade, entre os anos de 1985 e 2023. Após considerar apenas estudantes do curso de Ciência da Computação, a base de dados apresentou 69.943 linhas, com cada linha correspondente a uma disciplina atendida pelo aluno durante sua jornada na universidade.

A presença de cotas é uma variável socioeconômica de interesse neste estudo. Portanto, é importante ressaltar que em 2012 o Congresso Nacional decretou a Lei número 2.711, conhecida como Lei de Cotas. Essa legislação estabelece que as instituições federais de educação superior vinculadas ao

Ministério da Educação reservarão, em cada concurso seletivo para ingresso nos cursos de graduação, por curso e turno, no mínimo 50% de suas vagas para estudantes que tenham cursado integralmente o ensino médio em escolas públicas. Além disso, metade dessas cotas (25% da quantidade total de vagas) deverão ser reservados aos estudantes oriundos de famílias com renda igual ou inferior a 1,5 salário-mínimo per capita, ou para estudantes auto-declarados pretos, pardos ou indígenas [28].

Dentro deste contexto, foram considerados somente os estudantes que ingressaram na UnB a partir do primeiro semestre de 2013, que foi o primeiro período com implementação do sistema de cotas atual. O último período de ingresso considerado foi o primeiro semestre de 2020. Este marco final foi definido com base na duração do curso, sendo os ingressantes do primeiro semestre de 2020 o último período com dados de alunos que conseguiram concluir o curso.

Também foram desconsiderados todos os alunos ativos na universidade ou que faleceram. A filtragem foi realizada pois o grupo de interesse se restringe apenas aos alunos que concluíram ou evadiram o curso.

A filtragem resultou em uma base de dados com 517 alunos, cada um representado por uma linha com as seguintes colunas: IRA, Sexo, Data de Nascimento, Estado de Nascimento, Cota, Tipo da Escola de Segundo Grau, Período que Ingressou na UnB, Forma de Ingresso na UnB, Nome da Disciplina e Forma de Saída do Curso. Cada variável será explicada nos parágrafos a seguir.

A variável IRA (Índice de Rendimento Acadêmico) contém valores numéricos de 0 a 5. O dado é calculado pela UnB a partir das notas em cada disciplina, ponderadas pela carga horária e semestre cursado [29]. Assim, o valor da variável corresponde ao IRA de cada estudante no último semestre em que esteve ativo na universidade.

Em seguida, a coluna ”Sexo” se refere ao sexo do aluno, com valores ”F” (feminino) e ”M” (masculino). A coluna ”Data de Nascimento” possui a data de nascimento de cada aluno. E a coluna ”Estado de Nascimento” possui a UF do estado de nascimento dos alunos.

Sobre o sistema de cotas, atualmente, a UnB reserva 50% das vagas da Universidade para alunos de escolas públicas, como exigido pela Lei de Cotas, 5% das vagas para para candidatos que se auto-declaram negros, e 45% das vagas para candidatos do sistema universal (alunos que não entram pelo sistema de cotas).

Para concorrer pelo sistemas de cotas para Escolas Públicas, o candidato deve ter cursado integralmente o ensino médio em escolas públicas. Esses candidatos são divididos conforme os critérios abaixo:

- Baixa Renda - Estudantes que possuem renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo.
- Alta Renda - Estudantes que possuem renda familiar bruta per capita superior a 1,5 salário mínimo.
- PPI - Estudantes que se identificam como pretos, pardos ou indígenas.
- Não PPI - Estudantes que não se identificam como pretos, pardos ou indígenas.
- PCD - Estudantes com deficiência.

Dessa forma, a coluna Cota possuía os seguintes valores: ”Universal”, ”Escola Púb. Alta Renda-Não PPI”, ”Escola Pública Baixa Renda-PPI”, ”Escola Pública Alta Renda-PPI”, ”Candidato Negro”, ”Escola Púb. Baixa Renda-Não PPI”, ”Escola Pública Alta Renda-PPI”, e ”Escola Púb. Alta Renda-PPI-PCD”.

