Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Análise de Caracterização Quantitativa e Predição da Evasão Escolar nos Cursos da Área de Computação do ICEA por meio de Técnicas de *Data Science*

Edgar Henrique Alves Rodrigues

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Alexandre Magno de Souza

Junho, 2022
João Monlevade–MG

Edgar Henrique Alves Rodrigues

Análise de Caracterização Quantitativa e Predição da Evasão Escolar nos Cursos da Área de Computação do ICEA por meio de Técnicas de *Data Science*

Orientador: Alexandre Magno de Souza

Monografia apresentada ao curso de Engenharia de Computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação.

Universidade Federal de Ouro Preto

João Monlevade

Junho de 2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

R696a Rodrigues, Edgar Henrique Alves.
Análise de caracterização quantitativa e predição da evasão escolar nos cursos da área de computação do ICEA por meio de técnicas de Data Science. [manuscrito] / Edgar Henrique Alves Rodrigues. - 2022.
71 f.: il.: color., gráf., tab..

Orientador: Prof. Me. Alexandre Magno de Sousa.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Computação .

1. Aprendizado de computador. 2. Mineração de dados (Computação). 3. Análise descritiva. 4. Análise preditiva. 5. Evasão escolar. I. Sousa, Alexandre Magno de. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.85

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Edgar Henrique Alves Rodrigues

Análise de Caracterização Quantitativa e Predição da Evasão Escolar nos Cursos da Área de Computação do ICEA por meio de Técnicas de *Data Science*

Monografia apresentada ao Curso de Engenharia de Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de bacharel em Engenharia de Computação.

Aprovada em 22 de junho de 2022.

Membros da banca

Prof. Mestre Alexandre Magno de Sousa - Orientador - Instituto de Ciências Exatas e Aplicadas (Icea)
Prof. Doutor Fernando Bernardes de Oliveira - Instituto de Ciências Exatas e Aplicadas (Icea)
Prof. Doutor Luiz Carlos Bambirra Torres - Instituto de Ciências Exatas e Aplicadas (Icea)

Prof. Alexandre Magno de Sousa, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 19/07/2022.

Documento assinado eletronicamente por **Alexandre Magno de Sousa**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 19/07/2022, às 15:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0364522** e o código CRC **8E744AAF**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.009503/2022-51 SEI nº 0364522

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: (31)3808-0819 - www.ufop.br

Este trabalho é dedicado à minha família e a todos aqueles colegas de curso que concluíram, ou não, o curso de Engenharia da Computação.

Agradecimentos

Agradeço primeiramente à meus pais, minha mãe, Vera, por seu amor incondicional e meu pai, Joel, por todo o apoio. À minha irmã, Adrielle, por todos os incentivos e direcionamentos. À minha avó, Vera, por todo o cuidado, carinho e orações. Agradeço à todos meus familiares pela compreensão e por serem meu porto seguro durante todo esse longo percurso de minha vida, mesmo com a distância e saudade crescendo a cada etapa que supero.

Agradeço aos amigos que fiz nessa jornada, meus companheiros da República Bino e meus colegas de curso, por vivenciarem de perto essa fase e por todo o companheirismo proporcionado, seja este nos momentos bons ou nos mais árduos. Agradeço também à meu orientador, Alexandre, por toda a paciência e apoio propiciado para superar as adversidades da vida acadêmica.

Por fim, a todos aqueles que me apoiaram e me motivaram nessa fase da vida, deixo meu muito obrigado e minha eterna gratidão!

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where —" said Alice.
"Then it doesn't matter which way you go." said the Cat.

— Lewis Carroll,
Alice's Adventures in Wonderland

Resumo

A carência de profissionais de áreas relacionadas à Computação é presenciada globalmente, situação que tende a se agravar visto que a demanda de profissionais no setor é crescente e pode intensificar o cenário. Um dos grandes motivos dessa carência de profissionais são as altas taxas de evasão de alunos em cursos relacionados à Tecnologia de Informação. No ICEA, por exemplo, a taxa de evasão é de 69,21%. Dessa forma, aumentar a retenção de estudantes é uma necessidade eminente. Para isso, neste trabalho são investigados modelos de classificação preditivos que auxiliam na identificação de alunos que são propensos à evadir do curso. Foram obtidos dados dos alunos de Engenharia de Computação e Sistemas de Informação da UFOP, dados os quais foram tratados, analisados e caracterizados a fim de revelar o comportamento dos alunos e identificar as principais variáveis que impactam a evasão. Durante o processo de caracterização dos dados, foram realizadas análises de correlação, de componentes principais e de regressão linear múltipla para definir as variáveis que melhor se ajustavam para treinamento e teste dos modelos. Em seguida, foram construídos e validados modelos preditivos por meio de técnicas de aprendizado de máquina como Regressão Logística, XGBoost, SVM, Árvores de Decisão e *Random Forest*. Por fim, foi realizada a etapa de análise preditiva, onde foram aplicados experimentos com diferentes configurações de treino e teste. Os modelos alcançaram resultados satisfatórios, alcançando mais de 95% de acurácia na predição da evasão de discentes em determinados cenários. Dessa forma, o trabalho permite melhor entendimento acerca dos fatores que impactam a evasão no ICEA para os cursos de computação. Além disso, abrem-se oportunidades de investigação da evasão em cursos de outras áreas de conhecimento ou em modelos generalizados, o que pode permitir uma comparação e avaliação de qual nível de generalização é oportuno para o ajuste dos modelos.

Palavras-chaves: Evasão Escolar. Análise Descritiva e Preditiva. Ciência dos Dados. Aprendizado de Máquina. Classificação.

Abstract

The shortage of professionals in areas related to Technology is witnessed globally, a situation that tends to worsen as the demand for professionals in the sector is growing and may intensify the scenario. One of the main reasons for this lack of professionals is the high dropout rates of students in courses related to Information Technology, which is 69.21% at ICEA. Thus, increasing student retention is an eminent need. For this, we developed predictive models that help in the identification of students who are likely to drop out of the course. We obtained data from students of Computer Engineering and Information Systems at UFOP, data which were processed, analyzed and characterized in order to reveal student behavior and identify variables associated with dropout. During the data characterization process, we performed correlation, principal components and multiple linear regression analyzes to define the indicators that best fit for the following steps. Then, predictive models were built and validated using machine learning techniques such as Logistic Regression, XGBoost, SVM, Decision Trees and Random Forest. Lastly, we performed the predictive analysis step, where experiments with different training and test configurations were applied. The models achieved satisfactory results, reaching more than 95% accuracy in predicting student dropout in certain scenarios. Thus, this work allows a better understanding of dropout at ICEA, revealing factors that impacted data variance and student decision. In addition, opportunities to investigate dropout in courses from other areas of knowledge or using generalized models are opened up, allowing the comparison and assessment of which level of generalization is appropriate for the adjustment of the models.

Key-words: Student Dropout. Student Retention. Descriptive and Predictive Analysis. Data Science. Machine Learning. Classification. Prediction.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Brasscom Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação

CDF *Cumulative Distribution Function*

CRA Coeficiente de Rendimento Acadêmico

CRE Coeficiente de Rendimento Escolar

DECSI Departamento de Computação e Sistemas

DT *Decision Trees*

EC Engenharia de Computação

ENEM Exame Nacional do Ensino Médio

GPA *Grade Point Average*

HESA *Higher Education Statistics Agency*

ICEA Instituto de Ciências Exatas e Aplicadas

IES Instituição de Ensino Superior

INEP Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

MEC Ministério da Educação

PCA *Principal Component Analysis*

SI Sistemas de Informação

STEM *Science, Technology, Engineering and Mathematics*

SVM *Support Vector Machines*

TI Tecnologia da Informação

TIC Tecnologia da Informação e Comunicação

UFOP Universidade Federal de Ouro Preto

UnB Universidade de Brasília

Sumário

1 Introdução

A demanda por profissionais qualificados em áreas relacionadas à computação vem sendo crescente ao longo dos anos, dadas as constantes evoluções tecnológicas apresentadas globalmente na era atual a qual Schwab chama de "Quarta Revolução Industrial" (SCHWAB, 2019). De acordo com Fayer, Lacey e Watson (2017), a aréa de *Science, Technology, Engineering and Mathematics* (STEM) é a qual tem a maior demanda de profissionais nos Estados Unidos da América (EUA). Foi previsto que mais de um milhão de novas vagas possam ser abertas até 2024 para ocupações relacionadas à computação nos EUA. No Brasil a situação não é diferente, estudos da Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação (Brasscom) indicam a necessidade de 420 mil profissionais da área de Tecnologia da Informação (TI) entre 2018 e 2024, que se subtraídos pela quantidade de egressos em cursos de Tecnologia da Informação e Comunicação (TIC) resultaria em uma carência de 290 mil profissionais na área até 2024 (BRASSCOM, 2019).

Perante o exposto, a evasão escolar têm recebido atenção especial de universidade e governos, principalmente em relação à cursos superiores na área de TI. Olson e Riordan (2012) fizeram um estudo nos EUA que revelou que menos de 40% dos estudantes que entram em cursos de graduação na área de STEM conseguem concluí-los. Sendo que, especificamente no curso de Ciência da Computação 42,7% dos alunos deixaram o curso superior sem concluí-lo, a qual é a pior das taxas de evasão dentre os cursos analisados.

No Brasil as taxas de evasão também são preocupantes. De acordo com a Brasscom (2019), a taxa de evasão nos cursos superiores e tecnólogos relacionados à TIC, em 2017, foi de 32,7%. Já em relação ao curso de Ciência da Computação da Universidade de Brasília (UnB), em 2014, a taxa de evasão chegou a 55,76% (PALMEIRA; SANTOS, 2014).

Na Universidade Federal de Ouro Preto (UFOP) a realidade não é diferente. Nos cursos de computação ofertados, sendo Ciência da Computação, Engenharia de Computação e Sistemas de Informação, pode-se observar um baixo coeficiente de desempenho médio geral nos cursos e altas taxas de evasão. Dentre esses cursos, os dois últimos são ofertados no campus da cidade de João Monlevade, no Instituto de Ciências Exatas e Aplicadas (ICEA), os quais são alvos deste projeto. No departamento desses cursos, o Departamento de Computação e Sistemas (DECSI), existem estatísticas preocupantes em relação às disciplinas de Programação de Computadores I e Algoritmos e Estruturas de Dados I. Tais disciplinas estão presentes na grade curricular de todos os 4 programas de graduação do campus e possuem índices elevados de reprovação (i.e. igual ou superior a 50%). Ademais, nossos estudos demonstraram que para os ingressantes dos cursos Engenharia de Computação e Sistemas de Informação a taxa de evasão média dos cursos é de 69,21%,

taxa a qual deve ser preocupante para a universidade.

1.1 Motivação e Justificativa

Diante do contexto apresentado, fica claro o impacto que a evasão tem no mercado de trabalho relacionado à Tecnologia de Informação. Olson e Riordan (2012) apresentaram essa preocupação no relatório encaminhado ao presidente Barack Obama em 2012. Nesse relatório é indicada a carência de graduados em STEM para a década de 2021 à 2030, sendo que, para sanar tal demanda, foi indicado diminuir as taxas de evasão escolar nos cursos superiores da área de STEM. Assim, os estudos a fim de prover mais eficácia na retenção dos estudantes são primordiais para a área, a qual é uma das que mais sofre com a evasão escolar.

Visto isso, destaca-se a necessidade de identificar os principais fatores que influenciam e impactam de maneira direta e indireta a evasão nos cursos de computação do ICEA. Para isso, a análise e caracterização dos alunos dos cursos de Engenharia de Computação e de Sistemas de Informação pode revelar características importantes que impactam no decisão dos discentes quanto à evasão do curso. Além disso, a elaboração de modelos preditivos podem contribuir para políticas universitárias de combate a evasão. Por meio das predições dos modelos utilizados, pôde-se identificar os alunos que estão propensos à se evadirem do curso, dado o qual pode ser utilizado para focar em grupos específicos de alunos e proporcionar mais assertividade no combate à evasão.

1.2 Definição do Problema

O problema de pesquisa alvo é o combate a evasão nos cursos relacionados à computação do ICEA, mais especificamente a evasão nos cursos de Engenharia de Computação e Sistemas de Informação da UFOP. Tendo em vista os aspectos apresentados na seções anteriores, o combate à evasão nos cursos relacionados à TIC é essencial para a continuidade do desenvolvimento tecnológico no país, dado que é um dos principais meios de se combater a carência de profissionais no setor.

Portanto, a fim de colaborar no combate à evasão, esse trabalho visa analisar e caracterizar os discentes de forma a identificar características que impactam no comportamento do discente. Acrescido disso, a elaboração de modelos preditivos que identifiquem alunos que estejam predispostos à evadirem do curso cria ferramentas poderosas no combate à evasão.

1.3 Objetivos geral e específicos

O objetivo geral deste trabalho é definir modelos de classificação que possibilitem predizer a decisão de discentes de se evadirem do curso baseado-se no histórico escolar, nas características socio-econômicas e nos dados disponíveis na seção de ensino da UFOP. Para alcançar o objetivo geral, os seguintes objetivos específicos foram realizados:

- Identificar trabalhos relacionados e obter estatísticas acerca do tema;
- Obter dados relativos aos discentes, realizar triagem e selecionar características relevantes para o estudo da evasão;
- Caracterizar os dados: realizar sumarização estatística e análise de correlação;
- Identificar fatores que impactam na evasão e na variação do comportamento dos discentes;
- Construir e validar modelos de classificação que sejam capazes de predizer a evasão do discente.

1.4 Resultados e contribuições

Durante o processo de caracterização dos dados foi possível compreender o comportamento dos dados, identificar variáveis que impactassem na variância dos dados e revelar a real situação da evasão nos cursos relacionados à computação do ICEA. Evidenciou-se que em média 69,21% dos discentes de EC e SI evadem do curso, e que somente 27,44% destes chegam ao final do curso e se tornam bacharéis. Dessas estatísticas, o curso de Engenharia de Computação foi o que apresentou os piores índices em relação à evasão. Apontou-se, também, que mais de 50% dos discentes evadem até o 5º semestre de curso, aproximadamente na metade do curso.

Em relação ao conjunto de dados utilizado, percebeu-se que as variáveis que mais explicam a variação dos dados foram: Código do Curso, Gênero, Ano de Admissão e Classificação no Vestibular. Sendo que, somente a variável Código do Curso explica 21% da variância dos dados. Em relação às variáveis que mais impactaram na decisão dos modelos quanto à evasão, pode-se observar, a partir do modelo XGBoost, que o maior impacto na decisão do modelo foi dado pelas variáveis: Número de Aprovações, Ano de Admissão, Código do Curso, Média no Curso e Pontuação no Vestibular. Esse resultado mostrou, novamente, que os alunos que mais evadem são aqueles dos semestres iniciais dos cursos, e também que o desempenho do discente no curso de graduação e no Vestibular para ingressar na universidade tem papel importante no futuro acadêmico do discente.

Na etapa de predição da evasão foram construídos 5 modelos de classificação e sumarizados seus resultados, dos quais podem ser citados os seguintes modelos: Regressão Logística, XGBoost, *Support Vector Machines* (SVM), *Decision Trees* e *Random Forest*. Os modelos obtiveram altos níveis de acurácia na identificação de discentes que evadiram do curso, sendo que, em testes utilizando 1 ano de registros para avaliar o desempenho preditivo dos modelos. Foi possível obter até 98,70% na predição da evasão em modelos como a Regressão Logística e SVM. Dessa forma, por meio dos resultados obtidos, este trabalho contribui em expor o real cenário da evasão nos cursos de tecnologia do ICEA. Comprovou-se que a evasão do discente está relacionada à fatores como o curso de graduação que o discente está inserido e, também, a fatores relacionados ao desempenho do aluno, como a nota média no curso e a pontuação no vestibular utilizado para ingressar na universidade.

Por fim, os modelos preditivos treinados apresentaram desempenho preditivo superior aos resultados obtidos nos trabalhos apresentados na Seção 2.3 com acurácia muitas vezes superior à 95%. Isso traz a hipótese de que treinar os modelos exclusivamente para cada área de conhecimento pode induzir em um melhor ajuste no modelo e, consequentemente, melhor acurácia, dado que, em sua maioria, os trabalhos relacionados utilizaram um conjunto maior de dados com diversos cursos de áreas distintas.

1.5 Organização da monografia

No Capítulo 2 são apresentadas informações à respeito da evasão no contexto internacional e nacional na primeira seção e a definição do termo evasão para este trabalho na segunda seção. Na última seção são apresentados trabalhos que abordam a evasão em universidades, bem como utilizam modelos de predição a fim de identificar os alunos propensos a evadir do curso. A última seção foi dividida em subseções, sendo cada uma relacionada à um trabalho e a última apresentando as considerações finais do seção. Em seguida, no Capítulo 3, são apresentadas os conjuntos de dados utilizados para a elaboração deste trabalho na primeira seção, seguida da contextualização da evasão especificamente no ICEA e da caracterização dos alunos dos cursos de EC e SI. O Capítulo 4 é dividido em quatro seções. A primeira se refere ao passos executados no pré-processamento de dados, a segunda apresenta as análises de correlação, que é seguida da análise descritiva e, por fim, da análise de predição. Finalmente, o Capítulo 5 trata das conclusões do trabalho. Nele são apresentadas as conclusões, limitações, contribuições e trabalho futuros.

2 Fundamentação Teórica

Este capítulo aborda os conhecimentos necessários para a compreensão deste trabalho. A primeira seção descreve o cenário da evasão no contexto Internacional e Nacional. A Seção 2.2 descreve os principais conceitos relacionados à evasão escolar. Por fim, a seção 2.3 apresenta separadamente trabalhos relacionados à evasão universitária nas Seções 2.3.1, 2.3.2, 2.3.3 e 2.3.4 e expressa a relação desses trabalhos com este na Seção 2.3.5.

2.1 Evasão no contexto Internacional e Nacional

Com a crescente transformação digital no mercado de trabalho presenciada no século XXI, fortemente intensificada na pandemia de Covid-19 em 2020 pelo regime de trabalho remoto, a demanda por profissionais qualificados em áreas relacionadas à computação se torna iminente, dada a necessidade destes para possibilitar tal avanço. No estudo da *U.S. Bureau of Labor Statistics* (FAYER; LACEY; WATSON, 2017), divisão federal estadunidense que faz análises anuais do mercado de STEM, prevê-se que o emprego em vagas relacionadas à computação aumente 12,5% no mercado estadunidense entre 2014 e 2024, bem mais do que qualquer outro grupo relacionado à ciência, tecnologia, engenharia e matemática. Esse aumento resultaria em quase 500 mil novos postos de trabalho para profissionais da área de computação, e que, se somados com o número de empregos de outras áreas que serão substituídos por profissionais de computação, o número de novas vagas de emprego para computação pode chegar à 1 milhão entre 2014 e 2024, como demonstrado na Figura 1.

