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

Definição de um Modelo de Inteligência Artificial para a Identificação do Padrão Curricular dos Alunos do ICEA

Anna Paula Figueiredo Gonçalves

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:

Fernando Bernardes de Oliveira

COORIENTAÇÃO:

Felipe Coelho Silva

Junho, 2022
João Monlevade–MG

Anna Paula Figueiredo Gonçalves

Definição de um Modelo de Inteligência Artificial para a Identificação do Padrão Curricular dos Alunos do ICEA

Orientador: Fernando Bernardes de Oliveira

Coorientador: Felipe Coelho Silva

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

Universidade Federal de Ouro Preto

João Monlevade

Junho de 2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

G635d Gonçalves, Anna Paula Figueiredo.
Definição de um modelo de inteligência artificial para a identificação do padrão curricular dos alunos do ICEA. [manuscrito] / Anna Paula Figueiredo Gonçalves. - 2022.
73 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Fernado Bernardes de Oliveira.
Coorientador: Prof. Me. Felipe Coelho Silva.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Inteligência artificial - Aplicações educacionais. 2. Programação orientada a dados (Computação). 3. Mineração de dados (Computação). 4. Evasão universitária. I. Oliveira, Fernado Bernardes de. II. Silva, Felipe Coelho. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 004.8:004.62

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

Anna Paula Figueiredo Gonçalves

Definição de um modelo de inteligência artificial para a identificação do padrão curricular dos alunos do ICEA

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação

Aprovada em 24 de junho de 2022

Membros da banca

Prof. Dr. Fernando Bernardes de Oliveira - Orientador (Universidade Federal de Ouro Preto)
Prof. MSc. Felipe Coelho Silva - Coorientador (Minds Digital / PUC Minas)
Profa. Dra. Helen de Cássia Sousa da Costa Lima - Avaliadora (Universidade Federal de Ouro Preto)
Prof. Dr. Luiz Carlos Bambirra Torres - Avaliador (Universidade Federal de Ouro Preto)

Fernando Bernardes de Oliveira, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/07/2022

Documento assinado eletronicamente por **Fernando Bernardes de Oliveira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 14/07/2022, às 13:29, 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 **0362541** e o código CRC **14777E60**.

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 todas as pessoas que acreditam na educação como instrumento para transformar vidas, assim como a democratização e acesso à informação.

Agradecimentos

Em primeiro lugar agradeço a Deus pelo sopro de vida, por guiar-me a tomar decisões que me trouxeram até aqui, da mesma forma em que foi meu sustento para vencer os obstáculos no meu caminho.

Agradeço ao corpo discente, em especial à cada monitor e aos colegas de curso que muita das vezes dedicaram seu conhecimento e atenção para me ajudar nos momentos de dificuldade. Reconheço também a competência do corpo docente do curso de Sistemas de Informação, pelo empenho e dedicação ao desenvolver o trabalho de Educador com maestria.

Agradeço aos meus amigos e familiares pela compreensão da minha ausência em momentos festivos. Aos familiares mais próximos, agradeço ao entendimento da minha irritabilidade ao final dos períodos. Em especial ao Lucas pelo incentivo, apoio e paciência, desde sempre e oferecer-me abrigo nos dias difíceis.

Reconheço a atuação dos meus psicólogos durante o período acadêmico, no qual aprendi a arte de manter a calma e confiar nas minhas habilidades. Eles foram um apoio de extrema importância para a autogestão mental e emocional, para que eu conseguisse atingir meus objetivos.

Agradeço à Minds Digital, por oferecer o trabalho remoto e flexibilidade de horário, pois, sem esses aspectos, concluir a graduação seria mais difícil. Aos meus colegas de trabalho, pelo compartilhamento de conhecimento, e vivência de trabalho em equipe.

Ademais, agradeço ao meu orientador e coorientador pelo comprometimento com o trabalho e por oferecer o suporte técnico e moral para a conclusão do mesmo.

"Não basta ter a informação, é preciso saber o que fazer com ela."

— Mario Sergio Cortella

Resumo

Este trabalho apresenta uma metodologia capaz de identificar o perfil dos alunos do Instituto de Ciências Exatas e Aplicadas (ICEA) em relação ao comportamento curricular com foco no sucesso acadêmico. Através da utilização de técnicas de *Artificial Intelligence* (AI) e *Educational Data Mining* (EDM) é possível identificar precocemente os discentes com maior probabilidade de evadir do Instituto. Este problema foi abordado neste trabalho com o desenvolvimento de Modelos de *Machine Learning* (ML) baseados na *Random Forest*. Foram desenvolvidos diversos modelos em diferentes cenários de tal modo a atender de forma eficiente o contexto acadêmico do ICEA. Alguns métodos de *Feature Selections* foram utilizados para compreender as características importantes na tomada de decisão do modelo, de modo a aperfeiçoar o desempenho e os resultados. O cenário que teve o melhor resultado foi o que combinou as variáveis pré e pós universidade, com o método de seleção de características *Recursive Feature Elimination* (RFE), que atingiu aproximadamente 80% de acurácia. Para compreender os resultados do modelo desenvolvido, foi empregue a visualização das variáveis que tiveram impacto relevante no desenvolvimento do modelo por meio da estratégia *Shapley Aditive Explanations* (SHAP).

Palavras-chaves: *Artificial Inteligence. Data Driven. Educational Data Mining.* Evasão.