A coluna do ”Tipo da Escola de Segundo Grau” indica o tipo de escola em que o aluno cursou o ensino médio, categorizada como ”Público” ou ”Particular”. A coluna ”Período que Ingressou na UnB” possui os valores referentes ao semestre que o aluno ingressou na UnB.

Quanto a forma de entrada na UnB, as principais formas são ENEM-UnB/SISU, PAS e Vestibular. O SISU é um sistema que permite aos candidatos utilizar suas notas do ENEM (Exame Nacional do Ensino Médio) para concorrer a vagas em diversas instituições públicas de ensino superior em todo o país. A partir de 2020, a Universidade de Brasília introduziu um processo seletivo específico, substituindo o uso do SISU pelo seu próprio edital, conhecido como ENEM-UnB, para o preenchimento das vagas.

O PAS é um sistema de avaliação seriada composto por três etapas aplicadas ao longo do ensino médio, com maior peso para a última etapa da seleção no 3º ano. Por fim, o Vestibular é um processo seletivo tradicional que abrange provas específicas elaboradas pela própria universidade para avaliar os conhecimentos e habilidades dos candidatos.

Além desse valores, a coluna referente à forma de ingresso na UnB também continha mudança de curso, transferências, convênios, dupla diplomação e matrícula cortesia.

A coluna do ”Nome da Disciplina” possui o nome das disciplinas cursadas pelo aluno no decorrer da sua trajetória acadêmica.

E, por fim, a variável resposta utilizada consistiu na forma de saída da UnB. Após a exclusão dos estudantes ativos e falecidos, as categorias incluíam: ”Conclusão”; ”Integralização de Discente”; ”Abandono”; ”Efetivação de Novo Cadastro”; ”Transferência”; ”Solicitação Espontânea”; ”Cancelamento Judicial”; ”Desligamento”; ”Mudança de Curso”; ”Reprovação em uma Disciplina 3 Vezes”; ”Novo Vestibular”, ”Saída Judicial” e ”Término de Convênio”.

C. Preparação dos dados

Para utilizar cada uma dessas variáveis, foi necessário realizar algumas transformações nos dados. O procedimento visou obter dados mais coerentes e variáveis mais balanceadas, sem a presença de *outliers*.

Todas as transformações e aplicações foram feitas utilizando a linguagem de programação Python (versão 3.10) no ambiente de paginação, o código pode ser visualizado em [30].

Primeiramente, foi tratada a coluna referente à forma de ingresso na UnB. Considerando os valores de entrada mais relevantes numericamente, essa coluna foi agrupada em quatro

categorias distintas: ”ENEM/SISU”, ”PAS”, ”Vestibular” e ”Outro”, que engloba todas as outras formas de ingresso.

Em seguida, a coluna referente ao estado de nascimento foi transformada, categorizando os dados em apenas duas alternativas: nascidos no Distrito Federal e nascidos em outros estados.

Posteriormente, foi transformada a coluna referente às cotas. Para simplificar a análise, esses dados foram reagrupados em quatro categorias: ”Universal” - alunos que não entraram pelo sistema de cotas, ”Candidato Negro” - alunos que entraram pela cota de candidatos que se auto-declararam negros, ”Escola Pública de Baixa Renda” - alunos que entraram pelas cotas de escolas públicas com renda per-capita igual ou inferior a 1,5 salário-mínimo e ”Escola Pública de Alta Renda” - alunos que entraram pelas cotas de escolas públicas com renda per-capita superior a 1,5 salário-mínimo.

As datas de nascimento e de entrada na UnB foram utilizadas para criar uma nova variável, referente à idade em que o aluno ingressou na UnB. O cálculo foi realizado a partir da diferença entre o ano de ingresso na UnB e o ano de nascimento do estudante. A coluna foi dividida em cinco grupos, baseado na classificação em [22]: menores de 17 anos, de 18 a 20 anos, de 21 a 25 anos, de 26 a 30 anos e mais de 30 anos.