Dadas as projeções para o mercado de STEM, a hegemonia das vagas relacionadas a computação já era vista em 2015, sendo que de acordo com Fayer, Lacey e Watson (2017), dentre os 10 cargos com maior quantidade de vagas em STEM no ano, 7 deles eram relacionados a empregos na área de computação.

Porém, para que essa enorme quantidade de vagas sejam preenchidas, deve-se aumentar substancialmente o número de egressos nos cursos de ensino superior relacionados as áreas de STEM. Essa preocupação é apresentada pelo governo dos Estados Unidos no relatório encaminhado ao presidente Barack Obama, em 2012, pelo Conselho de Assessores de Ciência e Tecnologia do Presidente (OLSON; RIORDAN, 2012), no qual projeções econômicas demonstram uma carência de um milhão de graduados em STEM para a década de 2021 à 2030. O relatório ainda indica que para sanar tal demanda, deve-se diminuir as taxas de evasão escolar nos cursos superiores da área de STEM.

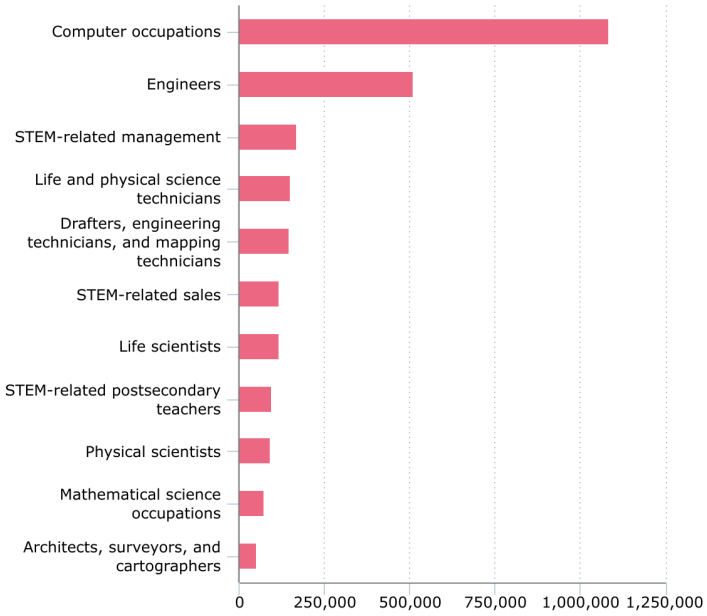


Figura 1 – *Projected job openings for types of STEM occupations, 2014 to 2024.*

Fonte: Fayer, Lacey e Watson (2017)

Diante disso, a evasão escolar têm recebido atenção de muitas universidades e governos. Dados coletados no Reino Unido pela *Higher Education Statistics Agency* (HESA) demonstram que a taxa de evasão nos cursos da área de Ciência da Computação vem decaindo devido aos esforços governamentais e das universidades privadas. A taxa de evasão média em Ciência da Computação no Reino Unido, que era de 20,6% no ano letivo de 2007/2008, decaiu para 9,8% no ano letivo de 2017/2018. Uma queda de 10,8%, que no entanto ainda coloca a área de Ciência de Computação com a maior taxa de evasão no ensino superior do Reino Unido (HESA, 2020).

Em contrapartida, nas instituições de ensino superior estadunidenses as taxas de evasão não são tão animadoras. Olson e Riordan (2012) demonstram que menos de 40% dos estudantes que entram em uma graduação em STEM conseguem concluir a mesma. Após um acompanhamento dos alunos durante 6 anos, tempo geralmente adotados nas pesquisas sobre evasão, foi observado que a taxa de alunos que deixaram o ensino superior sem algum título de conclusão foi de 34,5% em 2009, para aqueles que iniciaram na universidade no ano letivo de 2003/2004. Sendo que a pior taxa ficou para os cursos relacionados à Ciência de Computação, alcançando 42,7% de alunos que deixaram o ensino superior sem concluí-lo, ou seja, evadiram não só de Ciência da Computação, mas do ensino superior como um todo.

Acresce que o cenário é mais agravado ainda ao notar na Figura 2 que apenas 24,6% dos alunos que entraram no ensino superior, na área de Ciência da Computação, concluíram o curso ao final do acompanhamento de 6 anos, sendo que o curso tem duração prevista de 4 anos. Já em 2018, dentre os alunos os quais entraram na universidade no

outono de 2012, a taxa de evasão do ensino superior nos Estados Unidos caiu para 30,6% de acordo com o *National Student Clearinghouse Research Center* (SHAPIRO et al., 2018).

Degree Attainment and Persistence as of 2009						
	Completers		Persisters		Leavers	
Estimates (%)	Attained a STEM degree	Attained non-STEM degree	No degree; enrolled in a STEM field	No degree; enrolled in a non-STEM field	Left postsecondary education without a degree	Total
Total	8.1	42.2	1.8	13.5	34.5	100
Major field of study in 2003-04						
STEM	35.1	21.6	5.7	8.9	28.7	100
• Math	40.3	27.3	0.0	1.9 ‼	30.5 !	100
• Life Sci	37.8	31.9	3.7	9.7	16.8	100
• Physical Sci	41.3	28.1	2.1 ‼	11.5 !	16.9	100
• Eng/ Eng Tech	41.8	16.9	7.2	7.9	26.2	100
• Comp/ Info Sci	24.6	16.7	6.6 !	9.3	42.7	100
• Science Tech	‡	‡	‡	‡	‡	100
Non-STEM	3.1	48.7	1.0	13.6	33.6	100
Undeclared	6.1	39.1	1.6	15.0	38.3	100

Figura 2 – *STEM and Non-STEM Attainment by Major Field of Study in 2003-04.*

Fonte: Olson e Riordan (2012)

Analogamente, estudos da Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação (Brasscom) também demonstram a necessidade de profissionais de tecnologia no mercado brasileiro, de cerca de 70 mil novos profissionais por ano até 2024. O que resultaria em 420 mil profissionais demandados entre 2018-2024, sendo que o país obteve uma média, em 2017, de 46 mil egressos em cursos de Tecnologia da Informação e Comunicação (TIC). Um cenário no qual se estima uma carência de 290 mil profissionais no setor em 2024 (BRASSCOM, 2019).

Ademais, as altas taxas de evasão escolar no país induzem a essa carência de profissionais. Nos cursos superiores e tecnólogos relacionados à TIC a taxa de evasão, em 2017, foi de 32,7% de acordo com a Brasscom (2019). E especificamente no curso de Ciência da Computação da Universidade de Brasília (UnB), por exemplo, a taxa de evasão em 2014 chegou a 55,76% (PALMEIRA; SANTOS, 2014). Assim sendo, constata-se taxas de evasão preocupantes no Brasil, se comparadas ao padrão apresentado no Reino Unido.

2.2 Definição da Evasão

A evasão escolar pode parecer um conceito autoexplicativo, no qual se entende pela evasão do aluno de sua instituição acadêmica. No entanto, existem diversos conceitos que definem a evasão de formas diferentes, podendo ser tanto generalistas quanto com o escopo bem definido e detalhado. Lee e Choi (2011) elaboram um compilado envolvendo 19 definições diferentes adotadas na literatura internacional entre os anos de 1999 e 2009, o que deixa claro que a comparação entre taxas de evasão presentes em publicações diferentes deve ser um trabalho minucioso, a fim de possibilitar um bom entendimento e futura

normalização dos dados para um possível acompanhamento e elaboração de comparações estatísticas.

Vargas[1] (2007 apud Almeida, 2008) explora outros 3 conceitos de evasão, de acordo com a abrangência de cada um, como demonstrado na Tabela 1.

Tabela 1 – Definições de evasão.

Autores	**Definição**	**Amplitude do conceito**
Utiyama e Borba (2003)	Evasão é entendida como a saída definitiva do aluno de seu curso de origem, sem concluí-lo.	Ampla. Não foi estabelecido nenhum critério de tempo no curso para a saída do aluno.
Maia e Meireles (2005)	Evasão consiste em alunos que não completam cursos ou programas de estudo, podendo ser considerada como evasão aqueles alunos que se matriculam e desistem antes mesmo de iniciar o curso.	Especifica que mesmo os alunos que nunca começaram o curso devem ser considerados no cálculo das taxas de evasão.
Abbad, Carvalho e Zerbini (2005)	Evasão refere-se à desistência definitiva do aluno em qualquer etapa do curso.	Não deixa claro se evasão se aplicaria apenas aos alunos que chegaram a iniciar o curso ou se abrangeria também àqueles que apenas se matricularam e nunca iniciaram o curso.

Fonte: Vargas (2007 apud Almeida, 2008)

A UFOP, assim como outras Universidades Federais, utiliza a definição da evasão cunhada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). O INEP é o instituto responsável por censos escolares e avaliações periódicas no sistema educacional brasileiro, o qual define a evasão da seguinte forma:

> *Evasão: saída antecipada, antes da conclusão do ano, série ou ciclo, por desistência (independentemente do motivo), representando, portanto, condição terminativa de insucesso em relação ao objetivo de promover o aluno a uma condição superior a de ingresso, no que diz respeito à ampliação do conhecimento, ao desenvolvimento cognitivo, de habilidades e de competências almejadas para o respectivo nível de ensino. Obviamente, a interrupção do programa em decorrência de falecimento do discente não pode ser atribuída como insucesso, dado que, de forma geral, se trata de caso fortuito e não se pode presumir uma intencionalidade do indivíduo em interromper o curso, cessá-lo ou uma incapacidade do indivíduo de manter-se no programa educacional (INEP, 2016).*

O conceito de evasão do INEP acompanha o conceito de evasão do aluno, descrito por Lobo (2012). Para Lobo (2012), devemos primeiramente explicitar de qual evasão estamos tratando, tendo o seguinte:

> *Em primeiro lugar, ao estudar a evasão do ensino superior é preciso ter clareza e explicitar de qual evasão estamos falando, pois podemos citar*

[1] VARGAS, Miramar Ramos Maia. "Implantação de programas de educação a distância." Material Didático do Curso de Pós-Graduação em Educação a Distância. Centro de Educação a Distância, Universidade de Brasília (2007).

> *alguns diferentes tipos a evasão do curso, a evasão da IES e a evasão do sistema, todas derivadas de diferentes cálculos da evasão dos alunos (LOBO, 2012).*

Sendo assim, podemos separar as evasões nos seguintes tipos, de acordo com Lobo (2012):

- *Evasão do Sistema*: é aquela em que o aluno deixa de estudar e abandona o sistema de ensino, ou seja, não se encontra mais estudando em nenhuma Instituição de Ensino Superior (IES), de qualquer tipo dentro do sistema estudado;
- *Evasão da Instituição de Ensino Superior (IES)*: trata-se da evasão na qual o aluno deixa a IES, mas não deixa o Sistema de Ensino Superior (ou seja, muda de instituição);
- *Evasão do Curso*: é aquela em que o aluno deixa um curso por qualquer razão: muda de curso, mas permanece na IES, muda para outro curso de outra IES ou abandona os estudos universitários;
- *Evasão do Aluno*: essa é a evasão que origina todas as outras, ou seja, a evasão do aluno gera a evasão do curso, da IES, ou do sistema e só é possível medi-la com precisão por meio do acompanhamento da coorte, isto é, do acompanhamento da evolução da situação individualizada de cada estudante.

Diante disso, o objetivo de pesquisa desse trabalho trata da **Evasão do Curso**, sendo analisadas as evasões dos cursos de Engenharia de Computação e de Sistemas de Informação da Universidade Federal de Ouro Preto (UFOP).

2.2.1 Calculo da Evasão

Cientes dos diferentes tipos de evasão, devemos adotar um cálculo padrão para realizar a análise da evasão escolar. O Ministério da Educação (MEC) define a seguinte fórmula para o acompanhamento do percentual de evasão:

$$E = \left(N_i - N_d - \frac{N_r}{N_i} \right) \times 100, \tag{2.1}$$

em que E indica o percentual de evasão, N_i é o número de ingressantes, N_d é o número de diplomados e N_r indica o número de retidos.

Nossa realidade de pesquisa se enquadra bem nessa fórmula e, de acordo com os nossos dados, a Equação 2.1 pode ser simplificada pela Equação 2.2, a qual converge no mesmo resultado em nosso cenário, exemplificado na Tabela 4.

$$E = \frac{N_e}{N_i} \times 100, \tag{2.2}$$

em que E indica o percentual de evasão, N_e é o número de evadidos e N_i é o número de ingressantes.

A taxa de conclusão do curso também é de grande valia para nosso cenário, pois, além de nos oferecer uma perspectiva diferente, possibilita a comparação com dados de outras instituições analisadas na bibliografia. A equação da taxa de conclusão é descrita a seguir:

$$C = \frac{N_d}{N_i} \times 100, \tag{2.3}$$

em que C indica o percentual de conclusão, N_d é o Número de diplomados e N_i é o número de ingressantes

Assim, as análises de evasão do curso e da taxa de conclusão do curso presentes no trabalho, como na Tabela 4, seguem a Equação 2.2 e Equação 2.3, respectivamente.

2.3 Trabalhos Relacionados

A evasão no ensino superior é um tema estudado há décadas, o qual têm ganhado destaque nos últimos anos. Trabalhos como os de Spady (1970), Tinto (1975) e Bean (1980) são frequentemente citados, pois eles foram precursores no tema e conseguiram nos atentar ao alto grau de influência que a evasão escolar provoca na sociedade. Eles fazem uma análise aprofundada sobre o assunto e buscam entender e descrever os motivos pelos quais os alunos tendem à evadir dos cursos de ensino superior em universidades americanas. Dada a importância desses trabalhos no tema, começaram a surgir novas abordagens a fim de predizer a persistência dos alunos nos cursos de ensino superior e assim evitar antecipadamente a evasão. Essa abordagem é vista em Pascarella e Terenzini (1980), no qual foi construído um modelo para predizer a persistência dos alunos do primeiro ano da graduação (*Freshman*) a partir dos modelos teóricos elaborados por Spady (1970) e Tinto (1975).

Nesta seção serão abordados trabalhos relacionados que foram considerados relevantes quanto ao objetivo de pesquisa desse trabalho. Dada a evolução nas tecnologias adotadas, e consequentemente nas abordagens realizadas acerca do tema, como a inserção de *Data Science* e técnicas de *Machine Learning*, foram buscados artigos atuais, que

demonstrassem esses caminhos de exploração. Assim, as Seções 2.3.1, 2.3.2, 2.3.3 e 2.3.4 descrevem artigos relacionados ao tema e a Seção 2.3.5 apresenta as considerações finais relacionando os trabalhos expostos na seção com este trabalho.

2.3.1 Gestão da Evasão na Unisinos: combate e prevenção

No trabalho apresentado pela Universidade do Vale do Rio dos Sinos (Unisinos) (VITELLI, 2010), o problema de estudo tratado foi identificar os alunos que evadem e o porquê dessa decisão. Assim, foram realizadas pesquisas qualitativas com o público envolvido e quantitativas utilizando o banco de dados da instituição. Caracterizou-se o perfil do aluno evadido e foi possível identificar fatores e variáveis associados à evasão. Diversos fatores tiveram grande influência no estudo da evasão, como: Instituição (Qualidade de Suporte provido na mesma); Professores; Cursos de Graduação (se são flexíveis e atualizados); Profissão e Mercado de Trabalho (relevância e valorização da profissão); Desempenho Acadêmico; Aspectos financeiros e sociais do discente. Após essa análise, foi utilizado o modelo estatístico de Regressão Logística para tentar predizer os alunos que possivelmente evadirão do curso. Inicialmente, utilizou-se 26 variáveis no modelo que, após aperfeiçoamentos, resultou em um modelo final utilizando 8 variáveis relacionadas aos discentes para predizer a possibilidade de evasão dos mesmos. Obteve-se boas capacidades preditivas nos modelos e análises importantes quanto ao comportamento específico dos discentes dos Cursos de Informática. Percebeu-se que 30% a 40% desses alunos evadem no primeiro semestre após o ingresso, e 50% a 60% evadem até o terceiro semestre. Além disso, 36% evadem para outro curso na área de Informática, 4% para exatas, 4% para áreas distintas e o restante (56%) evadem da instituição.

2.3.2 Compreendendo e prevendo a retenção e o abandono de graduandos de TI

Platt, Fan-Osuala e Herfel (2019) buscam entender como as variáveis de tipo de estudante e o auxílio financeiro influenciam na probabilidade de completar a graduação, além de quais variáveis são mais correlacionadas ao tempo gasto para se formar. No trabalho, eles utilizam um dataset com 10 anos de registros de estudantes da área de TI (2008–2017), totalizando 10.456 registros. Foram utilizados 3 modelos para a predição da evasão, sendo eles a Regressão Logística, Árvores de Decisão (Decision Trees) e o modelo Naive Bayes. Primeiramente, foram utilizados somente o tipo de estudante e a variável auxílio financeiro como entrada para os modelos. Contatou-se que essas variáveis sozinhas não resultam em bons resultados preditivos nos modelos, porém, foi possível chegar em outras conclusões pela análise dos dados. Observou-se que estudantes não tradicionais (com mais de 25 anos) tem maior probabilidade de não se formar, porém, gastam menos tempo para se formar quando isso ocorre. Já para os alunos bolsistas, percebeu-se que eles

tem maior probabilidade de se formarem, entretanto, em um tempo médio de curso mais longo.

2.3.3 Uma análise comparativa de técnicas de aprendizado de máquina para gerenciamento de retenção de alunos

Delen (2010) desenvolveu modelos analíticos para identificar os estudantes do 1º ano acadêmico que tem maior probabilidade de evadir. Ele identificou as variáveis mais importantes para essa predição aplicando análise de sensitividade nos modelos desenvolvidos. Foi utilizado um dataset com mais de 23.000 estudantes universitários, obtido em 5 anos de dados, o que resultou em 16.066 registros para o uso em seus métodos. Delen utilizou a metodologia de *data mining* CRISP-DM (*Cross Industry Standard Process for Data Mining*) para desenvolver o trabalho, juntamente com 4 métodos de classificação, os quais são: *Multi-layer Perceptron* (MLP), Árvores de Decisão , *Support Vector Machines* (SVM) e Regressão Logística. Após a execução dos métodos em dois datasets distintos, um original e outro balanceado previamente, obteve-se boas capacidades preditivas nos modelos, sendo variável o modelo que melhor previu a evasão dependendo da métrica adotada. Ademais, em seus resultados os fatores mais importantes para a predição foram variáveis relacionadas ao rendimento acadêmico, como coeficiente semestral e créditos aprovados, e o recebimento de bolsas pelo discente.