Abstract

This work presents an efficient methodology to describe the students profile regarding the students curriculum behavior, focused on academic success. Through the use of techniques of *Artificial Intelligence* (AI) and *Educational Data Mining* (EDM), is possible early identification of students with higher chances to drop out of the Institute. This problem has been addressed in this work as development of *Machine Learning* (ML), the base algorithm implemented was the Random Forest. Several Models were developed in different scenarios in order to better serve the Instituto de Ciências Exatas e Aplicadas (ICEA) academic context. Some methods of Feature Selection were used, to understand the important characteristics in the decision making of the Model, in order to optimize the performance and the results. The scenario that had the best result was the one that combined the pre and post university variables with the *Recursive Feature Elimination* (RFE), feature selection method, which reached approximately 80% accuracy. In order to understand the model's results, the view of variables that had relevant impact on its decisions was applied, using the strategy of *Shapley Aditive Explanations* (SHAP).

Key-words: Artificial Inteligence. Data Driven. Educational Data Mining. University Dropout.

Lista de ilustrações

Figura 1 – Histórico de Evasão x Diplomação 16
Figura 2 – Evasão por Período do Curso 17
Figura 3 – Distribuição do Coeficiente Semestral nos Dois Primeiros Anos de Curso 18
Figura 4 – Reprovações nos Períodos Iniciais 18
Figura 5 – *Word Cloud* dos Motivos de Evasão 19
Figura 6 – *Workflow* do projeto 25
Figura 7 – Dados ausentes: registros 29
Figura 8 – Dados ausentes: notas 29
Figura 9 – Exemplo de Árvore de Decisão 41
Figura 10 – Exemplo de *Random Forest* 42
Figura 11 – Exemplo de Matriz de Confusão 43
Figura 12 – Matriz de Confusão Cenário 1. 49
Figura 13 – *Feature Importance* Cenário 1. 50
Figura 14 – Matriz de Confusão Cenário 2. 51
Figura 15 – *Feature Importance* Cenário 2. 51
Figura 16 – Correlações Fortes da Base de Dados. 52
Figura 17 – Matriz de Confusão Cenário 3. 53
Figura 18 – *Feature Importance* Cenário 3. 53
Figura 19 – Seleção de Características - Dados de Treino - Cenário 4. 54
Figura 20 – Matriz de Confusão Cenário 4. 56
Figura 21 – *Feature Importance* Cenário 4. 56
Figura 22 – Matriz de Confusão Cenário 5. 58
Figura 23 – *Feature Importance* Cenário 5. 58
Figura 24 – Matriz de Confusão Cenário 6. 60
Figura 25 – *Feature Importance* Cenário 6. 60
Figura 26 – Matriz de Confusão Cenário 7. 62
Figura 27 – *Feature Importance* Cenário 7. 62
Figura 28 – Matriz de Confusão Cenário 8. 63
Figura 29 – *Feature Importance* Cenário 8. 64
Figura 30 – Matriz de Confusão Cenário 9. 65
Figura 31 – *Feature Importance* Cenário 9. 65
Figura 32 – Explicabilidade do Modelo - Valor SHAP. 66

Lista de tabelas

Tabela 1 – Características Gerais 27
Tabela 2 – Variáveis de Registro 28
Tabela 3 – Variáveis de Nota 28
Tabela 4 – Proporção da Classes do Problema 31
Tabela 5 – Características Gerais Após Limpeza 31
Tabela 6 – Proporção da Classes do Problema Após Limpeza 32
Tabela 7 – *Map* dos dados ausentes. 33
Tabela 8 – Geração de Novos Atributos. 34
Tabela 9 – Tratamento das *Features* 35
Tabela 10 – *One Hot encoding* 35
Tabela 11 – *One Hot encoding* - Modo de Admissão 36
Tabela 12 – Dados sem Agregação 37
Tabela 13 – Agregação por Período e Matrícula 37
Tabela 14 – Agregação por Matrícula. 37
Tabela 15 – Resultado da Modelagem dos Dados 38
Tabela 16 – Divisão Treino, Teste e Validação. 40
Tabela 17 – Parametrização *Random Forest*. 48
Tabela 18 – Variáveis *Select k-best*. 55
Tabela 19 – Variáveis RFE. 57
Tabela 20 – Variáveis Pré Universidade. 59
Tabela 21 – Variáveis Pós Universidade. 61
Tabela 22 – Métricas de Diferentes Classificadores - Treinamento. 67

Lista de abreviaturas e siglas

AI *Artificial Intelligence*

CPF Cadastro de Pessoa Física

CSV *Comma-separated Values*

DECEA Departamento de Ciências Exatas e Aplicadas

DECSI Departamento de Computação e Sistemas

DEELT Departamento de Engenharia Elétrica

DEENP Departamento de Engenharia de Produção

EC Engenharia da Computação

EDM *Educational Data Mining*

EE Engenharia Elétrica

ENEM Exame Nacional do Ensino Médio

EP Engenharia de Produção

IC Inteligência Computacional

ICEA Instituto de Ciências Exatas e Aplicadas

IES Instituto de Ensino Superior

KDD *Knowledge-Discovery in Databases*

LGPD Lei Geral de Proteção de Dados

MEC Ministério da Educação

ML *Machine Learning*

PCA *Principal Component Analysis*

RFE *Recursive Feature Elimination*

SHAP *Shapley Aditive Explanations*

SI Sistemas de Informação

SISU Sistema de Seleção Unificada

SMOTE *Synthetic Minority Oversampling Technique*

UFOP Universidade Federal de Ouro Preto

Sumário

1 INTRODUÇÃO 15
1.1 O Problema 16
1.2 Objetivo 19
1.3 Justificativa 20
1.4 Metodologia 20
1.5 Organização do trabalho 21

2 REVISÃO BIBLIOGRÁFICA 22
2.1 Desempenho Acadêmico 22
2.2 Evasão 22
2.3 Trabalhos correlatados 23
2.4 Considerações finais 23