A variável ”Nome da Disciplina” foi utilizada para calcular a quantidade de vezes que o aluno cursou a disciplina ”Algoritmos e Programação de Computadores” (APC). Tratando-se da primeira matéria de programação do curso, APC representa um grande desafio para os estudantes [31]. Assim, a variável foi computada a partir do agrupamento do aluno com a coluna de ”Nome da Disciplina”, considerando apenas APC ou a outra disciplina equivalente, ”Computação Básica”. Após o cálculo, verificou-se que os valores variavam de 0 a 8.

De acordo com a política da UnB, alunos que reprovam na mesma disciplina três vezes são desligados da Universidade. Assim, os alunos que cursaram APC mais de três vezes representam casos de reintegração. Dado que alunos que cursaram APC entre cinco e oito vezes somavam apenas 12 casos, para evitar dados desbalanceados, foram considerados apenas os alunos que cursaram a disciplina no máximo quatro vezes.

Também foram excluídos da análise os alunos que não possuem APC ou Computação Básica no histórico (i.e. com valor 0). Optou-se pela exclusão em vista os objetivos do trabalho e os resultados de diversos testes com os modelos preditivos. Esses casos são, majoritariamente, alunos que pediram o reaproveitamento de disciplinas, de modo que essas disciplinas não constam no histórico da UnB, embora tenham sido cursadas em outros cursos ou instituições. Sendo APC uma variável importante no modelo, a ausência deste valor afeta o desempenho geral das predições e pode comprometer a possibilidade de análise das informações obtidas.

As colunas de ”Sexo” e ”Segundo Grau Tipo Escola” não exigiram transformações adicionais. E, finalmente, a variável de forma de saída da UnB foi reagrupada. Com o objetivo de prever apenas conclusão ou evasão, os valores originalmente descritos como ”Conclusão” e ”Integralização de Discente” foram agrupados como ”Conclusão”, e os demais valores foram categorizados como ”Evasão”.

D. Modelagem dos dados

Após a filtragem e transformação dos dados, o *dataset* final continha 393 registros, com 7 variáveis distintas. Cada observação representou características e informações de alunos do curso de Ciência da Computação que ingressaram entre o primeiro período de 2013 ao primeiro período de 2020, e que, no momento da conclusão deste trabalho, não estavam mais ativos na universidade.

Com a formatação desses dados, foi possível aplicar os modelos de predição. Para cada modelo foram consideradas as colunas de ”Forma de Ingresso na UnB”, ”Cota”, ”IRA”, ”Sexo”, ”Idade”, ”Estado de Nascimento”, ”Tipo Escola” e a quantidade de vezes que o aluno fez a primeira matéria de programação (”APC”).

Os algoritmos utilizados foram Gradient Boosting Machine (GBM), Support Vector Machine (SVM) e Random Forest (RF). A escolha dos algoritmos foi baseada no sucesso previamente reportado em [32–35]. Os algoritmos foram implementados utilizando 70% dos dados como treinamento e 30% como teste de acordo com [22].

Em seguida, foi realizada uma análise da importância de cada variável. Objetivou-se identificar quais atributos mais influenciam o desempenho dos estudantes no curso de Ciência da Computação e a probabilidade de evasão. A análise de importância das variáveis foi implementada a partir do modelo RF, que apresentou maior acurácia preditiva. Assim, os resultados calculados foram a importância de cada valor dentro de uma categoria e o peso de importância total da categoria.

Além da construção dos modelos, foram calculadas as taxas de evasão dos estudantes para cada valor dentro de cada variável, afim de entender quais valores estão mais associados à evasão do curso. Assim, a taxa foi calculada a partir do percentual dos alunos que evadiram em relação ao total de cada grupo.