2.3.4 *Framework* baseado em Análise de Sobrevivência para Previsão Antecipada de Evasão Escolar

Ameri et al. (2016) objetivaram criar um modelo capaz de predizer os estudantes que iriam evadir e em qual semestre isso ocorreria. Para isso, foram utilizados métodos baseados na Análise de Sobrevivência, como o modelo de riscos proporcionais de Cox e o *time-dependent* Cox (TD-Cox). Para realização do trabalho foi utilizado um conjunto de dados com informações de 11.121 estudantes de 2002 à 2009, sendo utilizados dados somente de estudantes em sua primeira graduação, descartando assim os advindos de transferência. Na metodologia utilizada foi optado por realizar uma reestruturação dos dados do aluno por semestre e utilizar uma estratégia para censurar os dados. Dessa forma, é feito o acompanhamento dos alunos somente até o 6º semestre, de modo que os alunos que não evadiram até esse momento tem seus dados censurados. Além dos métodos de Análise de Sobrevivência, foram realizados adicionalmente testes com os métodos AdaBoost, Regressão Logística, Regressão Linear, Árvores de Decisão e SVM. Dentre os resultados, os métodos Cox e TD-Cox obtiveram maior eficiência preditiva que os demais, tanto para predizer a evasão quanto para predizer o semestre que o aluno irá evadir. Além disso, as variáveis que tiveram maior influência na evasão foram: Coeficiente escolar no

Ensino Médio; Coeficiente Semestral; Créditos aprovados; Créditos reprovados; e atributos financeiros.

2.3.5 Considerações Finais

De modo geral, os trabalhos apresentados buscam realizar a predição da evasão escolar em universidades. Quando comparado com os trabalhos relacionados, este projeto se diferencia principalmente na abordagem dos dados, por especificar os dados de análise para somente cursos da área de Computação e atuar no contexto do Brasil. Tal especificação trouxe desempenhos diferentes neste trabalho, quando comparado com os trabalhos que utilizaram dados de diversos cursos para predizer a evasão.

Vitelli (2010) teve como objetivo realizar pesquisas qualitativas e elaborar um plano de gestão para o combate à evasão. O trabalho se assemelha com este trabalho por utilizar Regressão Logística com um conjunto de variáveis similar no conjunto de dados. No entanto, o trabalho não se estende em questão de modelos preditivos. Por sua vez, Platt, Fan-Osuala e Herfel (2019) tiveram como foco em entender se o auxílio financeiro influencia na probabilidade do aluno completar a graduação e no tempo gasto na graduação. O trabalho também se restringe à cursos de TI e utiliza modelos de *machine learning*, porém tem como objetivo identificar se somente com as varáveis de tipo de estudo e variáveis de auxilio financeiro são capazes de predizer o tempo gasto do aluno na graduação. Já o trabalho de Delen (2010) é muito similar à este em termos de metodologia de desenvolvimento e métodos utilizados, compartilhando diversos processos e modelos preditivos. No entanto, ele se difere por focar somente em alunos do 1º ano acadêmico e generalizar ao utilizar dados de todos os cursos presentes na universidade. Por fim, Ameri et al. (2016) se difere por ir mais a fundo e objetivar identificar o semestre no qual o aluno irá evadir. Dessa forma, o estudo trilhou outro caminho e elaborou modelos de análise de sobrevivência que utilizam regressão. Estes modelos não são abordados neste trabalho por se tratarem de naturezas diferentes, dado que neste projeto tratou-se de um problema de classificação e Ameri et al. (2016) tratam de um problema de regressão.

3 Definição e Caracterização do Problema

Este capítulo aborda a evasão nos cursos de EC e SI do ICEA e detalha os dados utilizados para a elaboração deste trabalho. A Seção 3.1 descreve os conjuntos de dados utilizados neste trabalho. Em seguida, é apresentada a situação da evasão no ICEA na Seção 3.2. Por fim, na Seção 3.3 é realizada a caracterização dos alunos dos cursos de EC e SI.

3.1 Descrição do *Dataset*

A obtenção de dados individuais de cada aluno ingressante no curso de Engenharia de Computação ou Sistemas de Informação foi essencial para todo o desenvolvimento do trabalho. Os dados foram disponibilizados pela seção de ensino da UFOP, com os nomes dos discentes anonimizados para preservar a privacidade de cada um.

As análises referentes aos discentes da UFOP, apresentadas no Capítulo 2, onde caracterizamos os ingressantes e analisamos as evasões dos cursos, foram possíveis devido à utilização de um conjunto de dados referente aos ingressos e evasões nos cursos de EC e SI. O *dataset* conta com mais de 40 características individuais de cada aluno, as quais são descritas no dicionário de dados apresentado na Tabela 2.

Além desse *dataset*, também foi disponibilizado um conjunto de dados referente ao histórico escolar de cada discente dos cursos de Engenharia de Computação e de Sistemas de Informação. Nesse conjunto de dados pode-se consultar o desempenho de cada aluno em cada disciplina de forma isolada, o que será extremamente valioso para análise preditiva desenvolvida no Capítulo 4. Assim, as características presentes no conjunto de dados são: Matrícula criptografada do discente; Ano dos estudos; Semestre; Código do Departamento; Código da Disciplina; Nome da Disciplina; Código da Turma; Código do Curso; Caráter da Disciplina; Média Final; Exame Especial; Faltas; Situação do Discente; Quantidade de Aulas Dadas; Professor; e Data do Registro. O dataset, juntamente com suas características, é descrito no dicionário de dados apresentado na Tabela 3.

Dessa forma, os dados apresentados foram essenciais para atingir os objetivos do trabalho. A partir da análise dos conjuntos de dados foi possível desenvolver não apenas as análises estatísticas já apresentadas nesse capítulo, como também o desenvolvimento dos modelos de predição apresentados no Capítulo 4.

Tabela 2 – Dicionário de dados para o *dataset* de ingressantes.

Campo	**Definição**	**Tipo de dado**	**Exemplo de Valor**	**Notas adicionais**
Matrícula	Matrícula do discente	Texto (9)	19.1.8010	Formato “00.0.0000”
Curso	Nome do curso de graduação	Texto	Engenharia de Computação	Nome sem caracteres especiais
Sexo	Gênero do discente	Texto (1)	F	Valores: F - Feminino M - Masculino
Cidade Nascimento	Cidade natal do discente	Texto	João Monlevade	Cidade sem caracteres especiais
Estado de Nascimento	Estado natal nasceu	Texto	Minas Gerais	Estado sem caracteres especiais
País Nascimento	País natal do discente	Texto	Brasil	País sem caracteres especiais
Endereço Familiar	Endereço da família do discente	Texto	Rua Niterói - 99	Nome da rua e número da residência
Bairro Familiar	Bairro da família do discente	Texto	Baú	Bairro sem caracteres especiais
CEP Familiar	CEP da família do discente	Texto (10)	35.930-326	Formato “00.000-000”
Cidade Familiar	Cidade da família do discente	Texto	João Monlevade	Cidade sem caracteres especiais
Estado Familiar	Estado da família do discente	Texto	Minas Gerais	Estado sem caracteres especiais
País Familiar	País da família do discente	Texto	Brasil	País sem caracteres especiais
Endereço Acadêmico	Endereço residencial do discente	Texto	Avenida Armando Fajardo, 4100	Nome da rua e número da residência
Bairro Acadêmico	Bairro residencial do discente	Texto	Loanda	Bairro sem caracteres especiais
CEP Acadêmico	CEP residencial do discente	Texto (10)	35.931-073	Formato “00.000-000”
Cidade Acadêmico	Cidade residencial do discente	Texto	João Monlevade	Cidade sem caracteres especiais
Estado Acadêmico	Estado residencial do discente	Texto	Minas Gerais	Estado sem caracteres especiais

Continuação da tabela 2

Campo	**Definição**	**Tipo de dado**	**Exemplo de Valor**	**Notas adicionais**
País Acadêmico	País residencial do discente	Texto	Brasil	País sem caracteres especiais
República	Nome da república onde o estudante reside	Texto	Monastério	Campo vazio se o discente não residir em república
Tipo República	Tipo da República (Particular ou Federal)	Texto	Particular	Campo vazio se o discente não residir em república
Data Nascimento	Data de nascimento do discente	Data	31/12/1995	Formato DD/MM/AAAA
Ano de Admissão	Ano de ingresso do discente	Texto	2019	
Semestre de Admissão	Semestre de ingresso do discente	Texto	1	Valores: 1 ou 2
Cód. Modo Admissão	Código do modo de admissão	Texto (1)	V	
Descrição Modo Admissão	Descrição do modo de admissão	Texto	Vestibular	
Classificação Vestibular	Classificação do discente no vestibular	Inteiro	10	Campo vazio se o discente não ingressou por vestibular
Pontuação Vestibular	Pontuação do discente no vestibular	Float	720.10	Pontuação de 0 - 1000 para ENEM e 0 - 120 para Vestibular da UFOP
Origem	Universidade de origem do discente, em caso de transferência	Texto	Faculdade Doctum	
Turno	Turno das aulas do discente	Texto (1)	V	Vespertino (V) ou Noturno (N) para os cursos de CJM e SJM
Data Ingresso	Data de ingresso do discente	Data	19/02/2015	Formato DD/MM/AAAA

Continuação da tabela 2

Campo	**Definição**	**Tipo de dado**	**Exemplo de Valor**	**Notas adicionais**
Participou Política Afirmativa	Participação de políticas afirmativas de cotas para o discente	Texto	Não	
Usou Política Afirmativa	Uso de políticas afirmativas de cotas para o discente	Texto	Não	
Modalidade Concorrência	Modalidade de ingresso optada pelo discente	Texto	AC	Diversos código de políticas afirmativas (AC - Ampla concorrência)
Modalidade Concorrência Homologada	Modalidade de ingresso aceita ou recusada	Texto	SIM	
Ano diplomação	Ano de diplomação do discente	Texto	2019	"null" se o discente não se formou até o momento
Semestre Diplomação	Semestre de diplomação do discente	Texto	1	"null" se o discente não se formou até o momento

Tabela 3 – Dicionário de dados do histórico escolar dos discentes.

Campo	**Definição**	**Tipo de dado**	**Exemplo de Valor**	**Notas adicionais**
Ano	Ano dos estudos	Texto (4)	2019	Valores de 2005 à 2019
Cód. Departamento	Código do departamento	Texto (5)	DECSI	Códigos utilizados na UFOP
Descrição	Nome da disciplina	Texto	Programação de Computadores I	O nome da disciplina pode ser alterado ao longo do tempo, devido à readequações
Cód. Turma	Código da turma	Inteiro	11	
Cód. Curso	Código do curso	Texto	CJM	Formato "AAA"
Caráter	Carater da displina para o discente	Texto (1)	O	Valores: O - Obrigatória E - Eletiva F - Facultativa
Média Final	Media final obtida pelo discente	Float	9.5	Valores de 0.0 à 10.0
Exame Especial	Nota obtida no exame especial, caso realizado	Float ou Null	6.0	Substitui a Média Final obtida pelo discente, se realizado
Falta	Quantidade de faltas às aulas pelo discente	Inteiro	2	Valor em horas/aula (Ex.: 2 H/A)
Situação	Situação do discente na disciplina	Texto	Aprovado	Valores: Aprovado; Cursando; Trancado; Reprovado por Nota; Reprovado por Nota e Falta; Cancelado
Aula Dada	Quantidade de horas aula semanal para a disciplina	Inteiro	4	Valor em horas/aula (Ex.: 4 H/A)
Professor	Nome dos professores responsáveis e tipo das aulas lecionadas	Texto	Alexandre Magno de Sousa(T+P)	Pode haver mais de um professor por disciplina. Valores entre parênteses: T - Aula Teórica P - Aula Prática
Gravação	Data de gravação do registro	Data	19/02/2020	Formato DD/MM/AAAA
Matrícula	Matrícula criptografada do discente	Texto	2ba1983edd604819-7c22823d3fc7f055	A matrícula foi criptografada por uma função hash para preservar a privacidade dos discentes

3.2 Contextualização do Problema: Evasão no ICEA

No ICEA, mais especificamente no DECSI, departamento o qual engloba os cursos de Engenharia de Computação e de Sistemas de Informação, a evasão é um tema que merece atenção. Analisamos dados relacionados aos discentes desde o início dos dois cursos de graduação na área de tecnologia, sendo o curso de Sistemas de Informação instaurado em 2005 e o de Engenharia de Computação em 2008. Assim, obtemos conjuntos de dados extraídos da Seção de Ensino da UFOP, os quais contém informações relacionadas aos ingressos, egressos e evasões desses cursos desde seus respectivos inícios até o final do ano de 2019, totalizando 15 anos de acompanhamento para Sistemas de Informação e 12 anos para Engenharia de Computação.

A partir dos conjuntos originais, foram identificados os dados relevantes e executou-se um pré-processamento nestes a fim de filtrá-los, remodelá-los de acordo com o objetivo de pesquisa e tratar peculiaridades que poderiam gerar disformidades nos relatórios e experimentos. Inicialmente, para as etapas presentes nas Seções 3.2 e 3.3, esse pré-processamento se resume em selecionar o escopo de estudo e executar a limpeza dos dados. Para as etapas presentes no Capítulo 4 foram incluídas outras etapas de pré-processamento, descritas na seção 4.1, dada a necessidade das técnicas utilizadas no capítulo citado. O detalhamento das definições adotadas para selecionar o escopo de estudo é descrito a seguir.

Foram definidos prazos mínimos para o acompanhamentos dos alunos, a fim de disponibilizar tempo hábil para o aluno concluir o curso. Para isso, foi considerado o tempo limite de conclusão de cada curso de acordo com a resolução CEPE Nº 7.322 da UFOP, que define o prazo máximo de integralização curricular como 1,5 vezes o tempo estabelecido na matriz curricular do curso. Sendo assim, obtém-se um prazo máximo de 15 semestres letivos, em média 7,5 anos, para Engenharia de Computação e de 12 semestres letivos, em média 6 anos, para Sistemas de Informação. Após esse prazo o aluno normalmente é jubilado do curso, gerando um caso de evasão, salvo exceções cabíveis de recurso. Logo, acompanhou-se os ingressantes de Engenharia de Computação por 7,5 anos e os de Sistemas de Informação por 6 anos. O que resulta nos ingressantes do início do curso até os ingressantes do segundo semestre de 2012 para Engenharia de Computação, e os ingressantes até o primeiro semestre de 2014 para Sistemas de Informação. Assim, atendemos o prazo mínimo de acompanhamento considerando nosso limite temporal no final de 2019.

Ademais, para que sejam contemplados todos os casos de jubilamento, foi adicionada uma exceção ao limite temporal para incluirmos os jubilamentos do primeiro semestre de 2020, pois um aluno que excede o prazo máximo de integralização do curso tem seu jubilamento protocolado no semestre letivo posterior ao prazo limite. Ou seja, os discentes que atingem o prazo máximo do curso no segundo semestre de 2019 têm seus jubilamentos

protocolados no primeiro semestre de 2020.

Outro ponto encontrado foi de alunos que tiveram evasões protocoladas mais de uma vez. Isso ocorre devido a casos que o aluno é submetido ao jubilamento segundo alguma resolução da UFOP, como coeficiente de rendimento insuficiente durante um prazo determinado ou limite de tempo para integralização do curso, o discente interpõe um recurso que, se aceito, o proporciona a reingressar-se na universidade. Porém, após reingressar no curso o discente pode evadir-se novamente, ocasionando em uma segunda evasão para a mesma matrícula. Por serem casos especiais e poderem ocasionar incongruência de resultados, como um número total de egressos somados à evasão maior que o número de ingressos, preferimos filtrar esses casos e contabilizar somente uma evasão por matrícula.

Além disso, para que haja um acompanhamento desde o início acadêmico dos discentes, somente estão sendo considerados alunos que ingressaram na UFOP via vestibular ou Exame Nacional do Ensino Médio (ENEM). Sendo assim, estão exclusos alunos que entraram nos cursos por meio de transferência acadêmica, seja ela interna ou externa. Logo, evita-se adversidades que alunos dessa modalidade podem encontrar, como readequação acadêmica e grades curriculares não padronizadas devido ao reaproveitamento de disciplinas.

Assim, ao final da implementação dessas definições, foram realizadas análises estatísticas que convergiram nos resultados apresentados na Tabela 4.

Tabela 4 – Evasão nos cursos de tecnologia do ICEA.

Curso	**Ingressos**	**Egressos (Diplomados)**	**Evadidos**	**Taxa de Conclusão do Curso**	**Taxa de Evasão do Curso**
Engenharia de Computação	256	48	195	18,75%	76,17%
Sistemas de Informação	559	202	348	36,14%	62,25%
Média				27,44%	69,21%

A partir da observação dos dados, é possível constatar a necessidade de políticas para a redução da evasão nos cursos de tecnologia do ICEA. Pode-se notar uma taxa de evasão de 62,25% em Sistemas de Informação e, surpreendentes, 76,17% no bacharelado em Engenharia de Computação. Em relação à taxa de conclusão do curso, somente 18,75% dos ingressantes no curso de Engenharia de Computação conseguem completar a graduação e obter o título de bacharel. Ou seja, a cada 40 alunos que ingressam semestralmente no curso de Engenharia de Computação na UFOP, menos de 8 chegam até o final do curso e se tornam bacharéis.

3.3 Caracterização do Desempenho Acadêmico dos Discentes

Afim de entender a natureza dos ingressantes dos cursos de EC e SI da UFOP realizou-se análises estatísticas que auxiliaram no processo de caracterização do conjunto de dados. Foram analisadas as características individuais de cada ingressante, como sua origem, gênero, curso de ingresso e, principalmente, as notas obtidas nos vestibulares de avaliação para ingresso na UFOP. As formas de ingresso via vestibular foram o vestibular tradicional da UFOP, utilizado até o final de 2010, e o ENEM, o qual é utilizado desde 2011 até o momento. As análises são apresentadas a seguir.