3 DESENVOLVIMENTO 25
3.1 Ambiente de Desenvolvimento 26
3.2 Análise dos Dados 26
3.3 Coleção dos Dados 26
3.3.1 Fonte dos dados 26
3.3.2 Informações Gerais 27
3.3.3 Tipo das Variáveis 27
3.4 Limpeza 28
3.4.1 Dados Ausentes 28
3.4.2 Definição do intervalo de tempo estudado 30
3.4.3 Desbalanceamento das classes 30
3.4.4 Resultante da Etapa de Limpeza dos Dados 31
3.5 *Feature Engineering* 32
3.5.1 Tratamento de Dados Ausentes 32
3.5.2 Novas Características 33
3.5.3 Pré-Processamento 35
3.5.4 Modelagem 37
3.5.5 Normalização dos Dados 38
3.6 Aprendizado de Máquina 38
3.6.1 Divisão dos Dados 39
3.6.2 *Random Forest* 40
3.6.3 Métricas de Avaliação 43
3.7 Seleção de Variáveis 45

3.7.1 Correlação . . . 45
3.7.2 *Select k-best* . . . 46
3.7.3 *Recursive Feature Elimination* . . . 46
3.7.4 *Shapley Additive Explanations* . . . 46
3.8 Considerações Finais . . . **47**

4 ANÁLISE DOS RESULTADOS . . . **48**
4.1 Cenário 1: Todas as Variáveis . . . **49**
4.2 Cenário 2: Remoção *Tags* de Evasão . . . **50**
4.3 Cenário 3: Remoção da Multicolinearidade . . . **52**
4.4 Cenário 4: Aplicação Seleção de *Features*: *Select k-best* . . . **54**
4.5 Cenário 5: Aplicação Seleção de *Features*: RFE . . . **57**
4.6 Cenário 6: Variáveis Pré Universidade . . . **59**
4.7 Cenário 7: Variáveis Pós Universidade . . . **61**
4.8 Cenário 8: Variáveis Destaques . . . **63**
4.9 Cenário 9: Classes Binárias . . . **64**
4.10 Cenário 10: Outros Algoritmos . . . **67**
4.11 Considerações Finais . . . **67**

5 CONCLUSÃO . . . **69**
5.1 Propostas para trabalhos futuros . . . **70**

REFERÊNCIAS . . . **71**

1 Introdução

Atualmente no mercado competitivo, as instituições de ensino precisam oferecer uma educação eficiente e de qualidade para os alunos (RAMASWAMI; SUSNJAK; MATHRANI, 2022). A crescente disponibilidade de dados no âmbito educacional traz consigo diversos desafios e aprendizados. Entretanto, com a adesão da cultura *data-driven*, no qual os dados são caracterizados como o insumo principal das organizações para tomada de decisão Hartmann et al. (2016), o meio acadêmico também se faz presente nessa realidade. Essa metodologia pode ser usada para melhorar a eficiência organizacional e também para a construção de melhores práticas em sala de aula (HALVERSON et al., 2006).

No ensino superior é de suma importância prever o desempenho dos alunos, seja para oferta de bolsa acadêmica ou a identificação precoce dos alunos que possam vir a reprovar e/ou a evadir (ACHARYA; SINHA, 2014). A aplicação de *Artificial Intelligence* (AI) no campo da Educação ainda está em fase inicial, no qual adotado juntamente às técnicas de *Educational Data Mining* (EDM) David, Clare e Doleck (2021), trouxe resultados significativos, de modo a apoiar a tomada de decisão no contexto universitário.

O Instituto de Ciências Exatas e Aplicadas (ICEA) é um campus da Universidade Federal de Ouro Preto (UFOP) que oferta os cursos de Engenharia da Computação (EC), Engenharia Elétrica (EE), Engenharia de Produção (EP) e Sistemas de Informação (SI). Desde o período letivo de 2011.1, época em que o Exame Nacional do Ensino Médio (ENEM) passou a ser considerado o principal meio de ingresso na Universidade, o instituto conta com o total de 3.654 alunos até o período letivo de 2020.2. Aproximadamente 51% desses alunos não conseguiram o sucesso acadêmico, ou seja, não finalizaram seu respectivo curso e estão em situação de evasão.

Um discente que não possui um bom desempenho no decorrer de sua trajetória universitária, principalmente nas esferas Federal e Estadual interfere negativamente no desenvolvimento do país, uma vez que recursos financeiros são distribuídos para este fim (MUSSLINER et al., 2021). O mau desempenho acadêmico, ocasiona abertura de novas turmas, atrasos no período de formação, sobrecarga de alunos para com os docentes, disciplinas com baixos alunos inscritos, desmotivação, por conseguinte a evasão. São diversos os recursos necessários para manter um aluno matriculado como, por exemplo, mão de obra, disponibilização de equipamentos, custos básicos de funcionamento como água, luz, internet, entre outros. Sendo assim, fica evidente a necessidade de desenvolver uma solução adequada para auxiliar a continuidade dos acadêmicos nas instituições. (IQBAL et al., 2017).