E. Avaliação

O objetivo dessa fase é garantir a qualidade dos modelos gerados. O desempenho dos modelos foi avaliado a partir da acurácia, que é calculada pela divisão do número de predições corretas pelo total de predições.

Além disso, foi também analisado a eficácia deste modelo e da análise das importâncias para trazer soluções para os problemas da alta taxa de evasão do curso de Ciênia da Computação da UnB.

V. Resultados e Discussão

Após a realização de todas as filtragens e transformações, o número total de estudantes analisados durante o trabalho foi de 393, com distribuições diversas entre as 7 variáveis observadas (Tabela I). Desses alunos, 153 concluíram o curso e 240 evadiram, resultando em uma taxa de evasão de 61,06%.

Todos os modelos preditivos apresentaram acurácias acima de 90% (Tabela II). O algoritmo GBM resultou em uma

Tabela I: Tabela com a quantidade e porcentagem equivalente de cada variável

Variável	Valor	Qtd	(%)
Sexo	F	39	9,9%
	M	354	90,1%
Idade	Ate 17	27	6,9%
	Entre 18 e 20	273	69,5%
	Entre 21 e 25	61	15,5%
	Entre 26 e 30	20	5,1%
	Mais de 30	12	3,0%
Estado de Nascimento	DF	270	68,7%
	Fora DF	123	31,3%
Cota	Candidato Negro	23	5,9%
	Pub Alta Renda	85	21,6%
	Pub Baixa Renda	43	10,9%
	Universal	242	61,6%
Tipo Escola	Pública	160	40,7%
	Particular	233	59,3%
Forma de Ingresso UnB	PAS	112	28,5%
	ENEM/SISU	23	5,9%
	Vestibular	160	40,7%
	Outros	98	24,9%
APC	1	168	42,8%
	2	153	38,9%
	3	29	7,4%
	4	43	10,9%

acurácia de 91,11%, o algoritmo SVM, 91,45% e, com o melhor desempenho obtido, o RF com 92,31% de acurácia. Os resultados dos modelos revelam que as características dos alunos utilizadas nos algoritmos impactam significativamente a probabilidade de evasão do curso de Ciência da Computação na UnB.

Tabela II: Acurácia dos modelos de predição

Modelo	Acurácia
GBM	91,11%
SVM	91,45%
RF	92,31%

Uma vez que obteve uma maior acurácia, o algoritmo RF foi utilizado para calcular os valores de importância por variável (Figura 1) e por categoria em cada variável, conforme apresentado na Tabela III, juntamente com a respectiva taxa de evasão.

A partir dos dados de distribuição apresentados na Tabela I e das importâncias e taxas de evasão da Tabela III foram feitas as análises a seguir.

A variável mais relevante para a predição de evasão dos estudantes foi o valor do IRA, com importância de 67% (Figura 1). Os resultados indicam que o aumento do valor corresponde a uma menor probabilidade de evasão do curso.

O segundo dado que mais afetou a predição foi a quantidade de vezes em que o aluno cursou APC ou Computação Básica no decorrer do curso, com importância de 10%. Com valores que variavam de 1 a 4, identificou-se uma maior taxa de evasão (89,65%) em alunos que cursaram as disciplinas 3 vezes. Cabe destacar, contudo, que esse grupo representava uma pequena parcela dos estudantes avaliados (29 de 393).

A distribuição dos dados na amostra apresenta que mais de 80% dos estudantes cursaram APC apenas 1 ou 2 vezes (Tabela I). Os alunos que cursaram a disciplina somente uma vez apresentaram a segunda maior taxa de evasão (78,57%) e foram o grupo mais representativo da amostra com 168 estudantes (42,8%) e, desses estudantes, apenas 68 conseguiram a aprovação na disciplina. Aqueles que atenderam à matéria 2 vezes (38,9% ou 153 estudantes) corresponderam a uma taxa de evasão menor (49,01%).