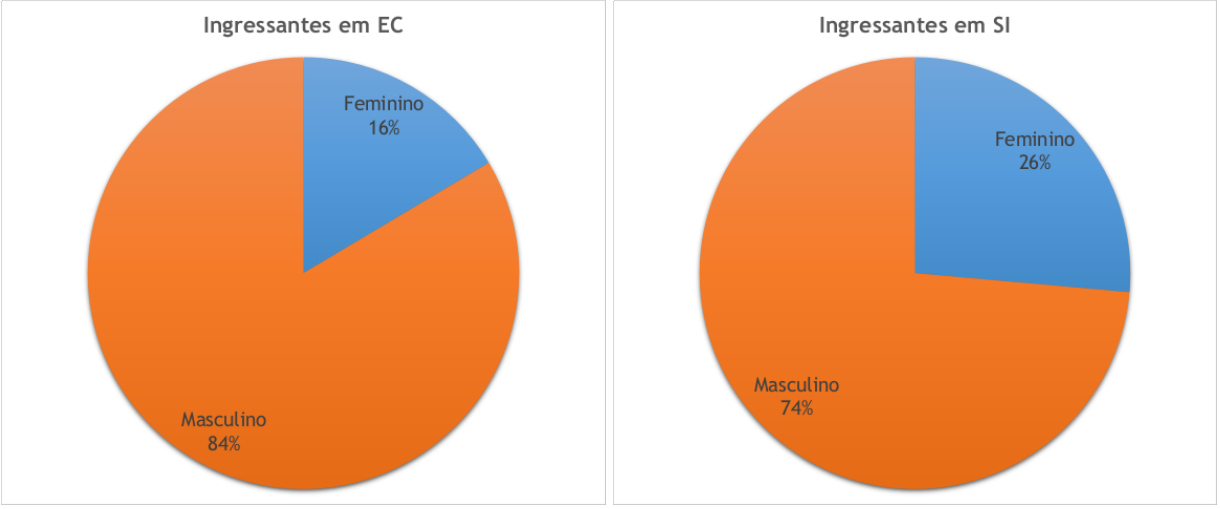
Na Figura 3, pode-se observar que a grande maioria do ingressantes optam pelo vestibular como forma de ingresso. Uma minoria opta por transferência externa e outros processos mais específicos, como o processo para portadores de diploma de graduação.

Figura 3 – Formas de Ingresso nos cursos de EC e SI da UFOP.

Quanto ao gênero dos ingressantes, as Figuras 4a e 4b indicam que a grande maioria dos ingressantes nesses cursos de tecnologia são do gênero masculino. Sendo que no curso de Engenharia de Computação o grupo é mais homogêneo ainda, sendo mais de 3/4 dos ingressantes do gênero masculino.

Dado o tamanho continental de nosso país, e sendo a UFOP a universidade de escolha para muitos estudantes, pode-se observar as diferentes origens dos estudantes na Figura 5. Há uma grande diversidade de culturas na universidade, havendo ingressantes de mais de 15 estados brasileiros nos cursos de EC e SI, desde suas devidas instaurações. Acresce que não somente brasileiros ingressam nesses cursos, estudantes de países de língua portuguesa, como Moçambique e Guiné-Bissau, também ingressaram nesses cursos, além de outras nacionalidades.

Em relação às pontuações obtidas no vestibular pelos ingressantes, as Figuras 6 e 7 apresentam, respectivamente, o histograma das pontuações obtidas pelos ingressantes nos cursos de EC e SI no vestibular tradicional da UFOP e no ENEM. Essas pontuações



(a) Ingressantes em EC por gênero. (b) Ingressantes em SI por gênero.

Figura 4 – Gênero dos ingressantes em EC e SI.

Figura 5 – Ingressantes em EC e SI por estado.

foram separadas devida à natureza de cada prova, com estruturas e escalas de pontuações diferentes, como pode-se observar nos histogramas. Assim, é possível deduzir a partir dos gráficos que a grande maioria dos ingressos tem pontuação de aproximadamente 60 pontos no vestibular e 640 pontos no ENEM.

Dada a diversidade quanto à origem dos ingressantes, é possível agrupa-los por estado e avaliar o desempenho de cada estado separadamente. As Figuras 8 e 9 fazem o uso de *boxplots* para agrupar as pontuações por estado, detalhando as características individuais de cada um.

Pode-se observar que para o vestibular da UFOP, na Figura 8, os estados de Minas Gerais e Espírito Santo apresentam maior dispersão nos dados e maior simetria em relação à mediana. Por outro lado, os estados de São Paulo e Rio de Janeiro apresentam menor

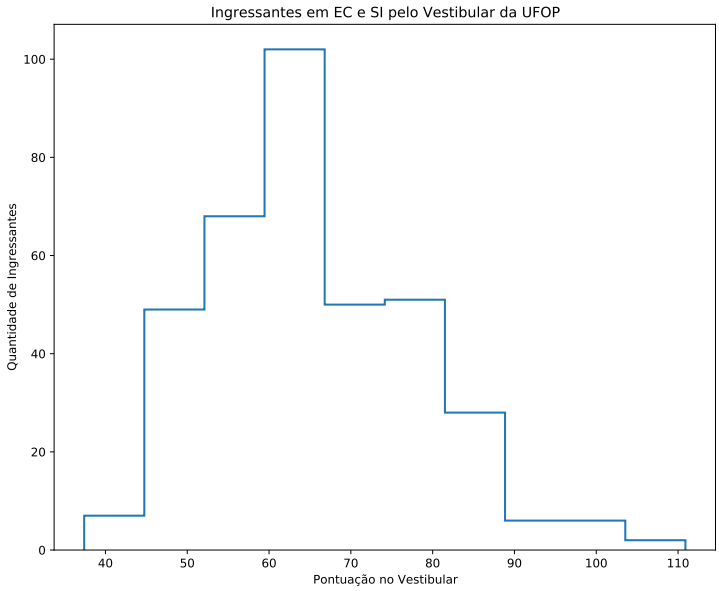


Figura 6 – Histograma da pontuação dos ingressantes pelo vestibular tradicional da UFOP para EC e SI.

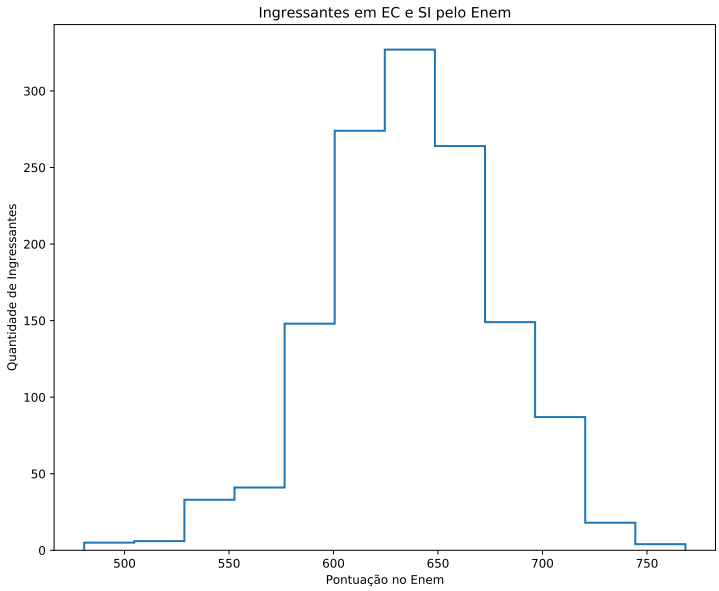


Figura 7 – Histograma da pontuação dos ingressantes pelo ENEM para EC e SI.

dispersão nos dados, porém com alta assimetria em relação à mediana.

Já para a pontuação no ENEM, na Figura 9, ao observar o intervalo interquartil de cada *boxplot* pode-se perceber que o Distrito Federal e o estado do Rio de Janeiro tem dados concentradas em um intervalo com pontuação superior aos demais estados, e com uma certa simetria em relação a mediana. Quanto a dispersão das notas, Minas gerais se destaca pela elevada dispersão nos dados e um alto número de *outliers*. É importante ressaltar, que para essas comparações deve ser levado em conta a predominância na quantidade de ingressantes de Minas Gerais, sendo estado de origem de 92% dos ingressantes, como apresentado na Figura 5.

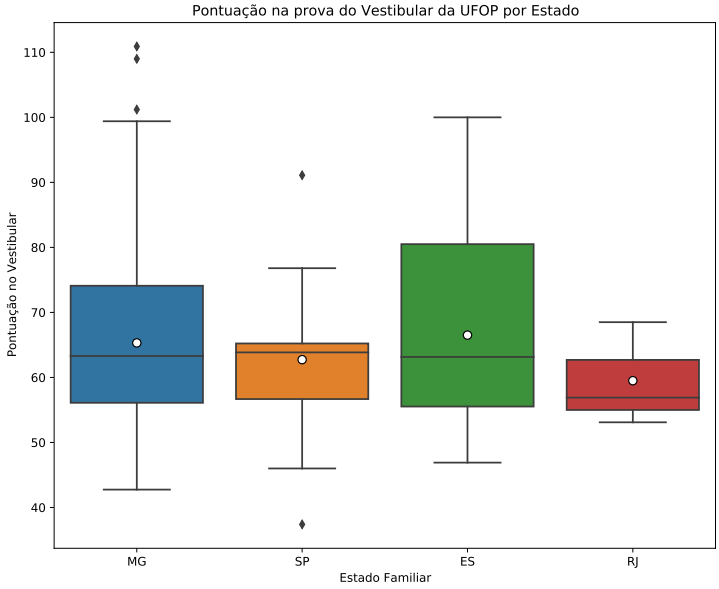


Figura 8 – *Boxplot* da pontuação dos ingressantes pelo vestibular tradicional da UFOP para EC e SI.

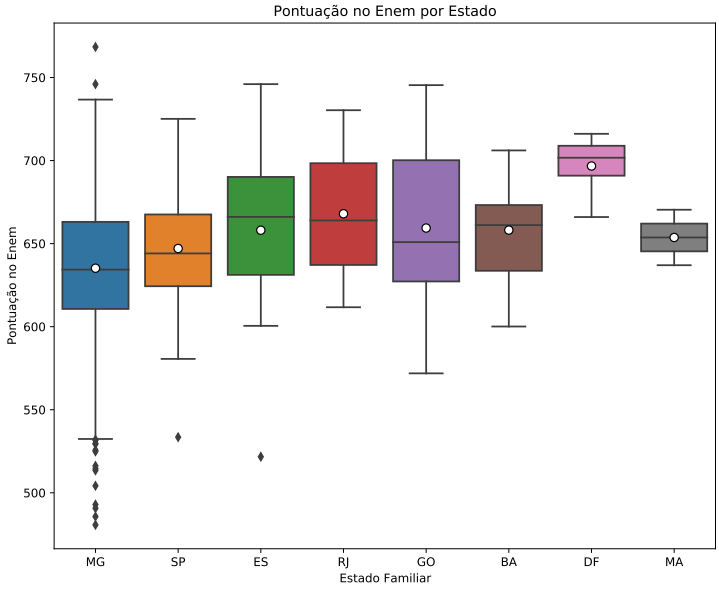


Figura 9 – *Boxplot* da pontuação dos ingressantes pelo ENEM para EC e SI.

Em relação aos números absolutos de ingressos nos cursos de EC e SI, as Tabelas 5 e 6 apresentam a quantidade de ingressantes pelo vestibular da UFOP e pelo ENEM, respectivamente. Para a vestibular da UFOP deve-se ressaltar, novamente, que o curso de EC é mais recente do que o de SI, o que explica a diferença na quantidade de dados. Já para os ingressantes pelo ENEM pode-se observar um equilíbrio na quantidade de ingressos em ambos os cursos, o qual não é apresentado quando se compara a quantidade de ingressantes do gênero masculino e feminino. Fato que já era previsto, levando em consideração os gráficos já apresentados nessa seção.

Com o auxílio da *Cumulative Distribution Function* (CDF), ilustrada nas Figuras

Tabela 5 – Ingressantes pelo Vestibular tradicional da UFOP (2005.1 até 2010.2).

Curso	Ingressos	Masculino	Feminino
Engenharia de Computação	95	81	14
Sistemas de Informação	274	185	89
Total	369	266	103

Tabela 6 – Ingressantes pelo ENEM (2011.1 até 2019.2).

Curso	Ingressos	Masculino	Feminino
Engenharia de Computação	676	564	112
Sistemas de Informação	680	518	162
Total	1356	1082	274

10 e 11, pode-se analisar e comparar de forma separada os cursos e gêneros dos discentes. A partir da interpretação das curvas, é possível perceber que a grande maioria dos ingressantes obtiveram pontuações entre 50 e 90 pontos para o vestibular e entre 580 e 700 pontos para o ENEM. Ademais, pode-se constatar que mais de 50% das notas foram inferiores a 64 pontos para o vestibular tradicional da UFOP e 650 pontos para o ENEM.

Acresce que, ao comparar-se as pontuações entre cursos e entre gêneros é possível obter análises relevantes. Para o ENEM, na Figura 11, a pontuação para os ingressantes do curso de Engenharia de Computação tende a ser 10 pontos maior do que para Sistemas de Informação, visto que mais de 50% das notas de EC são superiores a 640 pontos, enquanto SI apresenta 630 pontos no mesmo indicador. Analogamente, para os ingressantes do gênero masculino pode-se observar uma pontuação que tende a ser aproximadamente 10 pontos mais elevada do que a dos ingressantes do gênero feminino.

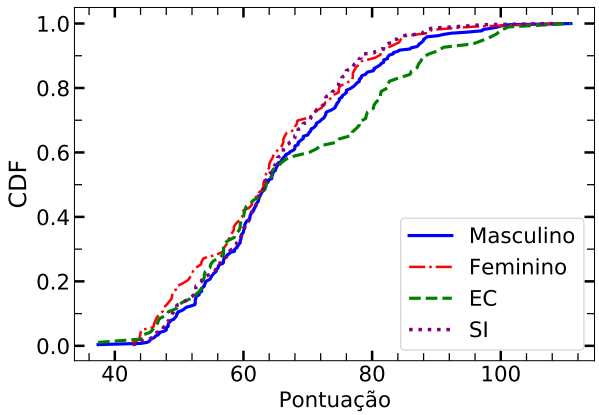


Figura 10 – CDF da pontuação dos ingressantes pelo vestibular tradicional da UFOP.

Além disso, essas análises são suportadas pelo teste de Kolmogorov–Smirnov (*KS Test*), o qual constatou que as distribuições das pontuações do ENEM para o curso de EC e de SI são significativamente diferentes (*KS Value*: 0,135; *P-value*: $6,710 * 10^{-6}$; *Alpha*: 0,001; com confiança de 99%). Similarmente, para a análise entre os gêneros na pontuação do ENEM também constatou-se que as distribuições são significativamente diferentes pelo *KS Test* (*KS Value*: 0,121; *P-value*: 0,003; *Alpha*: 0,001; com confiança de 99%).

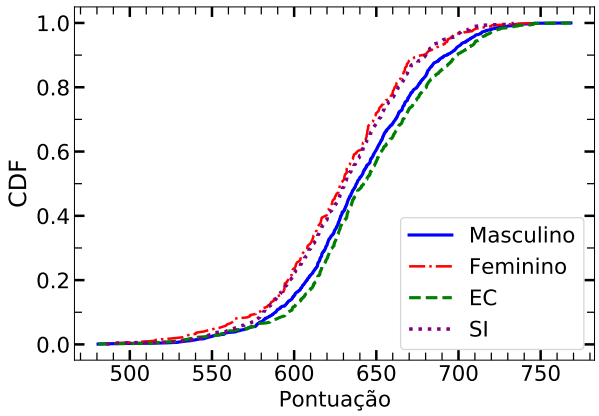


Figura 11 – CDF da pontuação dos ingressantes pelo ENEM.

Do mesmo modo, as Figuras 12 e 13 utilizam *violin plots* para explorar as notas dos ingressantes dos cursos de EC e SI de forma separada. Pode-se perceber que para as notas obtidas pelo vestibular tradicional da UFOP, na Figura 12, a pontuação dos ingressantes em Engenharia de Computação se concentram em 2 pontos, próximos aos 60 e 80 pontos, respectivamente. Por outro lado, os ingressantes em Sistemas de Informação têm notas mais uniformes em relação a mediana, com concentração próxima aos 60 pontos.

Em relação às notas do ENEM, o curso de Sistemas de informação mantém o mesmo padrão, têm pontuações uniformes em relação à mediana. Já para o curso de Engenharia de Computação o comportamento dos dados se altera em relação ao observado no vestibular tradicional da UFOP. Na Figura 13 é possível perceber que a pontuação no curso de EC se torna mais concentrada, sendo mais acentuada na região abaixo da mediana.

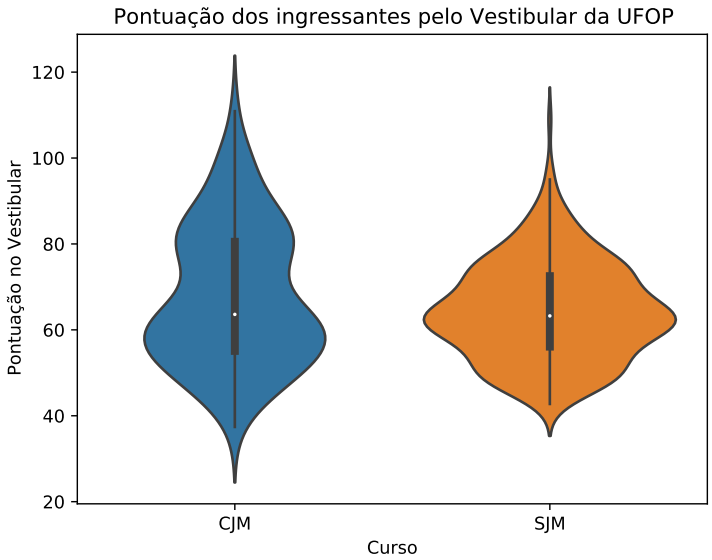


Figura 12 – *Violin plot* da pontuação dos ingressantes pelo vestibular tradicional da UFOP.

Ademais, a partir da leitura de trabalhos relacionados ao tema da monografia, os

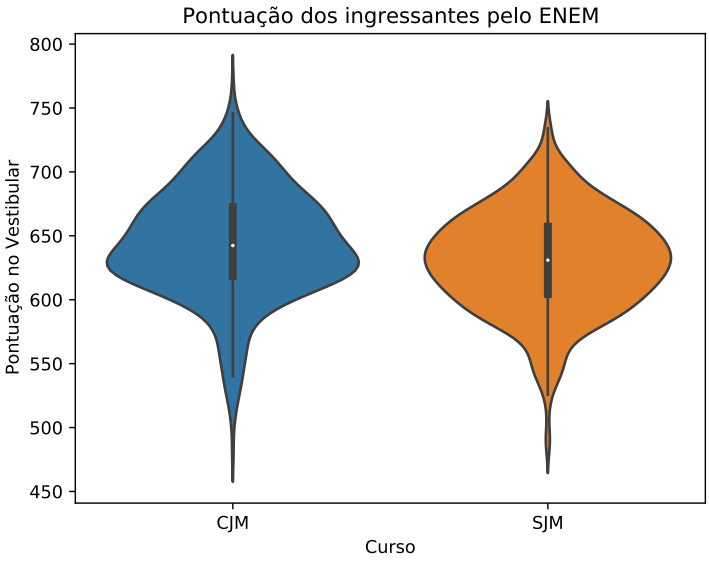


Figura 13 – *Violin plot* da pontuação dos ingressantes pelo ENEM.

quais são detalhados na Seção 2.3, pôde-se observar a importância de outras variáveis para a retenção do estudante no curso de graduação. Com base nos conjuntos de dados obtidos na seção de ensino da UFOP, descritos na Tabela 2 e Tabela 3, foi possível manipular os dados e derivar novas variáveis para auxiliar no entendimento dos dados e para a predição da evasão em si. Variáveis como Tempo de Curso, Media de Disciplinas por Semestre, Número de Reprovações, Número de Aprovações, Nota Média nas Disciplinas e a razão Aprovado/Matriculado foram criadas e examinadas. O detalhamento de cada variável é descrito a seguir.