A educação é um processo progressivo e contínuo, logo, pensando no futuro, as

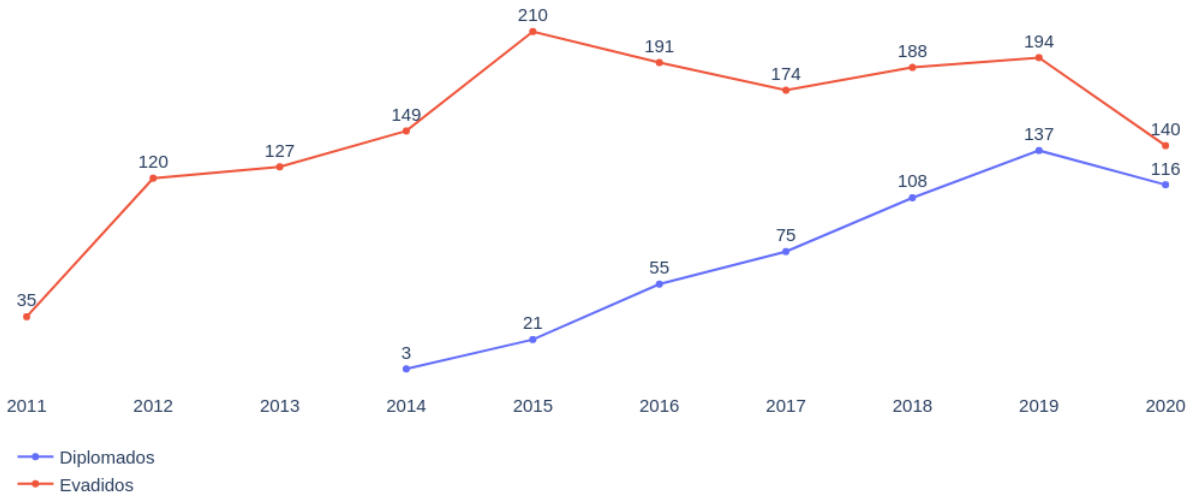
decisões tomadas, devem ser equipadas junto às técnicas de AI para promover uma melhor experiência na área da educação (MUTTATHIL; RAHMAN, 2018). Desse modo, mostra-se um caminho promissor desenvolver um Modelo de *Machine Learning* (ML) capaz de identificar precocemente os alunos com maior probabilidade de evadir.

1.1 O Problema

Esta seção abrange detalhadamente a caracterização do problema de evasão no ICEA e seus respectivos cursos ofertados. O desempenho acadêmico e a evasão são assuntos abordados abertamente desde os anos de 1997 pelo Ministério da Educação (MEC) de modo a apresentar uma educação igualitária e de qualidade para a sociedade. Baseado no censo realizado pelo MEC, Mussliner et al. (2021) identificou que os alunos universitários que iniciaram seus estudos em 2010 apontaram uma taxa de evasão de aproximadamente 56,8%. No instituto, essa temática foi abordada no último ano com a aplicação de técnicas de EDM por Caldeira (2021).

Este trabalho aborda uma análise completa dos dados do ICEA para desenvolver um Modelo de AI apto a diferir os alunos propensos a evadir. Os dados considerados datam do período de 2011 à 2020, totalizando 3.654 matrículas. Posto isso, veja a seguir como a situação dos alunos estão dispostas em relação à diplomação e evasão na Figura 1.

Figura 1 – Histórico de Evasão x Diplomação

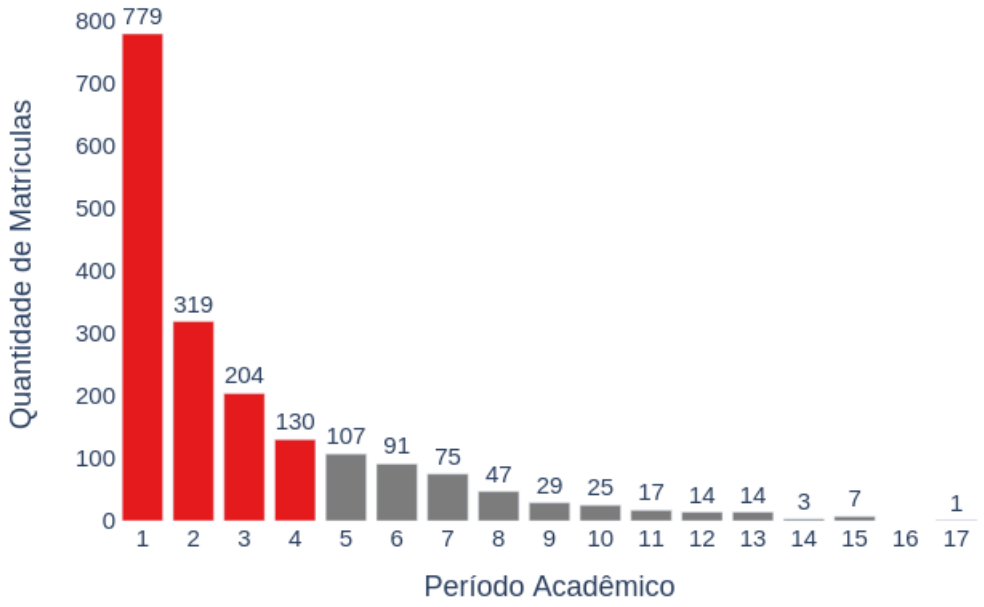


Fonte: Autora do trabalho

Assim, pode-se observar como as taxas de evasão se comparadas às de sucesso acadêmico é sempre superior. É importante frisar que nesta análise, devido à pandemia do