Também cabe ressaltar a menor taxa de evasão encontrada entre alunos que cursaram APC 4 vezes (16,27%), apesar de representarem uma parcela minoritária da amostra (10,9% ou 43 estudantes). Esse grupo compreende estudantes que após reprovação em APC pela terceira vez, pediram reintegração para retorno ao curso. O processo de reintegração reflete na determinação desses alunos em concluir o curso.

O terceiro fator de maior relevância para a predição foi a forma de ingresso na universidade. Verificou-se que estudantes que entraram na UnB pelo ENEM/SISU apresentaram a menor taxa de evasão (21,73%). Os alunos que ingressaram pelo PAS possuíram 58,92% de taxa de evasão. Os maiores valores foram identificados entre aqueles que entraram na UnB por Vestibular (65%) e outras formas de ingresso (66,32%), que representaram respectivamente 40,7% e 24,9% da base de dados.

A respeito da idade de ingresso na UnB, as pessoas que entraram entre as idades de 18 a 20 e 21 a 26 anos (334 estudantes) têm menor probabilidade de evasão. A menor taxa foi observada nos alunos que ingressaram com 21 a 26 anos (49,18%). Ademais, estudantes com idade de ingresso entre 18 e 20 anos apresentaram a segunda menor taxa de evasão (60,07%) e corresponderam ao maior grupo amostral (69,5% ou 273 estudantes). Por outro lado, os maiores percentuais de evasão foram observados entre os alunos com idades de 26 a 30 (85%) ou mais de 30 anos (100%). Apesar de representarem uma parcela minoritária dos estudantes (5,1% ou 20 estudantes e 3% ou 12 estudantes, respectivamente), o resultado sugere uma necessidade de políticas educacionais para essa faixa etária.

As demais variáveis não apresentaram relevância substancial ao modelo. A utilização do sistema de cotas (3%), o sexo dos estudantes (2,9%), o estado de nascimento (2,8%) e o tipo de escola de segundo grau (1,8%) possuíram, conjuntamente, cerca de 10% de importância. Apesar disso, é possível extrair informações interessantes da análise.

Quanto às cotas, verifica-se que alunos não cotistas apresentam menor probabilidade de evasão (58,26%) em comparação a estudantes cotistas (entre 62,79% e 69,56%). Em relação ao sexo, as mulheres apresentaram menor chance de evasão (43,58%) do que homens (62,99%). Contudo, o público feminino representa menos de 10% dos estudantes (Tabela I). O estado de nascimento implicou em uma taxa de evasão marginalmente superior para alunos que nasceram fora do DF (3,41% de diferença). Ainda, os dados apontam que estudantes que concluíram o ensino médio em escolas públicas têm maior chance de evasão (66,25%).

A partir da análise de importância, identifica-se que cada valor afeta a probabilidade de evasão do curso. É possível, as-