Em relação ao tempo de curso pode-se observar na Figura 14 que mais de 50% dos discentes de EC permaneceram no curso por até 7 semestres, resultando em sua grande maioria na evasão do aluno no curso já que são raros os casos de discentes que concluíram o curso antecipadamente — porém possíveis visto que o discente pode ter realizado algum curso de graduação similar anteriormente e requerido o aproveitamento de disciplinas, reduzindo a quantidade de disciplinas obrigatórias que o aluno deve cursar para concluir o curso de graduação. Já no curso de SI nota-se que mais de 50% do discentes permanecem por até 9 semestres no curso, em contraste com os 7 semestres para EC.

Esse dado alarmante para o curso de EC era presumível, dado que foi revelado na Tabela 4 que o curso de EC possui uma taxa de evasão de 76%. Ademais, como definido na Seção 3.2, é importante ressaltar que os dados de estudo são de discentes que dispuseram de ao menos 1,5 vezes o tempo de curso padrão para a conclusão do mesmo, ou seja, no mínimo 15 semestres para EC e 12 semestres para SI.

Ao analisar a quantidade média de disciplinas em que os discentes se matriculam por semestre, ilustrada na Figura 15, nota-se que os alunos do curso de EC realizam mais disciplinas em simultâneo que os alunos do curso de SI. Mais da metade dos discentes

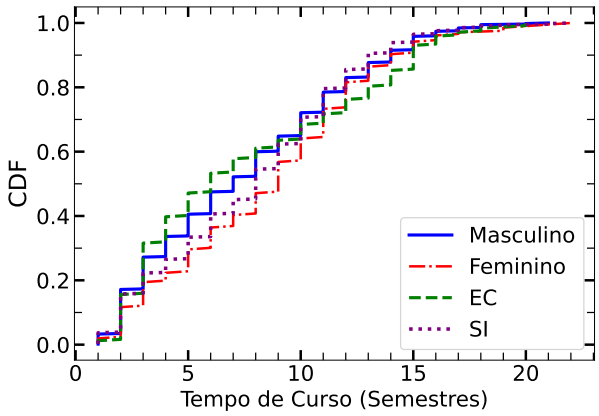


Figura 14 – Tempo de curso dos discentes.

de EC realizam 5 ou mais disciplinas por semestre em média, comparado à cerca de 4,5 disciplinas em média para os discentes de SI.

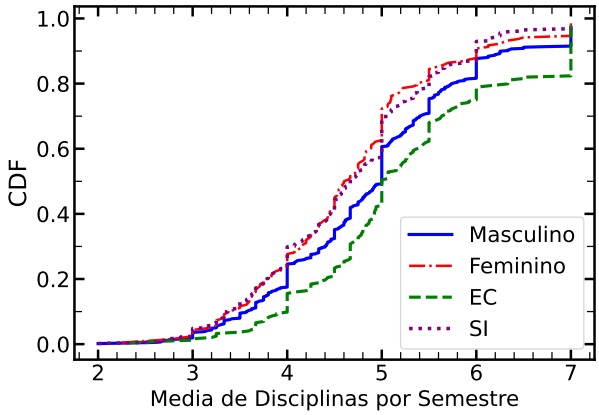


Figura 15 – Media de disciplinas matriculadas por semestre.

Na Figura 16 são detalhadas duas variáveis que, de acordo com os trabalhos relacionados presentes na Seção 2.3, tem grande influência na probabilidade de um aluno evadir ou não do curso de graduação. A Figura 16a apresenta a nota média (média aritmética) dos discentes nas disciplinas, variável a qual é similar ao Coeficiente de Rendimento Escolar (CRE) ou Coeficiente de Rendimento Acadêmico (CRA) ou *Grade Point Average* (GPA), porém se diferencia já que na média indicada na figura não é considerada a carga horária de cada disciplina para o cálculo, como geralmente é realizado nas outras métricas citadas.

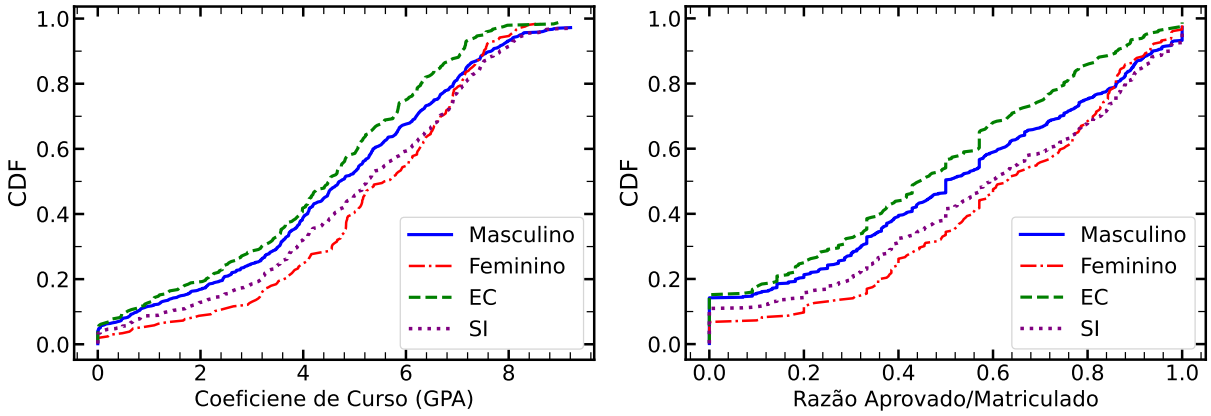
Nota-se que o cenário entre os cursos e gêneros se repete na figura, a qual indica que o curso de SI e o gênero feminino possuem desempenho superior ao curso de EC e gênero masculino. Observando a primeira metade da CDF ilustrada na Figura 16a, cerca de 55% dos discentes de EC tem uma nota média nas disciplinas inferior à 4,8 pontos (em uma escala de 0 à 10) comparados à 5,6 pontos para o curso de SI. Já em relação aos gêneros cerca de 55% dos alunos do gênero masculino tem uma nota média inferior à 5

pontos, já no grupo feminino constata-se 6 pontos para o mesmo cenário.

De maneira similar, a Figura 16b ilustra a razão entre quantidade de disciplinas em que o aluno foi aprovado com a quantidade de disciplinas no qual o mesmo se matriculou. Essa métrica obtida com base na leitura de Delen (2010), o qual refere a variável no texto como *Earned/Registered*. A variável teve grande influência no trabalho citado e nos ajuda a entender o comportamento do discente. A Equação 3.1 descreve o cálculo para obtenção da variável.

$$Aprovado/Matriculado = \frac{Quantidade\ de\ disciplinas\ aprovadas}{Quantidade\ de\ disciplinas\ matriculadas} \tag{3.1}$$

Assim, ao analisar os dados na figura Figura 16b percebe-se, novamente, um desempenho superior para o curso de SI e gênero feminino ao compará-los ao curso de SI e gênero masculino.

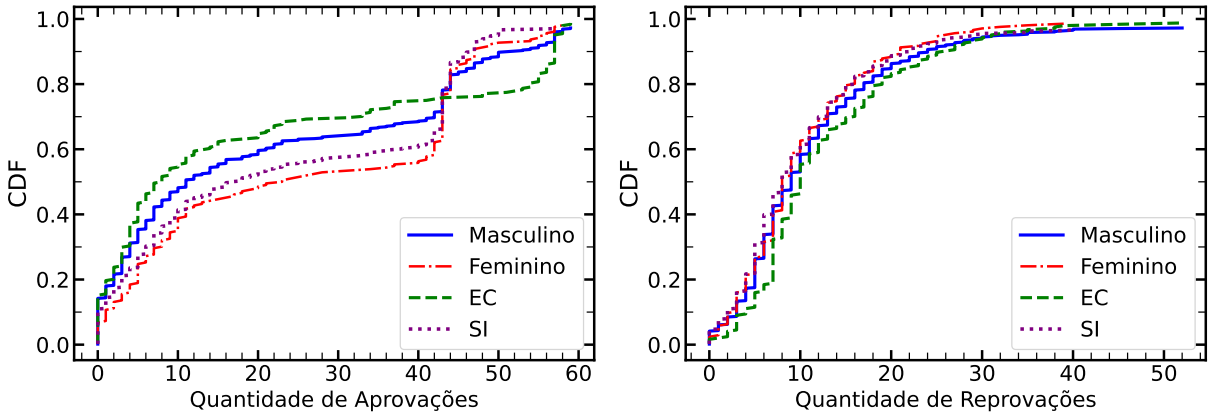


(a) Nota média nas disciplinas por aluno. (b) Razão Aprovado/Matriculado por aluno.

Figura 16 – Nota média nas disciplinas e razão Aprovado/Matriculado.

No que condiz a quantidade de disciplinas em que o discente foi aprovado ou reprovado, observa-se na Figura 17a que os discentes de SI tendem a ter mais aprovações do que os discentes de EC, visto que mais de 50% dos alunos de EC tem menos que 8 aprovações comparados às 24 aprovações para SI. O gráfico muda em torno de 43 aprovações para os cursos, mas isso é devido à extensão dos cursos serem diferentes e portando a quantidade de mínima de disciplinas aprovadas serem diferentes para a conclusão do curso. A situação negativa para o curso de EC é novamente observada na Figura 17b, a qual revela que os discentes de EC tendem a ter um número de reprovações ligeiramente maior do que os discentes de SI.

Em relação às evasões nos cursos de SI e EC, outro dado que é de extrema importância é o semestre em que o aluno evadiu. Na Tabela 7 são sumarizadas as evasões de acordo com o semestre na qual estas ocorreram. É possível observar que a evasão dos



(a) Disciplinas aprovadas por aluno.

(b) Disciplinas reprovadas por aluno.

Figura 17 – Quantidade de Aprovações e Reprovações nos cursos de EC e SI.

alunos ocorre em sua maioria nos semestres iniciais dos cursos, principalmente no 2º e 3º semestre. Nota-se, também, que mais de 50% das evasões ocorrem até o 5º semestre de curso, dado o qual é importante para políticas universitárias à respeito da evasão e também valioso para construção de modelos preditivos. Acresce que este comportamento é também verificado em outros estudos sobre a evasão universitária, apresentados na Seção 2.3, nos quais identificam-se maior frequência de evasões nos semestres iniciais dos cursos de graduação.

Tabela 7 – Semestre no qual os alunos evadiram.

Semestre da evasão	**% de alunos evadidos**	**% acumulada**
1º	4.41%	4.41%
2º	19.19%	23.61%
3º	14.20%	37.81%
4º	8.25%	46.07%
5º	10.36%	56.43%
6º	9.98%	66.41%
7º	6.14%	72.55%
8º	5.95%	78.50%
9º	4.22%	82.73%
10º	4.03%	86.76%
11º	3.45%	90.21%
12º	2.69%	92.90%
13º	1.73%	94.63%
14º	2.30%	96.93%
15º	1.34%	98.27%
16º ou superior	1.73%	100.00%

Dessa forma, com o auxílio dos dados apresentados nessa seção, é possível entender melhor o conjunto de dados de estudo. Assim, possibilita-se a caracterização desses dados, a qual será de extrema valia para as análises preditivas que serão desenvolvidas no Capítulo 4.

4 Metodologia e Desenvolvimento

Este capítulo descreve a metodologia utilizada para análise e caracterização e dos dados deste trabalho, seguida da análise preditiva da evasão utilizando técnicas de *machine learning*. A primeira seção 4.1 descreve as etapas realizadas no tratamento dos dados. A seção 4.2 apresenta as análises de correlação realizadas a fim de buscar correlações nos dados e compreender a variação desses. Na seção 4.3 é descrito o teste de Regressão Linear Múltipla realizado nos dados. Por fim, na seção 4.4 é apresentada a análise preditiva. São detalhados os modelos de classificação utilizados, as configurações de treino e teste e os resultados das predições.

Para o desenvolvimento deste trabalho, em termos de tecnologia, foi utilizada a linguagem de programação Python na versão 3.10 com o suporte da IDE PyCharm da JetBrains. Os principais pacotes Python utilizados foram: NumPy, Pandas, Scikit-learn, XGBoost, Matplotlib e Seaborn. A máquina utilizada para a execução dos modelos possui a seguinte configuração: processador Intel Core i5-5200U; 16GB de memória RAM; 256GB de armazenamento SSD; Sistema Operacional Linux Manjaro 21.3.2 Ruah; Linux Kernel versão 5.10.126-1-MANJARO 64 bits. .

4.1 Pré-processamento de dados

O pré-processamento de dados é uma parte essencial e de extrema importância para trabalhos que envolvem Mineração de Dados, Aprendizado de Máquina e/ou *Data Science*. Em relação aos fatores que podem impactar no sucesso de algoritmos de *Machine Learning* Kotsiantis, Kanellopoulos e Pintelas (2006) define que a representação e qualidade dos dados utilizados são os fatores de maior impacto no processo. Han, Pei e Kamber (2011) e García, Luengo e Herrera (2015) também citam que essa etapa é conhecida por ser uma das mais significantes dentro do famoso processo de Descoberta de Conhecimento a partir de Dados (*Knowledge Discovery from Data - KDD*).

Diversas etapas são executadas no pré-processamento de dados, podendo incluir a limpeza dos dados, normalização, transformação, extração de características e seleção, etc. De forma que, o produto final do pré-processamento de dados é o conjunto de dados final para treinamento (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

Dada a importância dessa etapa, foi realizada um tratamento extensivo dos dados ao longo do trabalho com o auxílio de diversas técnicas até que se obtivesse o conjunto final de dados. Conjunto o qual foi utilizado no treinamento dos modelos preditivos detalhados na seção 4.4.

As etapas realizadas no pré-processamento de dados são definidas, em ordem de execução, a seguir:

1. Limpeza dos dados e seleção do escopo de estudo;
2. Transformação e extração de características;
3. Padronização dos dados;
4. Tratamento de dados ausentes.

A primeira etapa do pré-processamento de dados foi a limpeza e extração dos dados de interesse. A partir dos dados originais obtidos na seção de ensino da UFOP, foram adotadas as definições detalhadas na seção 3.2. Dentre outros procedimentos detalhados na seção citada, essa etapa consistiu, principalmente, em selecionar somente os ingressantes via vestibular e selecionar os alunos que ingressaram em SI desde o semestre 2005.1 até 2014.1, e de 2008.1 até 2012.2 para o curso de EC.

Em seguida, dado que possuíamos dois *datasets* separados, apresentados nas tabelas 2 e 3, foi realizada combinação dos dados a fim de resultar-se em um único *dataset*. Para isso, foi realizada a transformação dos dados e extração de características relevantes ao estudo, além da criação de novas características, as quais foram possíveis a partir da manipulação dos dados. Realizados os procedimentos, prosseguiu-se com a união dos dados em um único *dataset*, tendo como referência a matrícula do discente.

Ao final da segunda etapa, produziu-se um único *dataset* com 17 características dos discentes, além do número de matrícula como índice. O conjunto de dados é apresentado na Tabela 8.

Tabela 8 – Conjunto de dados após pré-processamento.

Variável	**Tipo de dado**
Código do Curso	String
Gênero	String
Ano de Admissão	Int
Classificação no Vestibular	Int
Turno	String
Idade	Int
É de João Monlevade	Boolean
Tempo de Curso	Int
Média	Float
Aprovações	Int
Reprovações	Int
Aprovado/Matriculado	Float
Média de disciplinas por semestre	Float
Pontuação Vestibular	Float
Cidade Familiar	String
Estado Familiar	String
Modalidade de Concorrência	String

Nesse ponto, os dados foram utilizados para realizar a caracterização dos alunos, detalhada na seção 3.3, visto que os dados ainda mantinham seus valores originais, necessários para a análise citada.

Posteriormente, sucedeu-se com a padronização dos dados. A padronização dos dados é fundamental para que os algoritmos de classificação não fiquem enviesados com as variáveis de maior grandeza do *dataset*. No entanto, são necessários valores numéricos para que seja realizada a padronização das amostras, o que não era apresentado em todas as variáveis presentes no *dataset*. Sendo assim, as variáveis categóricas foram convertidas para valores ordinais. Para variáveis como Código do Curso, Gênero que somente assumiam dois valores foram atribuídos valores binários. E para variáveis que assumiam uma variedade de valores mais extensa como o caso de Cidade Familiar, Modalidade de concorrência foi realizada a transformação dessas para valores ordinais de acordo com a frequência de amostragem em ordem ascendente.

Ademais, a variável Pontuação Vestibular requiriu atenção extra na padronização dos dados, já que esta possuía uma variação de dados diferente de acordo com o período de ingresso do discente, dado que a pontuação no antigo vestibular da UFOP tem escala diferente das notas do ENEM, que foi adotado como forma de ingresso à partir do primeiro semestre de 2011. Sendo assim, foram separadas as pontuações advindas dos dois vestibulares para que seja realizada a padronização destas de forma separada, com base no ano de 2011.

Logo, levando em consideração os detalhes descritos anteriormente, foi utilizada a padronização dos dados de forma que ao calcular amostras cada característica presente no *dataset* resulte em média igual a 0 e desvio padrão igual à 1. Para isso foi utilizada o método *z-score*, descrito na Equação 4.1.

$$z = \frac{x - \mu}{\sigma}, \tag{4.1}$$

em que z é o valor padronizado, x é o valor da amostra, μ é a média aritmética e, por fim, σ é o desvio padrão. Em termos práticos, o valor de z é a medida de quantos desvios padrão um valor de amostra está acima ou abaixo da média aritmética.

Por fim, foi realizado o tratamentos de dados ausentes no *dataset*, como o caso da Média, Números de Aprovações e outros dados para alunos que ingressaram no curso e cancelaram sua matrícula ou tiveram-a cancelada antes de finalizar o primeiro semestre do curso, de forma à não gravar registros em relações às notas nas disciplinas já que as mesmas não foram finalizadas. Assim, como não é aceito pelos algoritmos de predição dados ausentes, nulos ou não definidos, foi realizado o tratamento desses dados. De forma a manter a integridade da variação dos dados, foi adotado o valor médio da respectiva variável para as amostras ausentes da mesma, que em nosso caso foi o preenchimento

desses dados com o valor 0, visto que os dados estavam padronizados pelo método *z-score* com média igual à 0 e desvio padrão igual à 1. O conjunto de dados final é apresentado na Tabela 9.