COVID 19, os períodos de 2020 foram realizados remotamente, o que pode ter influenciado a diminuição da taxa de evasão. Pela análise exploratória dos dados, também foi possível apontar o momento em que o aluno evade, ou seja, os períodos com maior taxa de abandono dos cursos. Isso pode ser observado na Figura 2.

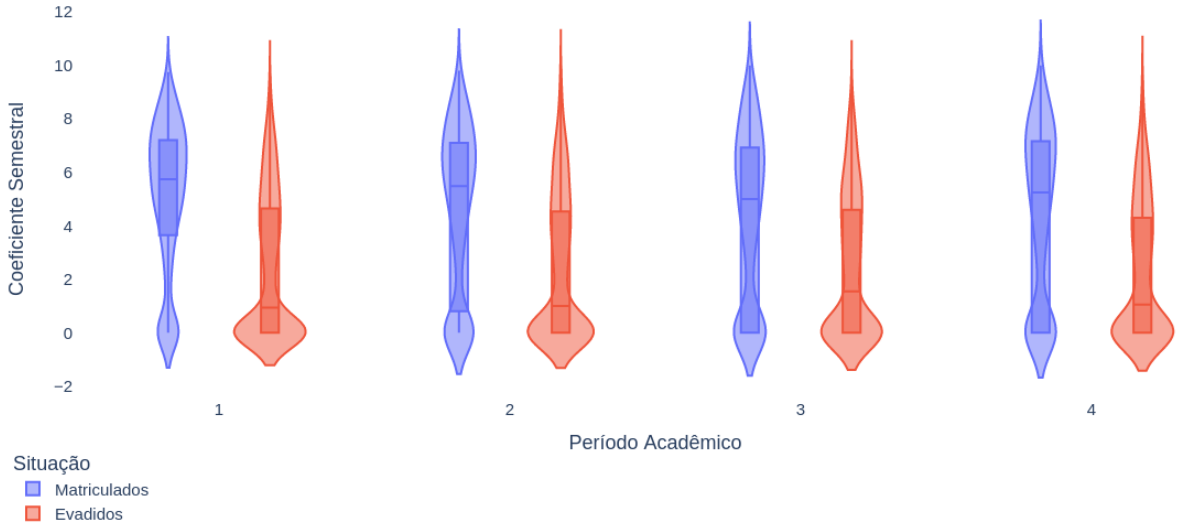
Figura 2 – Evasão por Período do Curso



Fonte: Autora do trabalho

Dessa maneira, é possível destacar que os dois primeiros anos iniciais dos cursos concentram grande parte do grupo de estudo. A autora Lehman (2005) relaciona tais acontecimentos com a falta de políticas de orientação acadêmica, principalmente no primeiro ano, o que se torna um fator importante para a evasão. Essa situação pode ser explicado pelo desempenho dos alunos nos primeiros dois anos, em que o coeficiente semestral dos alunos evadidos e dos matriculados reapresentam 2,32% e 4,61%, respectivamente. A distribuição dessa estatística pode ser observada a seguir na Figura 3.

Figura 3 – Distribuição do Coeficiente Semestral nos Dois Primeiros Anos de Curso



Fonte: Autora do trabalho

Esse comportamento negativo dos alunos nos períodos iniciais é refletido também na obtenção de sucesso nas disciplinas primárias, como apresentado na Figura 4. Dessa maneira, ocorre um excesso de reprovações que desencadeiam problemas como: a necessidade de abertura de novas turmas, sobrecarga de alunos para com os professores, e alta concorrência de vagas nas próximas classes. Possivelmente, esses problemas, também abordados por Mussliner et al. (2021) e Wen e Wen (2020), podem ter como consequência a evasão. Os autores ANDIFES, ABRUEM e SESu/MEC (1996) também relatam que a repetição em disciplinas contribui para tal fim.

Figura 4 – Reprovações nos Períodos Iniciais

Fonte: Autora do trabalho

Em síntese, muita das vezes, não é possível reconhecer ou comprovar os motivos que levaram à tal fato. Entretanto, da base de dados do Instituto de Ciências Exatas e Aplicadas (ICEA) com 1.865 alunos em situação de evasão, cerca de 801 alunos possuem em seus registros a causa que os levaram a tomar essa decisão. A seguir, de acordo com o exposto na Figura 5, é possível visualizar as palavras mais utilizadas pelos alunos ao indicar a razão de sua desistência.

Figura 5 – *Word Cloud* dos Motivos de Evasão

Fonte: Autora do trabalho

Ademais, entre as quatro motivações mais comuns estão:

- Não era o curso almejado
- Aprovação em outra instituição pública
- Motivos financeiros
- Distância entre a UFOP e a cidade natal

1.2 Objetivo

Este trabalho abrange a análise e modelagem dos dados educacionais do ICEA, de modo a identificar o padrão curricular dos alunos através do uso de AI, bem como predizer a evasão dos alunos.