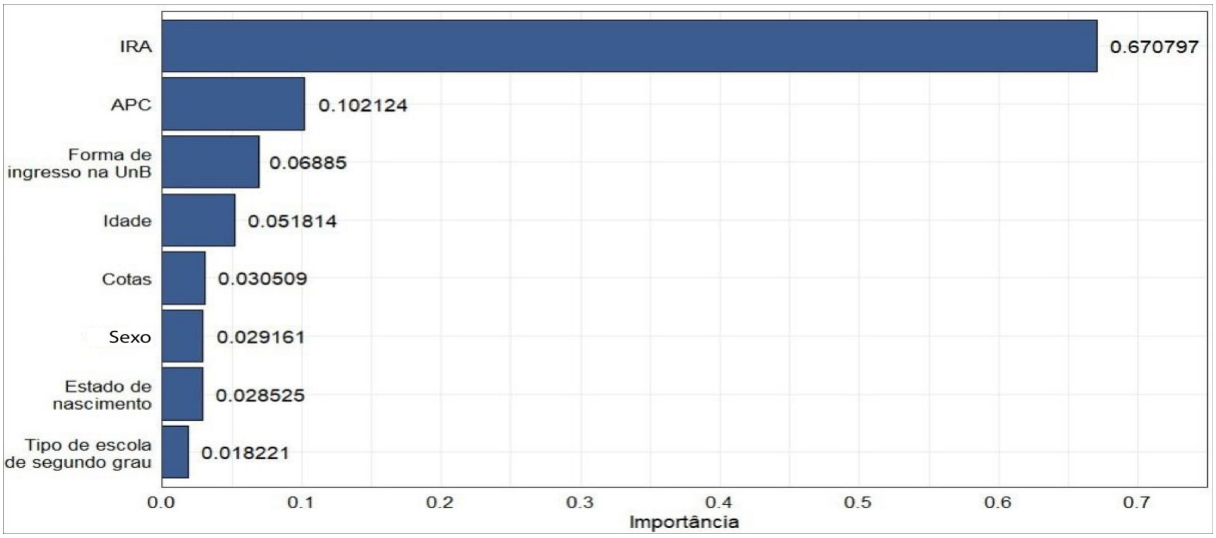

Figura 1: Importância das Variáveis

Tabela III: Taxas de evasão e importância

Variável	Valor	Taxa de Evasão	Importância Individual	Importância da Categoria
IRA	0-5	-	-	0.670797
APC	1	78,57%	-	0.102124
	2	49,01%	-	
	3	89,65%	-	
	4	16,27%	-	
Forma de Ingresso	ENEM/SISU	21,73%	0.022338	0.06885
	PAS	58,92%	0.012954	
	Vestibular	65%	0.018977	
	Outro	66,32%	0.022185	
Idade	Até 17 anos	62,96%	0.008800	0.051814
	Entre 18 e 20 anos	60,07%	0.019372	
	Entre 21 e 25 anos	49,18%	0.017469	
	Entre 26 e 30 anos	85%	0.014637	
	Mais de 30 anos	100%	0.013064	
Cota	Candidato Negro	69,56%	0.005687	0.030509
	Escola Pública Alta Renda	65,88%	0.009172	
	Escola Pública Baixa Renda	62,79%	0.006512	
	Universal	58,26%	0.015779	
Sexo	Feminino	43,58%	0.012928	0.029161
	Masculino	62,99%	0.013307	
Estado de Nascimento	DF	60%	0.017004	0.028525
	Fora do DF	63,41%	0.015779	
Tipo Escola	Pública	66,25%	0.012117	0.018221
	Particular	57,51%	0.009773	

sim, reconhecer padrões nas bases de dados e procurar formas de intervenção para alunos que ingressam na universidade com maiores desafios e maiores probabilidades de desistência.

As acurácias obtidas pelos modelos demonstram que essas características tem correlação relevante com a desistência do curso. A partir da avaliação da importância das variáveis nos modelos, identifica-se também que as principais características que afetam a evasão do estudantes foram: o IRA; a quantidade de vezes que o aluno fez a primeira matéria de programação; a forma de ingresso no curso; e a idade ao entrar na UnB.

Considerando a importância do IRA e da quantidade de vezes que o aluno cursou APC, é essencial implementar um monitoramento maior no caso de alunos com IRAs baixos ou grandes dificuldades nas primeiras disciplinas. Esse monitoramento permitirá proporcionar programas de suporte e tutorias para esses estudantes. O auxílio àqueles que consideram os primeiros semestres mais desafiadores pode reduzir a evasão no curso.

VI. Limitações

A alta taxa de evasão em cursos de Computação é um desafio que envolve diferentes fatores. Neste trabalho, os dados foram analisados com o objetivo de subsidiar políticas de suporte no Departamento de Ciência da Computação para estudantes com mais risco de evasão. Porém, devido ao período analisado, a quantidade de dados disponíveis foi limitada. Além disso, não foram avaliados dados como metodologia do professor, currículo utilizado, moradia e situação familiar do estudante, entre outros.