Tabela 9 – Conjunto de dados padronizado para treinamento

Variável	**Tipo de dado**
Código do Curso	Float
Gênero	Float
Ano de Admissão	Float
Classificação no Vestibular	Float
Turno	Float
Idade	Float
É de João Monlevade	Float
Tempo de Curso	Float
Média	Float
Aprovações	Float
Reprovações	Float
Aprovado/Matriculado	Float
Média de disciplinas por semestre	Float
Pontuação Vestibular	Float
Cidade Familiar	Float
Estado Familiar	Float
Modalidade de Concorrência	Float

4.2 Análise de Correlação

A análise de correlação foi realizada para todas as 17 variáveis presentes na Tabela 9. Dado a diversidade de características, a análise visou entender o relação entre as variáveis e a possibilidade de redução do *dataset* no caso de variáveis fortemente correlacionadas.

4.2.1 Matriz de Correlação

Para a elaboração das próximas análises foram utilizados dois coeficientes de correlação: *Pearson* e *Spearman*. O conjunto de dados apresentado na Tabela 9 e aplicados os métodos de correlação com o auxílio da biblioteca Pandas[1] para Python. Com as matrizes de correlação de *Pearson* e *Spearman* elaboradas utilizou-se da biblioteca Seaborn[2] para a produção de um mapa de calor que exibisse as matrizes de forma gráfica. As matrizes de correlação são apresentadas nas Figuras 18 e 19.

Ao analisar as matrizes de correlação *Pearson*, Figura 18, e de *Spearman*, Figura 19, percebe-se que ambas obtiveram resultados similares, variando levemente em grau de intensidade, em relação às variáveis que estão fortemente correlacionadas e são de interesse da análise. Utilizando como base a correlação de *Pearson*, pode-se notar algumas correlações pertinentes à uma análise mais detalhada, descrita a seguir.

[1] <https://pandas.pydata.org/>
[2] <https://seaborn.pydata.org/>

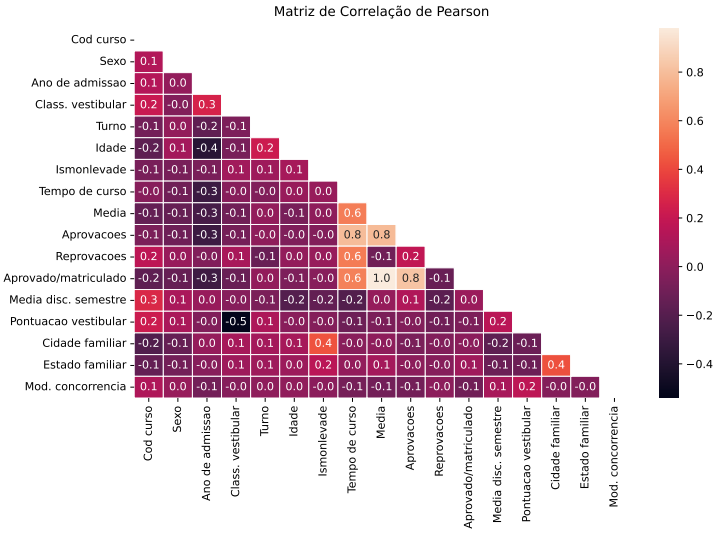


Figura 18 – Matriz de correlação de Pearson.

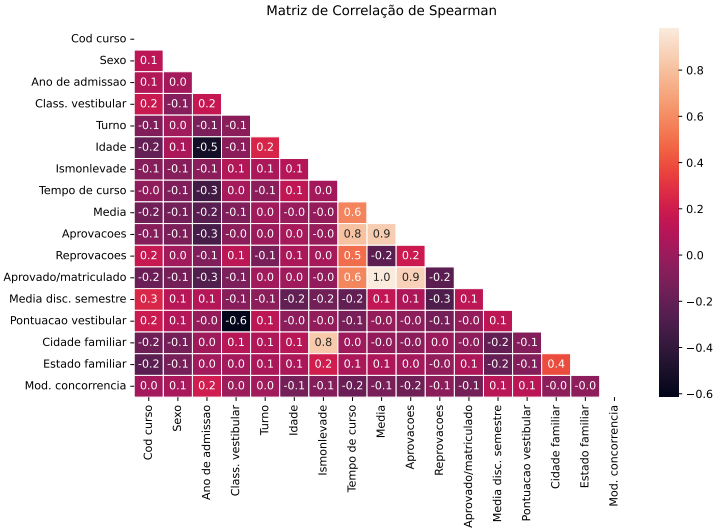


Figura 19 – Matriz de correlação de Spearman.

Observa-se que as variáveis Aprovações e Reprovações estão fortemente correlacionados de forma positiva com a variável Tempo de Curso. Apesar de ser uma correlação forte, esta correlação é evidente já que quanto maior for o tempo de curso do discente mais disciplinas este terá se matriculado e, portanto, ter sido aprovado ou reprovado nas mesmas. A mesma situação ocorre com a correlação entre Aprovações e Aprovado/Matriculado, onde nota-se uma correlação positiva forte entre as variáveis porém elas são dependentes e já era esperado esse comportamento.

Ainda em relação ao Tempo de Curso, nota-se que as variáveis Média e Aprovado/Matriculado estão fortemente correlacionadas de forma positiva ao Tempo de Curso. Essa correlação nos indica que os discentes tendem à melhorar o desempenho acadêmico ao decorrer do curso. São correlações fortes, porém com variáveis de naturezas diferentes e portanto necessárias no conjunto de dados.

Uma correlação que chama atenção em ambos as matrizes de correlação é entre as variáveis Média e Aprovado/Matriculado. Nota-se uma correlação linear perfeita para essas variáveis, indicando que elas estão fortemente correlacionadas. Essas métricas tem significados similares, já que ambas representam em sua essência o desempenho acadêmico do discente. Dado isso, deve-se analisar se é viável descartar uma das variáveis, visando obter melhor desempenho na etapa de treinamento dos modelos de predição que serão executados posteriormente. Assim, para avaliar as variáveis, será executado um teste de regressão linear múltipla de forma a identificar o impacto de cada no conjunto de dados. O teste é descrito na Seção 4.3.

Por fim, visualiza-se que as variáveis Pontuação Vestibular e Classificação Vestibular estão negativamente correlacionadas de forma moderada na Figura 18. Tal correlação tem fundamento visto que as variáveis são inversamente proporcionais. Quanto maior for a nota do discente no vestibular, menor (i.e. melhor) será sua classificação no vestibular, já que esta utiliza de números ordinais. Para essa correlação também é interessante avaliar se há a possibilidade de somente uma variável já descrever o comportamento do discente, o que necessitará ser avaliado em um teste de regressão linear múltipla.

Portanto, com o auxílio das matrizes de correlação de *Pearson* e *Spearman*, foi possível obter mais clareza no que concerne às relações entre as características presentes no conjunto de dados. 2 combinações de variáveis (Aprovado/Matriculado e Media; Pontuação Vestibular e Classificação Vestibular), resultando em 4 variáveis estão correlacionadas de modo a ser necessário avaliar a possibilidade de descartar variáveis análogas. Como mencionado anteriormente, para isso serão realizados testes de regressão linear múltipla para averiguar a possibilidade de uma variável do par poder explicar o comportamento do discente, não necessitando o uso de ambas as variáveis. Os testes de regressão linear são descritos na seção 4.3.

4.2.2 Análise de Componente Principal

Principal Component Analysis (PCA) ou Análise de Componente Principal, em português, é uma técnica multivariada que analisa uma tabela de dados na qual as observações são descritas por diversas variáveis quantitativas dependentes e inter-correlacionadas. Seu objetivo é extrair as informações importantes do conjunto de dados e representá-las como um conjunto de novas variáveis ortogonais chamadas componentes principais (ABDI; WILLIAMS, 2010).

O PCA é comumente utilizado na análise exploratória de dados como técnica de redução de dimensionalidade de dados, transformando um conjunto de dados com diversas variáveis em um conjunto menor de variáveis linearmente não correlacionadas, ou seja, os componentes principais. Dessa forma é possível agilizar o treinamento dos modelos de predição para altos volumes de dados. No entanto, o PCA não será utilizado com esse

intuito nesse trabalho, visto que não houve a necessidade de acelerar o treinamento de dados apresentado na subseção 4.4.1.

Adicionalmente, o PCA também é utilizada para analisar o impacto de cada variável na variação do conjunto de dados, como é realizado no Capítulo 6.6 de Jain (1991). Dessa forma, o PCA será utilizado neste trabalho como uma ferramenta para analisar como a variação dos dados é influenciada por cada variável do *dataset*. Assim, seguindo o procedimento definido em Jain (1991), foi realizada a análise quanto ao impacto de cada variável sobre a variação dos dados do *dataset*. Os resultados são apresentados na Figura 20.

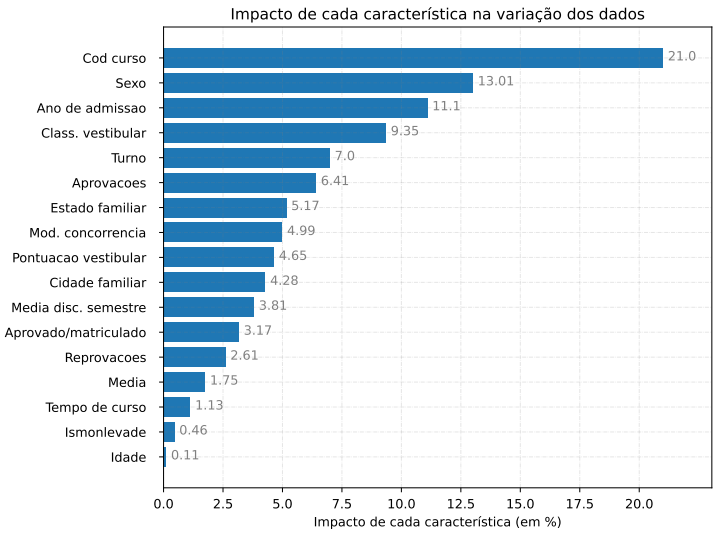


Figura 20 – Impacto de cada variável na variação dos dados a partir do PCA.

Ao analisar a Figura 20, percebe-se que o Código do Curso e Gênero do discente explicam grande parte da variação dos dados, o que poderá fazer com que essas variáveis tenham alto peso nos modelos preditivos. Esse impacto é suportado pelas análises realizadas na Seção 3.3, nas quais constatou-se comportamentos distintos para esses grupos em diversos cenários.

Outras características que chamam atenção na Figura 20 são o Ano de Admissão e a Classificação no Vestibular. O alto impacto do Ano de Admissão na variação dos dados nos indica que há uma inconstância entre as turmas de ingressantes nos cursos do ICEA, sendo que turmas de anos diferentes possuem características significativamente diferentes. Já em relação à Classificação no Vestibular, o impacto da variável sugere que o desempenho do aluno no vestibular conduz à presença de características diferentes no próprio.

Em situações em que o PCA é utilizado para redução de dimensionalidade, é uma prática comum utilizar as primeiras variáveis que somem 80% da variação explicada em um novo conjunto de dados e descartar as demais variáveis. No entanto, como explicado

anteriormente, não há a necessidade de reduzir o conjunto de dados neste trabalho, visto que o conjunto de dados não é suficiente grande para causar intempéries em relação à performance dos algoritmos.

4.3 Análise Descritiva

Conforme Jain (1991) um modelo de regressão permite estimar ou predizer uma variável aleatória como uma função formada por diversas outras variáveis. A variável estimada é chamada variável de resposta e as variáveis usadas para predizer a resposta são chamadas variáveis preditoras, preditores ou fatores. A análise de regressão assume que todas as variáveis preditoras sejam quantitativas para que a execução de operações aritméticas sejam possíveis, logo percebe-se a importância de obter um conjunto de dados padronizado como apresentado na Tabela 9. O modelo de regressão linear múltipla é descrito pela Equação 4.2.

$$Y_i = \beta_0 + \sum_{p=1}^{P} X_{i,p}\beta_p + \epsilon_i, \tag{4.2}$$

em que Y_i é a variável de resposta, que neste trabalho corresponde à variável binária "*IsEvadido*", e determina se o aluno evadiu ou não. $X_{i,p}$ são as variáveis preditoras, que neste trabalho são representadas pelas 17 características (i.e, variáveis) relativas ao discente, β_p representa o coeficiente estimado da regressão dados pelo método dos mínimos quadrados, por fim, ϵ_i é o erro para i termos. Assim, apresentado o modelo de regressão linear, recordamos que na Subseção 4.2.1 foi identificado que 2 pares de variáveis estavam fortemente relacionadas, de forma que poderia ser não proveitoso a inclusão dessas variáveis no conjunto de dados usado para treinamento dos modelos de predição. As variáveis identificadas foram: Aprovado/Matriculado; Media; Pontuação Vestibular e Classificação Vestibular.

Dada a necessidade de avaliar o impacto dessas variáveis para os modelos de predição, foram realizados testes de regressão linear múltipla utilizando diferentes conjuntos de variáveis como preditores de entrada no algoritmo. Para isso, foi aplicado o método *Backward Elimination* (LINDSEY; SHEATHER, 2010), o qual consiste em iniciar os testes com todas as variáveis e eliminar as variáveis de interesse progressivamente nos testes a fim de analisar o impacto dessas no modelo. Logo, foram realizados 7 testes com configurações diferentes que resultaram em diferentes coeficientes de determinação R^2.

O coeficiente de determinação R^2 explica o desempenho de uma regressão, quanto maior o valor de R^2, melhor é a regressão. Os resultados dos testes são apresentados na Tabela 10.

Tabela 10 – Resultados da Regressão Linear para as diferentes configurações de *dataset*.

Variáveis no *dataset*	R^2
Todas Variáveis (17 variáveis)	**0.8525**
Removida a Pontuação Vestibular (16 variáveis)	0.8523
Removida a Classificação Vestibular (16 variáveis)	0.8514
Removidas Pontuação Vestibular e Classificação Vestibular (15 variáveis)	0.8514
Removida a Média (16 variáveis)	0.8523
Removida a Aprovado/Matriculado (16 variáveis)	0.8487
Removidas Média e Aprovado/Matriculado (15 variáveis)	0.8361

Ao analisar os resultados da Tabela 10 nota-se que o melhor resultado foi obtido quando foram utilizadas todas as variáveis no modelo de regressão. Percebe-se que para determinadas variáveis, como na remoção da Pontuação Vestibular ou Média, a queda de desempenho foi baixa, o que cria a possibilidade de remoção dessas variáveis a fim de melhorar o tempo de execução dos modelos. No entanto, dado que o tempo de execução para o treinamento dos modelos na seção 4.4 foi satisfatório e como este trabalho visa o máximo desempenho preditivo, esse cenário não se aplica ao projeto.

Dessa forma, as hipóteses quanto à remoção no *dataset* de uma variável de cada par de variáveis correlacionadas foram descartadas e será utilizado o conjunto de dados completo para treinamento dos modelos, apresentado na Tabela 9.

4.4 Análise Preditiva

Nesta seção é detalhado o processo realizado para a construção e execução dos modelos preditivos. Os modelos seguiram a mesma metodologia de treino, utilizando *GridSearch* combinada com *K-fold Cross Validation*, detalhada na subseção 4.4.1. Dentre os modelos utilizados para predição, tem-se: Regressão Logística, XGBoost, *Support Vector Machines* (SVM), Árvores de Decisão e *Random Forest*. Os modelos são detalhados na subseção 4.4.2. Por fim, a metodologia para os testes e resultados da análise preditiva são apresentados na Subseção 4.4.3.

4.4.1 Metodologia de Treino e Teste

Definidas as variáveis para o *dataset* e tratados os dados, há de se definir como esses dados serão utilizados para o treinamento e teste dos modelos de predição. Logo, é habitual dividir os dados em duas partes, sendo: (a) *dataset* de treino; e (b) *dataset* de teste.

O *dataset* de treino é utilizado para a construção dos modelos de predição, ou seja, é a partir desses dados que os parâmetros internos de cada modelo são estimados. Por sua vez, o *dataset* de treino é utilizado para verificar o desempenho do modelo de predição

gerado à partir do *dataset* de treino, isto é, averiguar se o modelo tem bom desempenho em predizer se um aluno irá ou não evadir do curso de graduação.

Além dos parâmetros internos de cada modelo, também chamados de coeficientes ou pesos a depender do modelo, por se tratar de modelos de *Machine Learning*, os modelos deste trabalho também possuem hiper-parâmetros. Os hiper-parâmetros são configurações externas ao modelo cujos valores não podem ser estimados a partir do conjunto de dados. Essas configurações são geralmente ajustadas de acordo com o problema de modelagem preditiva e podem ser ajustadas de forma empírica, pela experiência do desenvolvedor, ou por heurísticas e processos que ajudam na seleção desses, como o *Grid Search* e *Random Search*.

Dada a importância dos hiper-parâmetros, neste trabalho foi utilizada a técnica *Grid Search Cross Validation* para a seleção destes. O *Grid Search* é uma técnica de optimização de que busca calcular os valores ótimos para os hiper-parâmetros (HUANG; MAO; LIU, 2012; WENWEN et al., 2014). Uma busca exaustiva é realizada de acordo com o conjunto de valores definido pelo desenvolvedor, que nesse trabalho foram selecionados com base nos possíveis valores de hiper-parâmetros detalhados na documentação de cada respectivo modelo preditivo.

Além disso, a técnica *Grid Search* é combinada com a *K-fold Cross Validation*. Na *K-fold Cross Validation* o conjunto de dados de treino é dividido em K subconjuntos (ou *folds*) mutuamente exclusivos. Cada subconjunto é usado uma vez para validar o desempenho do modelo de previsão que é gerado a partir dos dados combinados dos $K-1$ subconjuntos restantes, levando à K estimativas de desempenho independente. Neste trabalho foram testados os modelos de predição utilizando $K = 5$ e $K = 10$ *folds*. Assim, sendo que os modelos obtiveram, em geral, melhor desempenho com 5 *folds* e dado que a base de dados de alunos não é consideravelmente grande, foi definido $K = 5$ por fins de desempenho e proporção de dados.

Com exceção do último teste de predição dos modelos, o qual usa ordem cronológica dos dados e é descrito na seção 4.4, as demais etapas de treino e teste dos modelos foram elaboradas com os dados dispostos de forma aleatória. A aleatorização dos dados é possível devido a ausência de restrições de ordem cronológica no conjunto de dados, não provocando prejuízos aos modelos. Dessa forma, os modelos foram treinados com diversas configurações de hiper-parâmetros diferentes devido às combinações geradas pelo *Grid Search*, além da configuração padrão de cada modelo com base em sua respectiva biblioteca. Ao final do treinamento das diferentes configurações de modelo candidatas, foi selecionado a melhor configuração de hiper-parâmetros de acordo com a acurácia obtida pelo modelo na fase de treino e validação.