- Estudar e aplicar mecanismos de inteligência computacional para a identificação e classificação do padrão curricular dos alunos.
- Modelar e implementar uma estrutura para processar os dados desenvolvendo um Modelo de classificação supervisionado.

- Validar o Modelo proposto por experimentos computacionais utilizando a coleção de dados disponibilizada.

1.3 Justificativa

A problemática da evasão, amplamente discutido por ANDIFES, ABRUEM e SESu/MEC (1996), é caracterizada por definição referindo-se ao desligamento ou abandono do aluno pelo curso. Essa natureza não é um fato isolado, como aponta Filho et al. (2007), e pode variar de país para país. Segundo Mussliner et al. (2021), para o ano do 2018, as taxas de evasão nas Universidades Públicas no Brasil era de 18,5%. Como apresentado anteriormente, somente no ICEA essa taxa é superior a 50%.

Para manter o funcionamento dos Instituto de Ensino Superior (IES) Públicos, é necessário um alto investimentos de recursos financeiros, seja para a adequação de laboratórios, capacitação de professores e manutenção dos espaços físicos. Atualmente, com os frequentes cortes de gastos na área da Educação, desperdiçar recursos não está em pauta. Em razão dos fatos ditos, o sucesso do aluno está diretamente ligado ao uso eficiente de processos e de bens, uma vez que com o investimento realizado, espera-se como retorno a diplomação, não o abandono. Outrossim, o sucesso acadêmico, principalmente no curso superior é desejável o impacto de forma positiva o mercado econômico apto para oferecer mão de obra de qualidade, desenvolvimento de pesquisas, tecnologias e inovação (MUSSLINER et al., 2021).

Problemas que envolve EDM não são fáceis de modelar, uma vez que as regras de negócio são complexas e existe uma grande dimensionalidade de atributos, ou seja, um volume significativo de características (VILORIA et al., 2020). Todavia, autores como Behr et al. (2020) e Flores, Heras e Julian (2022) utilizam a Inteligência Computacional (IC) para auxiliar no reconhecimento precoce desses alunos e identificar padrões curriculares ocultos através dos dados acadêmicos.

Classificar os alunos que possuem um alto risco de evasão precocemente, permite uma intervenção única para cada indivíduo (IRAJI et al., 2012). Assim sendo, o presente trabalho abordará uma análise singular do contexto do ICEA. Fazer o uso da perspectiva EDM com técnicas de AI, possibilita desenvolver um Modelo de ML capaz de predizer os alunos com maior propensão à evasão. Este Modelo pode servir de apoio para a tomada de decisão durante o processo de ensino.

1.4 Metodologia

Esta seção apresenta a metodologia utilizada no desenvolvimento do trabalho, caracterizado como um estudo aplicado no contexto educacional, precisamente o do

ICEA. Os dados utilizados são referente à trajetória acadêmica dos alunos da instituição disponibilizados pela própria Seção de Ensino do Instituto. A amostra coletada se refere ao intervalo de tempo entre o primeiro período de 2011, ou seja, 2011.1 e do segundo período de 2020, 2020.2.

Devido às regras da Lei Geral de Proteção de Dados (LGPD) os dados sensíveis foram anonimizados, de tal modo não ser possível a identificação dos alunos nesta pesquisa. A partir disso, sucederam as análises e experimentos envolvendo algoritmos de ML com o objetivo de desenvolver um Modelo matemático apropriado para o contexto aplicado, classificando os alunos em situação de evadido, matriculado ou diplomado. Dessa forma, os passos para a elaboração do trabalho, são definidos a seguir:

- Revisar a literatura: apontar e estudar trabalhos correlatos, identificar pontos de similaridade e técnicas utilizadas.
- Tratamento dos dados: coletar os dados, anonimizá-los, buscar outras bases para enriquecer os *insights* obtidos, fazer a tratativa dos dados faltantes.
- Modelagem: criação de novas características, definição de algoritmos, seleção das *features*, dentre outros.
- Avaliação e testes: avaliar as métricas dos Modelos, testar e configurar os parâmetros de modo a maximizar os resultados.
- Analisar e discutir os resultados obtidos, além de identificar possíveis melhorias e considerações gerais sobre o projeto.

1.5 Organização do trabalho

O restante deste trabalho está estruturado do seguinte modo. No Capítulo 2 são apresentados os trabalhos correlatados. O Capítulo 3 aborda os processos utilizados bem como a modelagem do problema. Já no Capítulo 4, é apresentado a análise dos resultados e, ao findar, no Capítulo 5 é discutido as considerações finais e os trabalhos futuros.

2 Revisão bibliográfica

Este capítulo apresenta uma síntese da revisão bibliográfica, considerando os principais trabalhos correlatos na literatura. Os assuntos aqui tratados se referem ao EDM e técnicas de AI para predizer o desempenho dos alunos e a evasão, que é diretamente impactada pelo desempenho acadêmico.