VII. Conclusão e Trabalhos Futuros

A motivação deste trabalho foi identificar o impacto das características dos alunos de Ciência da Computação da UnB na sua evasão do curso. Além disso, buscamos realizar uma análise detalhada das variáveis influentes e construir modelos preditivos com acurácia satisfatória, capazes de serem aplicados a novos estudantes.

Os modelos preditivos desenvolvidos mostraram acurácias significativas, indicando que os dados podem ser utilizados para monitorar o desempenho acadêmico dos alunos no início do curso. Com isso, torna-se possível calcular a probabilidade de evasão de cada estudante, permitindo uma intervenção precoce e personalizada.

Adicionalmente, ao conhecer previamente as categorias e valores associados à evasão, o Departamento de Ciência da Computação poderia acompanhar os alunos que possuem maior risco de evasão ao decorrer dos primeiros semestres. Durante esses períodos críticos, o corpo docente pode implementar estratégias de estudos mais eficazes e fornecer orientação ativa aos estudantes. Isso resultaria em melhores ferramentas e incentivos para que os alunos enfrentem os desafios iniciais do curso.

Para trabalhos futuros, seria interessante considerar outros fatores da trajetória do aluno na UnB. Cita-se, como exemplo, avaliar o efeito do desempenho em todas as matérias obrigatórias do primeiro semestre do currículo de Ciência da Computação e a participação dos alunos em projetos de iniciação científica. Também se pode explorar diferentes subdivisões das variáveis em estudo, avaliando, por exemplo, se os resultados encontrados são mantidos ao analisar separadamente mulheres e homens.

Referências

[1] D. T. Cusciano, M. M. Laruccia, and L. F. S. Moraes, "Student motivation and 'dropout'rates in brazil," *Employability via Higher Education: Sustainability as Scholarship*, pp. 127–134, 2019.

[2] M. C. M. Lobo, "Panorama da evasao no ensino superior brasileiro: Aspectos gerais das causas e solucoes," *Instituto Lobo para Desenvolvimento da Educacao, da Ciencia e da Tecnologia*, 2012.

[3] C. Stephenson and A. D. Miller, "Retention in computer science undergraduate programs in the us: Data challenges and promising interventions," *ACM*, 2018.

[4] M. N. Giannakos, T. Aalberg, M. Divitini, L. Jaccheri, P. Mikalef, I. O. Pappas, and G. Sindre, "Identifying dropout factors in information technology education: A case study," in *2017 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1187–1194, 2017.

[5] R. M. Hoed, M. Ladeira, and L. L. Leite, "Influence of algorithmic abstraction and mathematical knowledge on rates of dropout from computing degree courses," *Journal of the Brazilian Computer Society*, vol. 24, pp. 1–16, 2018.

[6] G. Greefrath, W. Koepf, and C. Neugebauer, "Is there a link between preparatory course attendance and academic success? a case study of degree programmes in electrical engineering and computer science," *International Journal of Research in Undergraduate Mathematics Education*, vol. 3, pp. 143–167, 2017.

[7] I. O. Pappas, M. N. Giannakos, and L. Jaccheri, "Investigating factors influencing students' intention to dropout computer science studies," in *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '16, (New York, NY, USA), p. 198–203, Association for Computing Machinery, 2016.

[8] R. Takács, J. T. Kárász, S. Takács, Z. Horváth, and A. Oláh, "Successful steps in higher education to stop computer science students from attrition," *Interchange*, vol. 53, no. 3, pp. 637–652, 2022.

[9] B. Herring and R. St Jacques, "Using active learning to increase student retention in introductory computing courses," in *2019 ASEE Annual Conference & Exposition*, 2019.

[10] M. Rahaman, R. Ghosh, I. Dutta, T. Ensari, and R. Cunningham, "Enhancing the programming sequence for undergraduate computer science students: A program for improving learning outcomes," in *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–5, IEEE, 2023.

[11] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *Ieee Access*, vol. 5, pp. 15991–16005, 2017.

[12] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019.