A proporção utilizada para divisão da base de dados entre *dataset* de treino e *dataset* de teste é outro ponto importante para a construção dos modelos. É comumente

utilizado 80% dos dados para treinamento do modelos e 20% para etapa de testes, a fim de avaliar o desempenho dos modelos. Assim, nesse trabalho foi utilizada a proporção de 80/20 na divisão dos dados. No entanto, por padrão somente seria possível executar um teste com os 20% restantes dos dados para cada modelo. Dado isso, a fim de obter uma melhor confiabilidade estatística nos resultados obtidos pelos modelos, foi adotada uma metologia de teste mais sofisticada, possibilitando a obtenção da média de desempenho de cada modelo com seus respectivos intervalos de confiança.

Primeiramente, os modelos foram treinados com 80% dos dados a fim de otimizar os hiper-parâmetros de cada modelo com a técnica *Grid Search*. Descobertas as melhores configurações para cada modelo, os modelos foram submetidos à etapa de testes. Para cada modelo, foram realizadas 10 execuções em *datasets* de testes diferentes. Para isso, cada execução segue o seguinte processo:

1. Os *datasets* de treino e teste, mutuamente exclusivos, são gerados na proporção 80/20 a partir dos dados dos discentes embaralhados aleatoriamente (dessa forma a cada execução são gerados *datasets* de treino e teste diferentes);
2. Cada modelo preditivo é reconstruído do zero utilizando a melhor combinação de hiper-parâmetros;
3. Cada modelo é treinado com 80% dos dados;
4. Cada modelo é avaliado com 20% dos dados;
5. São emitidas as métricas de desempenho para cada execução.

Ao final do processo, é realizada a média das 10 execuções para cada algoritmo. Os resultados são sumarizados e apresentados na subseção 4.4.3 contendo a média, desvio padrão e intervalo de confiança do desempenho de cada modelo.

Além dessa avaliação de desempenho, foi adicionado outro teste independente para avaliar o desempenho dos modelos para dados cronologicamente contínuos. O conjunto de dados foi ordenado cronologicamente pela matrícula do discente e divido de forma que o *dataset* de teste fosse formado pelos ingressantes dos dois últimos semestres de nosso acompanhamento. Ou seja, o *dataset* de treino foi formado pelos discentes que ingressaram do semestre 2005.1 até 2013.1, e o *dataset* de teste pelos alunos que ingressaram em 2013.2 e 2014.2. Essa abordagem corresponde a aproximadamente 90% dos dados para treino e 10% dos dados para teste. Dessa forma, os modelos preditivos foram submetidos à essa abordagem e os resultados são apresentados na subseção 4.4.3.

As 3 etapas de treino e teste detalhadas nesta seção são ilustradas na Figura 21.

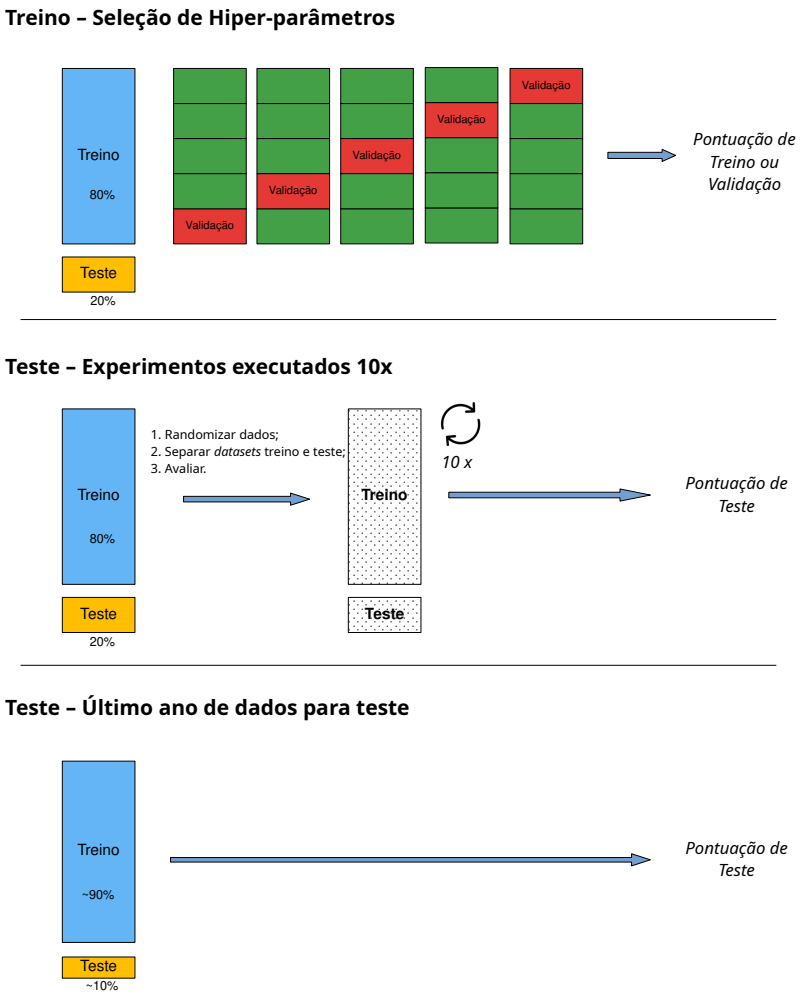


Figura 21 – Metodologia de experimentação da divisão do *dataset* em Treino e Teste.

4.4.2 Modelos de Predição

Dado que nesse trabalho objetivamos solucionar um problema de classificação, foram selecionados modelos preditivos para classificação com base nos resultados obtidos na literatura em problemas de evasão escolar, e também com base no estudo do estado da arte para modelos de classificação, de forma a encontrar modelos que revelem grande potencial para o problema. Dessa forma, foram selecionados os 5 modelos de classificação, sendo eles: *Logistic Regression*, XGBoost, *Support Vector Machines* (SVM), *Decision Trees* (DT) e *Random Forest*.

A técnica de Regressão Logística (do Inglês *Logistic Regression*) tem como objetivo construir um modelo que seja capaz de predizer os valores tomados por uma variável dependente categórica, geralmente de natureza dicotômica ou binária, mas que também pode ser multi-classe. Quanto às variáveis independentes, estas podem atribuir qualquer tipo de dado (JR; LEMESHOW; STURDIVANT, 2013).

A Regressão Logística é uma generalização da Regressão Linear. Como a variável de resposta é discreta, ela não pode ser modelada diretamente por regressão linear. Portanto,

ao invés de prever um estimativa pontual do evento em si, ela constrói o modelo para prever as chances de sua ocorrência. Embora a regressão logística tenha sido uma ferramenta estatística comum para problemas de classificação, sua suposições restritivas sobre a normalidade e independência levaram a um aumento no uso e popularidade de técnicas de aprendizado de máquina para problemas de predição do mundo real (DELEN, 2010).

Na fase de treinamento do modelo nesse trabalho, foram executadas 961 configurações diferentes de hiper-parâmetros para o modelo, totalizando em 4805 execuções considerando que cada modelo foi avaliado 5 vezes devido ao *5-Fold Cross Validation* presente no *Grid Search*. A combinação de hiper-parâmetros para a Regressão Logística é descrita a seguir:

- Conjunto de hiper-parâmetros da Regressão Logística:
  1. Conjunto 1:
     - "penalty": ['l2', 'none'];
     - "C": numpy.logspace(-4, 4, 20);
     - "solver": ['lbfgs', 'newton-cg', 'sag'];
     - "max_iter": [100, 1000, 2500, 5000].
  2. Conjunto 2:
     - "penalty": ['l1', 'l2'];
     - "C": numpy.logspace(-4, 4, 20);
     - "solver": ['liblinear'];
     - "max_iter": [100, 1000, 2500, 5000].
  3. Conjunto 3:
     - "penalty": ['l1', 'l2', 'elasticnet', 'none'];
     - ""C": numpy.logspace(-4, 4, 20);
     - "solver": ['saga'];
     - "max_iter": [100, 1000, 2500, 5000].

O XGBoost, que significa *eXtreme Gradient Boosting*, é uma biblioteca de aprendizado de máquina escalável e distribuída de *Gradient-Boosted Decision Tree* (GBDT). Ele é amplamente usado por cientistas de dados para obter resultados ao nível do estado da arte em muitos desafios de aprendizado de máquina, como em competições na plataforma *Kaggle*[3] (CHEN; GUESTRIN, 2016). Dado isso, o XGBoost é uma biblioteca que se destaca para problemas de regressão e classificação.

[3] <https://www.kaggle.com>

Nesse trabalho o XGBoost foi executado em 730 configurações distintas de hiper-parâmetros, até que se obtivesse a configuração que melhor se adaptasse ao conjunto de dados. A combinação de hiper-parâmetros para o XGBoost é descrita a seguir:

- Conjunto de hiper-parâmetros do XGBoost:
  - "subsample": [0.5, 0.75, 1];
  - "colsample_bytree": [0.5, 0.75, 1];
  - "max_depth": [2, 6, 12];
  - "min_child_weight": [1, 5, 15];
  - "learning_rate": [0.3, 0.1, 0.03];
  - "n_estimators": [100, 500, 1000].

As Máquinas de Vetores de Suporte (do Inglês *Support Vector Machines* (SVM)) constituem uma técnica de aprendizado de máquina supervisionado que é usada para problemas de classificação ou regressão. A SVM possui uma função de decisão que é baseada na combinação linear das características presentes no conjunto de dados utilizado para o treinamento do modelo. A função de mapeamento em SVMs pode ser uma função de classificação (usada para categorizar os dados, como é o caso deste trabalho) ou uma função de regressão (usada para estimar o valor numérico da variável desejada). Para classificação, funções de *kernel* não lineares são frequentemente usadas para transformar os dados de entrada (representando relações não lineares) para um espaço de características de alta dimensão em que os dados de entrada se tornam mais separáveis (ou seja, linearmente separáveis) em comparação com o espaço de entrada original (DELEN, 2010)(HEARST et al., 1998). Em outras palavras, o que uma SVM faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados de diferentes classes.

Na fase de treinamento, foram executadas 76 configurações diferentes de hiper-parâmetros para o modelo de classificação SVM. A combinação de hiper-parâmetros para o SVM é descrita a seguir:

- Conjunto de hiper-parâmetros do SVM:
  - "C": [0.1, 1, 10, 100, 1000];
  - "gamma": [1, 0.1, 0.01, 0.001, 0.0001];
  - "kernel": ['rbf', 'poly', 'sigmoid'].

As Árvores de Decisão (do Inglês *Decision Trees* (DTs)) são um método de aprendizado supervisionado não paramétrico introduzido por Quinlan (1986) e usado para

problemas de classificação e regressão. O objetivo do método é criar um modelo que preveja o valor de uma variável de destino aprendendo regras de decisão simples inferidas dos recursos de dados. Ao final da elaboração do modelo, este pode ser representado por uma tabela de decisão em forma de árvore (MORET, 1982). As árvores de decisão são algoritmos de classificação poderosos que estão se tornando cada vez mais populares devido às suas características intuitivas de capacidade de explicação (DELEN, 2010).

Nesse trabalho foram avaliadas 25 configurações diferentes para as Árvores de Decisão na etapa de treinamento. A combinação de hiper-parâmetros para as Árvores de Decisão é descrita a seguir:

- Conjunto de hiper-parâmetros do modelo Árvores de Decisão:
  - "criterion": ['gini', 'entropy'];
  - "max_depth": [10, 11, 12, 15, 20, 30, 40, 50, 70, 90, 120, 150].

No entanto, vale ressaltar que as Árvores de Decisão podem apresentar instabilidade em seus resultados, ou seja, criar árvores de complexidades distintas e apresentar resultados preditivos diferentes (LI; BELFORD, 2002). Para a maioria dos testes desse trabalho isso não causa distúrbios, já que é calculada a média do desempenho dos algoritmos para o conjunto de execuções, porém para testes de uma única execução, como a última etapa de testes apresentada na Tabela 18, esse comportamento pode ser importuno por apresentar a desempenho diferente a cada execução. Dado isso também foi adicionado o modelo *Random Forest*, o qual soluciona a instabilidade das Árvores de Decisão.

As Florestas Aleatórias ou Florestas de Decisão Aleatória (do Inglês *Random Forest*) é um método de aprendizado conjunto para classificação e regressão. Em sua essência, o *Random Forest* consiste em uma coleção (conjunto) de árvores de decisão simples, cada uma capaz de produzir uma resposta quando apresentada a um conjunto de valores preditores. Assim, o modelo opera construindo uma infinidade de árvores de decisão individuais em tempo de treinamento que, no caso de problemas de classificação, geram uma classe de saída que é selecionada pelas árvores de decisão individuais. A técnica de *Random Forest* soluciona problemas comuns nas Árvores de Decisão, como instabilidade e *over-fitting* (BREIMAN, 2001) (DELEN, 2010).

Na fase de treinamento, foram executadas 5601 configurações diferentes de hiper-parâmetros para o modelo de classificação *Random Forest*. A combinação de hiper-parâmetros é descrita a seguir:

- Conjunto de hiper-parâmetros do *Random Forest*:
  - “criterion”: ['gini', 'entropy'];

- "max_depth": [None, 10, 15, 20, 50, 90, 150];
- "max_features": ['sqrt', 'log2', 2, 3];
- "min_samples_leaf": [1, 2, 3, 4, 5];
- "min_samples_split": [2, 8, 10, 12];
- "n_estimators": [100, 200, 300, 500, 1000].

4.4.3 Resultados

Nessa seção são apresentados os resultados para os 5 modelos de predição descritos na subseção 4.4.2. A forma de avaliação segue o processo detalhado na subseção 4.4.1, logo, temos avaliações para 3 cenários diferentes, sendo:

a) Pontuação para a etapa de treino e validação (*Train/Validation Score*);

b) Pontuação média para a execução de 10 testes na configuração 80% dados para treino e 20% para teste (*Test Score*);

c) Pontuação para teste utilizando o último ano de registros (semestres 2013.2 e 2014.1 para testes e semestres anteriores para treino – *Test Score*).

Para que seja possível avaliar o desempenho dos modelos, esse trabalho conta com as seguintes métricas de desempenho:

- Acurácia: expressa a quantidade de itens (alunos nesse trabalho) que foram classificados de forma correta, em porcentagem;
- Precisão (do inglês *Precision*): dentre todas as classificações de Classe Positivo expressa quantas estão corretas. A precisão é dada pela equação:

$$Precision = \frac{TP}{TP + FP}, \tag{4.3}$$

  em que TP significa Verdadeiro Positivo e FP significa Falso Positivo;

- *Recall*: dentre todas as situações em que a Classe Positivo era o valor esperado, expressa quantas estão corretas. O *recall* é dado pela equação:

$$Recall = \frac{TP}{TP + FN}, \tag{4.4}$$

  em que TP significa Verdadeiro Positivo e FN significa Falso Negativo;

- *F1-Score*: também chamado de *F-Score* ou *F-measure*, é definido como a média harmônica entre precisão e *Recall*. Um alto valor de *F1-Score* indica que ambos precisão e *recall* são razoavelmente altos (AMERI et al., 2016). É dado pela equação:

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{4.5}$$

Em relação à precisão e ao *recall*, resumidamente, alta precisão significa que o modelo retornou substancialmente mais resultados relevantes que irrelevantes. Já alto *recall* significa que o algoritmo retornou a maioria dos resultados relevantes.

Na fase de treino e validação milhares de modelos foram avaliados pela técnica *Grid Search* utilizando *5-fold Cross Validation*. Ao final dos treinos, foi selecionada a configuração de hiper-parâmetros que obteve melhor desempenho para cada um dos 5 modelos preditivos. O resultado médio para as *5-folds* executados no melhor classificador de cada modelo preditivo é apresentado na Tabela 11.

Tabela 11 – Acurácia dos modelos preditivos na fase de treino e validação utilizando *5-Fold Cross Validation*.

Modelo de Classificação	**Acurácia (desvio padrão)**
Logistic Regression	**0.9744 (0.0163)**
SVM	0.9728 (0.0187)
Random Forest	0.9648 (0.0206)
XGBoost	0.9584 (0.0186)
Decision Trees	0.9489 (0.0109)

Percebe-se pelos resultados apresentados na Tabela 11 que todos os modelos preditivos tiveram bom desempenho já na etapa de treino e validação, em sua maioria acima de 95% de acurácia na predição da evasão dos alunos. Ademais, dois modelos de classificação se destacaram com base nos resultados, sendo eles Regressão Logística e SVM.

Além disso, a partir da etapa de treino e validação foi possível obter a relação das variáveis que tem mais impacto no processo de decisão do modelo XGBoost. Com base nos dados de treinamento, o XGBoost ajusta a importância de cada variável para a predição da evasão de forma facilmente compreendível por nós, o que não acontece em outros métodos como a Regressão Logística, já que esta ajusta os coeficientes (também chamados de pesos) de cada variável em uma escala de valores diferente. Dado isso, a importância de cada variável no processo de decisão do XGBoost é apresentada na Figura 22.

Ao analisar a Figura 22 nota-se que o XGBoost destaca a variável Número de Aprovações em seu processo de decisão. Geralmente um aluno com alto número de aprovações tem um longo tempo de curso também, o que indica que o aluno não está em seus semestres iniciais, os quais contém a maior parte das evasões como foi revelado na Tabela 7. Em seguida, o modelo considera relevante a variável Ano de Admissão, a qual

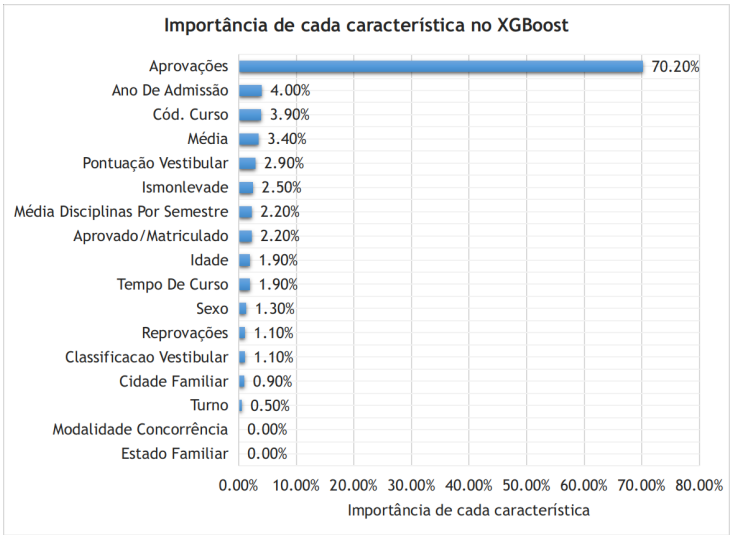


Figura 22 – Importância de cada característica no processo de decisão do XGBoost.

também é relacionada ao tempo de curso do aluno. Ademais, entre as 5 características que mais impactam na decisão do algoritmo é importante ressaltar as variáveis Média e Pontuação no Vestibular, pois, a partir dessas é possível inferir que o desempenho do discente no curso de graduação e no Vestibular para ingressar na universidade tem papel importante no futuro acadêmico do discente.