2.1 Desempenho Acadêmico

O desempenho acadêmico está diretamente ligado à qualidade do ensino e sucesso dos alunos, como estudado por (XU; MOON; SCHAAR, 2017). Este fato indica o quão crucial é a necessidade de identificar e monitorar os alunos com baixo desempenho. Assim, é possível realizar intervenções precoces para os alunos cujo desempenho provavelmente não atenderá aos critérios de graduação, por exemplo, a formação no prazo estipulado. Para atingir o sucesso acadêmico, são diversos os fatores que contribuem para esse objetivo, como seu histórico, *status* econômico, características familiares, características da instituição anterior, entre outros (ACHARYA; SINHA, 2014). As técnicas de EDM contemplam a exploração dos dados de sua respectiva área de abrangência, de modo a melhorar o processo de aprendizagem, a qualidade do ensino e das decisões gerenciais educacionais (JAIN et al., 2017).

2.2 Evasão

Segundo o estudo realizado por ANDIFES, ABRUEM e SESu/MEC (1996), a qualidade do ensino, não é contemplada apenas pelas taxas de evasão, mas igualmente com o índice de diplomação e de matriculado, como este trabalho busca diferenciar. Os autores também caracterizam a evasão em três grupos:

> "**Evasão de curso**: quando o estudante desliga-se do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional.
> **Evasão da instituição**: quando o estudante desliga-se da instituição da qual está matriculado.
> **Evasão do sistema**: quando o estudante abandona de forma definitiva ou temporária o ensino superior."

Posto isso, a abordagem utilizada neste trabalho trata-se de um misto desses tipos de evasão, já que não foi possível conhecer previamente toda a motivação dos alunos, já

evadidos. O estudo foca na construção de um modelo de classificação único, que inclua todos os cursos do ICEA.

2.3 Trabalhos correlatados

As principais técnicas utilizadas no EDM são: clusterização, padrão sequencial, predição, regras de associação e classificação (MOHAMAD; TASIR, 2013). Entretanto, o escopo de utilização desses dados se difere no tipo de abordagem. Alguns estudos apontam para aplicações voltadas para o *e-learning*, outros para dados do histórico escolar e outros baseados em formulários de pesquisa. Logo, os dados são utilizados de forma exclusiva, pois, em cada instituto existem problemas e conclusões distintas. Isto é, não existe uma regra fixa, por isso, muito se debate sobre as técnicas utilizadas. Desse modo, é possível aplicar a metodologia aos inúmeros contextos do EDM.

Um estudo realizado por Caldeira (2021) utilizou da técnica de clusterização com a aplicação dos algoritmos *K-means*, *Nearest Neighbors* e o método de *Principal Component Analysis* (PCA) para caracterizar o problema da evasão no ICEA. Dessa maneira, foi possível extrair características semelhantes entre os alunos, no qual se destaca a formação de grupos com a propriedade marcante de baixo rendimento escolar.

Já o estudo realizado por Flores, Heras e Julian (2022) investe nos algoritmos de classificação para distinguir os alunos com mais propensão a evasão da Universidade de Moquegua, no Peru. Foram utilizados algoritmos como *Decision Tree*, *Naive Bayes*, J48, *OneR* e seus decorrentes. Foi apresentado como base os dados referentes ao ingresso do discente, média semestral, dentre outros. Wen e Wen (2020), analisou o problema de forma similar, utilizaram algoritmos de classificação e regressão para o contexto de *e-learning*. O estudo possui como fonte de dados plataformas *online*, como Coursera e *Udacity*. Entre os dados coletados estão os *logs*, referentes às páginas acessadas, quantidade de cliques e tempo de uso das plataformas.

Com o mesmo objetivo de prever a possibilidade da evasão universitária, Viloria et al. (2020) emprega os algoritmos de classificação para categorizar os alunos em situação de evadido ou matriculado. O trabalho foca em condensar as informações referente ao primeiro ano escolar, geolocalização das residências dos alunos, dados referentes à escola anterior e quantidade de aprovações e reprovações. Os algoritmos utilizados foram J48, *Bayes Net* e *OneR*, que atingiram acurácia de 79.8%, 77.9% e 75.8%, respectivamente.

2.4 Considerações finais

A importância de mapear os indivíduos propensos à evasão inicia-se na fundamentação teórica de ANDIFES, ABRUEM e SESu/MEC (1996) até o recente trabalho