[13] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert systems with applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[14] C. Mehra and R. Agrawal, "Educational data mining approaches, challenges and goals: A review," *JIMS8I-International Journal of Information Communication and Computing Technology*, vol. 8, no. 2, pp. 442–447, 2020.

[15] S. Kausar, H. Xu, I. Hussain, W. Zhu, and M. Zahid, "Personalized e-learning system architecture using data mining approach," *MATHEMATICS & COMPUTER SCIENCE, preprint*, 2018.

[16] N. Valarmathy and S. Krishnaveni, "Performance evaluation and comparison of clustering algorithms used in educational data mining," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7, no. 6S5, 2019.

[17] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data mining for students' disposition analysis," *Education and Information Technologies*, vol. 23, pp. 957–984, 2018.

[18] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[19] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[20] R. dos Santos Oliveira and F. P. A. de Medeiros, "Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação," *Revista Brasileira de Informática na Educação*, vol. 32, pp. 1–21, 2024.

[21] Í. V. Alvim, R. A. Bittencourt, and R. S. Duran, "Evasão nos cursos de graduação em computação no brasil," in *Anais do IV Simpósio Brasileiro de Educação em Computação*, pp. 1–11, SBC, 2024.

[22] R. F. Silveira, M. Holanda, G. N. Ramos, M. Victorino, and D. Da Silva, "Analysis of student performance and social-economic data in introductory computer science courses at the university of brasília," in *2022 IEEE Frontiers in Education Conference (FIE)*, pp. 1–8, IEEE, 2022.

[23] J. B. F. Gomes, M. Holanda, C. C. Koike, M. T. L. Costa, *et al.*, "Study on computer science undergraduate students dropout at the university of brasilia," in *2023 IEEE Frontiers in Education Conference (FIE)*, pp. 1–7, IEEE, 2023.

[24] C. E. Durango Vanegas, J. C. Giraldo Mejía, F. A. Vargas Agudelo, and D. E. Soto Duran, "A representation based on essence for the crisp-dm methodology," *Computación y Sistemas*, vol. 27, no. 3, pp. 675–689, 2023.

[25] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, and P. Flach, "Crisp-dm twenty years later: From data mining processes to data science trajectories," *IEEE transactions on knowledge and data engineering*, vol. 33, no. 8, pp. 3048–3061, 2019.

[26] "A unb," 2024. Acessado em 17/06/2024, Disponível: https://unb.br/estudante/institucional/a-unb.

[27] "Ciência da computação," 2024. Acessado em 17/06/2024, Disponível: https://www.exatas.unb.br/index.php/pt/ciencia-da-computacao-1: :text=O

[28] "Brasil, "lei no 12.711"," 2012. Acessado em 17/06/2024.

[29] "Saa - perguntas frequentes." Acessado em 17/06/2024, Disponível: https://saa.unb.br/perguntas-frequentes.

[30] "Código tcc," 2024. Disponível: https://github.com/giovana100/TCC.git.

[31] J. P. Cohoon and L. A. Tychonievich, "Analysis of a cs1 approach for attracting diverse and inexperienced students to computing majors," in *Proceedings of the 42nd ACM technical symposium on Computer science education*, pp. 165–170, 2011.

[32] P. Sobreiro, P. Pinheiro, and A. Santos, "Performance in the prediction of dropout using the machine learning in sport services," 2018.

[33] F. Janan and S. K. Ghosh, "Prediction of student's performance using support vector machine classifier," in *Proc. Int. Conf. Ind. Eng. Oper. Manag*, pp. 7078–7088, 2021.

[34] S. K. Ghosh and F. Janan, "Prediction of student's performance using random forest classifier," in *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore*, pp. 7–11, 2021.

[35] E. E. Osemwegie and F. I. Amadin, "Student dropout prediction using machine learning," *FUDMA JOURNAL OF SCIENCES*, vol. 7, no. 6, pp. 347–353, 2023.