Realizada a etapa de treino e validação, e selecionadas as melhores configurações de hiper-parâmetros para os modelos, é iniciada a etapa de testes. Para a primeira etapa de testes os algoritmos são executados 10 vezes com diferentes *datasets* de treino e teste. Assim, além da acurácia e desvio padrão dos resultados, também é calculado o intervalo de confiança dos resultados. O intervalo de confiança é dado pela seguinte equação:

$$\overline{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \tag{4.6}$$

em que $\overline{x}$ é a média da amostra, σ é o desvio padrão da amostra, n é o tamanho da amostra e $z_{1-\alpha/2}$ é o quantil de uma variável norma unitária (JAIN, 1991). Neste trabalho será utilizado um nível de confiança de 95%, o que corresponde a $z_{1-\alpha/2} = 1.960$ para nossos cálculos.

Diante disso, os resultados médios para os 10 testes utilizando a proporção 80/20 para os *datasets* de treino e teste, respectivamente, são apresentados na Tabela 12.

Ao analisar a Tabela 12, notam-se altos níveis de acurácia para os modelos de classificação. Apesar da Regressão Logística e do modelo SVM se destacarem novamente, é visto que o intervalos de confiança dos resultados da maioria dos modelos se sobrepõem. Dessa forma, compreende-se o potencial dos modelos, porém os resultados mais elevados não são significativamente melhores a ponto de superar os demais modelos.

Acresce que, para cada um dos modelos utilizados, foi realizado um detalhamento dos resultados utilizando as métricas Precisão, *Recall* e *F1-Score*. Os resultados para cada

Tabela 12 – Acurácia média dos modelos preditivos para 10 testes com intervalo de confiança de 95%.

Modelo de Classificação	**Acurácia (desvio padrão)**	**Intervalo de Confiança**
SVM	0.9739 (0.0135)	(0.9655, 0.9823)
Logistic Regression	0.9732 (0.0102)	(0.9669, 0.9795)
XGBoost	0.9618 (0.0131)	(0.9537, 0.9699)
Random Forest	0.9618 (0.0107)	(0.9552, 0.9684)
Decision Trees	0.9535 (0.0173)	(0.9428, 0.9642)

modelo é apresentado separadamente nas Tabelas 13, 14, 15, 16 e 17.

Tabela 13 – Métricas de desempenho para a Regressão Logística em 10 testes com intervalo de confiança de 95%.

	Logistic Regression		
	Precision	**Recall**	**F1-Score**
Não Evadido	(0.9344, 0.9645)	(0.9586, 0.9842)	(0.9509, 0.9692)
Evadido	(0.979, 0.992)	(0.9662, 0.9823)	(0.9744, 0.9851)

Tabela 14 – Métricas de desempenho para a XGBoost em 10 testes com intervalo de confiança de 95%.

	XGBoost		
	Precision	**Recall**	**F1-Score**
Não Evadido	(0.9477, 0.9792)	(0.8957, 0.9477)	(0.9282, 0.9545)
Evadido	(0.9499, 0.9735)	(0.9732, 0.99)	(0.9647, 0.9781)

Tabela 15 – Métricas de desempenho para a SVM em 10 testes com intervalo de confiança de 95%.

	SVM		
	Precision	**Recall**	**F1-Score**
Não Evadido	(0.9342, 0.9705)	(0.9641, 0.9907)	(0.9534, 0.9753)
Evadido	(0.9788, 0.9946)	(0.9598, 0.984)	(0.9716, 0.9867)

A partir da análise das métricas de desempenho para cada modelo, pode-se observar que o método *Decision Trees* não obteve boa performance em determinados cenários comparado à outros modelos. Ao comparar o *F1-Score* para a classe Evadido entre a técnica de Regressão Logística e à técnica *Decision Trees* nota-se que o primeiro método é superior ao segundo nesse cenário, não apresentando sobreposição nos intervalos de confiança.

Tendo em conta que a métrica *F1-Score* indica a precisão e *recall* em uma única métrica, podemos utilizar essa única métrica para comparar os modelos de forma geral. Pode-se observar que os modelos SVM e Regressão Logística apresentam as mais altas faixas de valores para a *F1-Score*, tanto para a classe Evadido quanto para a classe Não Evadido. Esses modelos são seguidos pelo XGBoost, *Random Forest* e *Decision Trees* respectivamente, sendo que o modelo *Decision Trees* obteve os resultados mais baixos na métrica.

Tabela 16 – Métricas de desempenho para a técnica Árvores de Decisão em 10 testes com intervalo de confiança de 95%.

	Decision Trees		
	Precision	**Recall**	**F1-Score**
Não Evadido	(0.909, 0.9508)	(0.9152, 0.9591)	(0.9191, 0.9465)
Evadido	(0.9537, 0.98)	(0.9489, 0.9748)	(0.9544, 0.9739)

Tabela 17 – Métricas de desempenho para a técnica *Random Forest* em 10 testes com intervalo de confiança de 95%.

	Random Forest		
	Precision	**Recall**	**F1-Score**
Não Evadido	(0.9198, 0.9524)	(0.9441, 0.9682)	(0.9365, 0.9549)
Evadido	(0.9698, 0.9827)	(0.9547, 0.9745)	(0.9645, 0.9761)

Outra métrica específica revelante é o *Recall* da classe Evadido, pois a partir dela pode-se avaliar capacidade do modelo em classificar alunos como Evadido, mesmo que o modelo apresente alguns Falsos Positivos (alunos classificados como propensos à evadir, mas que não evadiram), os quais não são tão prejudiciais em nosso cenário de estudo. Para essa métrica observa-se que o modelo XGBoost possui os maiores valores, que no entanto não são suficientes para superar os outros modelos, dado que os intervalos de confiança se sobrepoem.

Acresce que, não foi possível definir o modelo com melhor performance geral nesse cenário de testes, visto que os intervalos de confiança de diversas métricas se sobrepõem.

Por fim, os modelos são submetidos à segunda etapa de testes, a qual consiste em realizar a predição da evasão dos discentes pertencentes aos dois últimos semestres de ingressantes do conjunto de dados. Logo, os modelos são treinados com cerca de 90% dos dados (ingressantes de 2005.1 à 2013.1) e testados nos 10% de dados restantes (ingressantes em 2013.2 e 2014.1). Os resultados para o teste são apresentados na Tabela 18.

Tabela 18 – Acurácia média dos modelos preditivos para *dataset* de teste com discentes dos 2 últimos semestres (2013.2 e 2014.1).

Modelo de Classificação	**Acurácia**
SVM	**0.9870**
Logistic Regression	**0.9870**
XGBoost	0.9740
*Decision Trees**	0.9610
Random Forest	0.9351

Pela análise da Tabela 18, nota-se que Regressão Logística e SVM obtiveram excelentes resultados preditivos, seguidos do XGBoost. As Árvores de Decisão aparecem em seguida, no entanto esse modelo é instável e pode apresentar resultados diferentes a cada execução do algoritmo, como explicado na subseção 4.4.2. Em seguida, é apresentado o resultado do modelo *Random Forest*, o qual corrige a instabilidade das Árvores de

Decisão e dispõe de um resultado estável assim como os outros modelos.

Apesar do modelo XGBoost ter se destacado recentemente na literatura, obtendo, muitas vezes, as melhores performances preditivas, ele não foi soberano neste trabalho. Isso pode ser explicado pelo conjunto de dados utilizado, já que o XGBoost é muito utilizado em projetos de *deep learning*, os quais contam com uma alta variedade de variáveis e com conjunto de dados que podem chegar na ordem de milhares de centenas de amostras. Visto que o conjunto de dados utilizado neste trabalho é imensamente menor do que o que geralmente é disponibilizado em *deep learning*, esse fator pode ter impactado negativamente o aprendizado do modelo XGBoost e não acionando todo o seu potencial.

Portanto, foi constatado bons resultados preditivos em todas as etapas para os modelos selecionados. Isso indica que foram encontradas boas configurações de hiperparâmetros na etapa de treinamento e que o conjunto de dados utilizado, aliado à um bom pré-processamento de dados, ofereceu informações valiosas para os modelos preditivos, de forma que esses foram capazes de modelar o comportamento dos discentes. Não houve um modelo que se destacasse significativamente em termos estatísticos na primeira etapa de testes, no entanto os modelos Regressão Logística e SVM, seguidos do XGBoost, obtiveram grande performance na última etapa de testes.

5 Conclusões e Trabalhos Futuros

Neste trabalho foi possível caracterizar os discentes e identificar características que impactam no comportamento destes na universidade, bem como o impacto de determinadas características na decisão do discente de evadir-se do curso. Também foi possível construir modelos de classificação que predissessem, com alta acurácia, a evasão de alunos do curso.

As diferentes análises dos dados trouxeram informações importantes sobre os discentes de EC e SI. Foi identificada uma alta taxa de evasão nesses cursos, em especial no curso de Engenharia de Computação. Percebeu-se também que os primeiros semestres dos cursos de graduação são nos quais se encontram a maior chance do aluno evadir do curso, sendo que mais de 50% dos discentes evadem até o 5º semestre de curso, o que representa aproximadamente a metade do curso.

Dentre as variáveis que mais impactaram na variância dos dados, observou-se que 4 características explicam mais de 50% da variação do perfil dos discentes, sendo elas: Código do Curso, Gênero, Ano de Admissão e Classificação no Vestibular. Já em relação ao impacto de cada variável na decisão dos algoritmos de classificar um aluno como evadido ou não, percebeu-se pelo modelo XGBoost que, além das variáveis Ano de Admissão e Código do curso que são em parte relativas aos cursos, as variáveis que mais impactaram na decisão do algoritmo foram: Número de Aprovações, Média no Curso e Pontuação no Vestibular. Esse resultado confirma que os alunos que mais evadem são aqueles dos semestres iniciais dos cursos e que o desempenho do discente no curso de graduação e no Vestibular para ingressar na universidade tem papel importante no futuro acadêmico do discente.

Por fim, foi possível observar o desempenho de diferentes modelos de classificação quanto a predição da evasão dos discentes. Observou-se que os modelos se adequaram bem aos dados fornecidos e obtiveram resultados satisfatórios, o que indica que as etapas de pre-processamento e otimização dos modelos foram bem executadas. Ademais, para testes em que os modelos foram desafiados a predizer a evasão dos ingressantes dos últimos 2 semestres do conjunto de dados pode-se observar um alto desempenho para os modelos Regressão Logística e SVM, sendo que esses foram melhores que os outros modelos em até 5,55%. A técnica XGBoost seguiu de perto esses modelos, obtendo uma acurácia ligeiramente inferior de 1,33% menor do que a dos modelos anteriores.

5.1 Limitações do Trabalho

Neste trabalho o conjunto de dados utilizado foi limitado aos cursos da área de computação do ICEA, ou seja, aos cursos de Engenharia de Computação e Sistemas de Informação oferecidos na UFOP. Po isso, pode-se utilizar os mesmos modelos para um conjunto maior de dados incluindo outros cursos desde que os dados sejam pré-processados a fim de obter resultados significativos no treinamento e testes dos modelos de classificação.

5.2 Contribuições

A partir deste trabalho é possível compreender a situação atual da evasão escolar nos cursos da área de computação do ICEA além de identificar modelos capazes de predizer a evasão do discente de forma assertiva e com alta acurácia.

Foi realizada a análise e caracterização dos discentes desses cursos de maneira que foi possível encontrar as variáveis que impactam no comportamento dos discentes (variação dos dados), e também as variáveis que tem alto impacto em relação à decisão de um aluno evadir do curso ou não. Ademais, o estudo investiga em detalhes o perfil dos discentes de Engenharia de Computação e de Sistemas de Informação na UFOP.

Ademais, esse trabalho descreve uma metodologia de sucesso para a predição da evasão. Foram descritas etapas que foram importantes para a obtenção de um conjunto de dados refinado, de maneira a potencializar o treinamento dos modelos de classificação. Também é exposto um modelo de otimização dos modelos e, por fim, a sumarização do desempenho de diferentes modelos de classificação na predição da evasão dos discentes. Essa sumarização com diferentes modelos é valiosa para trabalhos futuros pois pode direcionar os próximos passos deste trabalho.

5.3 Trabalhos Futuros

Existem possíveis caminhos para uma futura exploração do trabalho, como, por exemplo, maior abrangência de dados e/ou modelos. Uma via de pesquisa interessante é relacionada a obtenção de um conjunto de dados maior para ser utilizado no treinamento e teste dos modelos de classificação. Nessa linha, caso sejam adotados dados de diferentes cursos, é importante avaliar o desempenho dos modelos para uma maior heterogeneidade de cursos e, nesse caso, se é oportuno elaborar modelos separados para cada grande área de conhecimento, por exemplo, modelos com os dados separados entre os cursos relacionados à Humanas, Biológicas e Exatas.

Outro possível direcionamento é quanto aos métodos de classificação utilizados. Pode-se avaliar a utilização de outros métodos de classificação, dada a imensidão de

métodos disponíveis na literatura. Também é possível adotar outros meios de otimização de métodos, como o uso de *Random Search* ou *Bayesian Optimization*.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010. Citado na página 50.

ALMEIDA, O. C. d. S. de. Evasão em cursos a distância: análise dos motivos de desistência. In: *14 Congresso Internacional ABED de Educação a Distância. Santos–São Paulo, Brasil*. [S.l.: s.n.], 2008. v. 7. Citado na página 22.

AMERI, S. et al. Survival analysis based framework for early prediction of student dropouts. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. [S.l.: s.n.], 2016. p. 903–912. Citado 3 vezes nas páginas 26, 27 e 61.

ASSOCIAÇÃO BRASILEIRA DAS EMPRESAS DE TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO. *Formação Educacional e Empregabilidade em TIC*: Achados e recomendações. São Paulo, 2019. Disponível em: <https://brasscom.org.br/estudo-brasscom-formacao-educacional-e-empregabilidade-em-tic-achados-e-recomendacoes/bri2-2019-010-p02-formacao-educacional-e-empregabilidade-em-tic-v83/>. Acesso em: 7 ago. 2020. Citado 2 vezes nas páginas 15 e 21.

BEAN, J. P. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education*, Springer, v. 12, n. 2, p. 155–187, 1980. Citado na página 24.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 59.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Citado na página 57.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, Elsevier, v. 49, n. 4, p. 498–506, 2010. Citado 6 vezes nas páginas 26, 27, 43, 57, 58 e 59.

FAYER, S.; LACEY, A.; WATSON, A. Stem occupations: Past, present, and future. *Spotlight on Statistics*, p. 1–35, 2017. Citado 3 vezes nas páginas 15, 19 e 20.

GARCÍA, S.; LUENGO, J.; HERRERA, F. *Data preprocessing in data mining*. [S.l.]: Springer, 2015. v. 72. Citado na página 45.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 45.

HEARST, M. A. et al. Support vector machines. *IEEE Intelligent Systems and their applications*, IEEE, v. 13, n. 4, p. 18–28, 1998. Citado na página 58.

HIGHER EDUCATION STATISTICS AGENCY. *Non-continuation: UK Performance Indicators 2018/19*. United Kingdom, 2020. Disponível em: <https://www.hesa.ac.uk/data-and-analysis/performance-indicators/non-continuation-1819>. Acesso em: 7 ago. 2020. Citado na página 20.

HUANG, Q.; MAO, J.; LIU, Y. An improved grid search algorithm of svr parameters optimization. In: IEEE. *2012 IEEE 14th International Conference on Communication Technology*. [S.l.], 2012. p. 1022–1026. Citado na página 54.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Censo da Educação Superior 2010-2014*: Diretoria de estatísticas educacionais. Brasília, DF, 2016. Citado na página 22.

JAIN, R. The art of computer systems performance analysis-techniques for experimental design. *Measurement, Simulation, and Modeling*, John Wiley & Sons, 1991. Citado 3 vezes nas páginas 51, 52 e 62.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado na página 56.

KOTSIANTIS, S. B.; KANELLOPOULOS, D.; PINTELAS, P. E. Data preprocessing for supervised leaning. *International journal of computer science*, Citeseer, v. 1, n. 2, p. 111–117, 2006. Citado na página 45.

LEE, Y.; CHOI, J. A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, Springer, v. 59, n. 5, p. 593–618, 2011. Citado na página 21.

LI, R.-H.; BELFORD, G. G. Instability of decision tree classification algorithms. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2002. p. 570–575. Citado na página 59.

LINDSEY, C.; SHEATHER, S. Variable selection in linear regression. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 10, n. 4, p. 650–669, 2010. Citado na página 52.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, v. 25, 2012. Citado 2 vezes nas páginas 22 e 23.

MORET, B. M. Decision trees and diagrams. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 14, n. 4, p. 593–623, 1982. Citado na página 59.

OLSON, S.; RIORDAN, D. G. Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. report to the president. *Executive Office of the President*, ERIC, 2012. Citado 5 vezes nas páginas 15, 16, 19, 20 e 21.

PALMEIRA, L. B.; SANTOS, M. P. Evasão no bacharelado em ciência da computação da universidade de brasília: análise e mineração de dados. 2014. Citado 2 vezes nas páginas 15 e 21.

PASCARELLA, E. T.; TERENZINI, P. T. Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The journal of higher education*, Taylor & Francis, v. 51, n. 1, p. 60–75, 1980. Citado na página 24.

PLATT, A.; FAN-OSUALA, O.; HERFEL, N. Understanding and predicting student retention and attrition in it undergraduates. In: *Proceedings of the 2019 on Computers and People Research Conference*. [S.l.: s.n.], 2019. p. 135–138. Citado 2 vezes nas páginas 25 e 27.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986. Citado na página 58.

SCHWAB, K. *A quarta revolução industrial*. [S.l.]: Edipro, 2019. Citado na página 15.

SHAPIRO, D. et al. Completing college: A national view of student completion rates–fall 2012 cohort (signature report no. 16). *National Student Clearinghouse*, ERIC, 2018. Citado na página 21.

SPADY, W. G. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, Springer, v. 1, n. 1, p. 64–85, 1970. Citado na página 24.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975. Citado na página 24.

VITELLI, R. F. Gestão da evasão na unisinos: combate e prevenção. In: *XXX Congresso da Sociedade Brasileira de Computação (CSBC)*. [S.l.: s.n.], 2010. Citado 2 vezes nas páginas 25 e 27.

WENWEN, L. et al. Application of improved grid search algorithm on svm for classification of tumor gene. *International Journal of Multimedia and Ubiquitous Engineering*, v. 9, n. 11, p. 181–188, 2014. Citado na página 54.