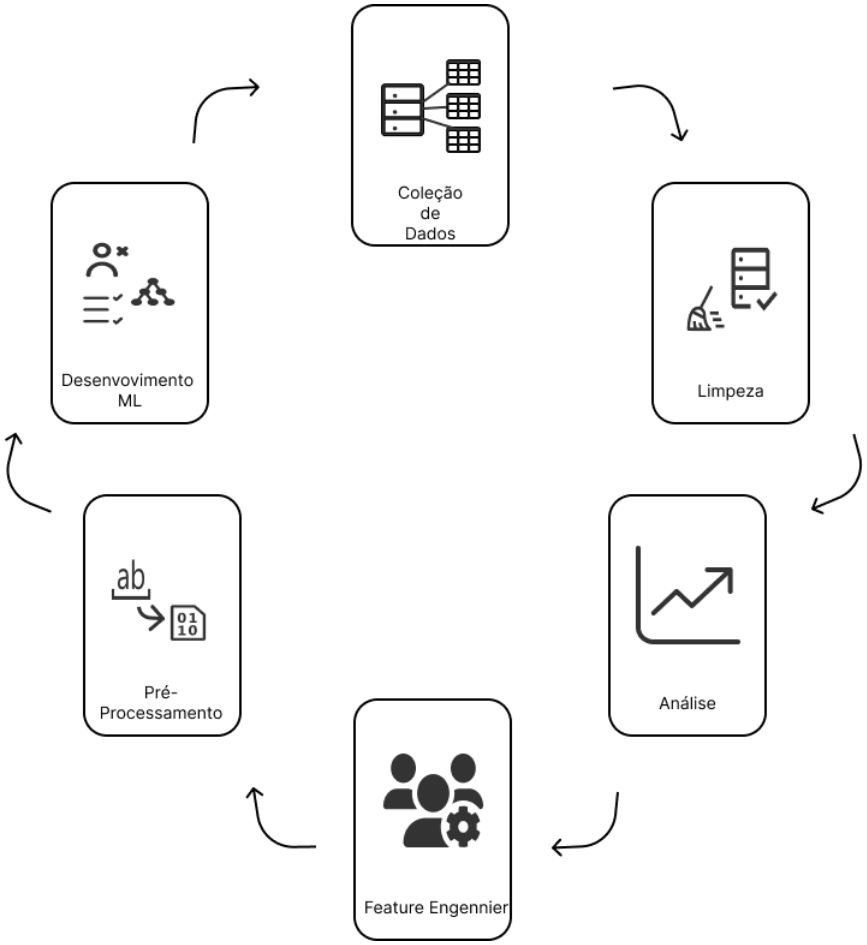
de (MUSSLINER et al., 2021). É um tema estudado há anos, que passou por diversas transformações e tem ganhado cada vez mais adeptos. Prontamente, este trabalho faz uma contribuição a esse extenso problema, relacionando o uso de AI para desenvolvimento de um modelo matemático, capaz de identificar precocemente os alunos que possuem maiores chances de evadir, bem como conhecer as características que mais impactam para a distinção desses alunos. A caracterização antecipada desses discentes, fundamenta a criação de ações de nível individual ou coletiva, de modo a combater esse tipo de prática no Instituto.

3 Desenvolvimento

Este capítulo apresenta o desenvolvimento do trabalho, iniciando pela configuração do ambiente, a definição do *workflow* do projeto, seguido do detalhamento de cada etapa dos processos desenvolvidos. O *workflow* apresenta de forma resumida as principais etapas realizadas neste trabalho, desde a etapa de fonte dos dados até o desenvolvimento e análise dos resultados.

A seguir, na Figura 6 é apresento a sistematização dos principais passos efetuados para desenvolver o projeto, os quais são detalhados nas próximas seções.

Figura 6 – *Workflow* do projeto



Fonte: Autora do Trabalho

3.1 Ambiente de Desenvolvimento

A Linguagem de Programação *Pyhton* é muito utilizada no contexto de AI por se tratar de ser uma linguagem de alto nível e possuir diversas bibliotecas e *frameworks* para processamento e análise de dados. Além disso, existem algoritmos de ML já implementados e disponíveis para uso. Por essa razão, foi a linguagem escolhida para o desenvolvimento deste trabalho.

O *Google Collaboratory* é uma ferramenta disponibilizada pelo *Google*, no qual é possível executar códigos em *Python* no ambiente em nuvem. Além da facilidade em desenvolver um processamento dinâmico e eficiente, é possível realizar o compartilhamento de código. Logo, por esses aspectos este ambiente foi escolhido para desenvolvimento deste trabalho.

3.2 Análise dos Dados

A etapa de análise dos dados consiste na exploração da base de dados, o objetivo é verificar algum padrão ou existência de tendência nos dados, bem como o levantamento de insumos estatísticos. Um exemplo prático desta fase é a quantificação dos alunos diplomados no instituto, dentre outros resumos estatísticos que circundam as análises apresentadas no Capítulo 1. É nesta etapa que começa a ser construída uma lógica para entender como os dados estão relacionados e aplicados ao cenário estudado. A partir do passo de investigação, é possível a tomada de decisão assertiva em relação à base ou obter *insights* capazes de direcionar o estudo.

3.3 Coleção dos Dados

Os dados são utilizados como principal insumo para a tomada de decisão, por isso é importante definir a fonte de referência, e o que será utilizado para realizar a análise e a modelagem do problema. Analisar a estrutura dos dados, a existência ou não de relacionamento entre eles, os tipos de variáveis que estão representadas, é fundamental para a aplicação das técnicas de AI.

3.3.1 Fonte dos dados

Os dados, bem como sua qualidade, são os principais insumos para desenvolver Modelos de ML eficientes para resolver um problema. Para que este trabalho fosse possível, a Seção de Ensino do instituto disponibilizou duas bases de dados em formato *Comma-separated Values* (CSV). Nessas bases existem informações sensíveis, como o Cadastro de Pessoa Física (CPF) e a matrícula do aluno, possibilitando a identificação do aluno.

Entretanto, em consonância à Lei Geral de Proteção de Dados (LGPD) de modo a preservar a identidade e a ética do trabalho, esses registros foram anonimizados, de modo a garantir a segurança e a privacidade dos discentes.

3.3.2 Informações Gerais

As tabelas recebidas foram nomeadas de maneira geral como uma base referindo-se às `notas`, e outra referindo-se aos `registros` dos alunos. A base de dados notas contém os dados relativos ao histórico escolar dos alunos, enquanto a base de dados registo inclui elementos a respeito da admissão do aluno. A Tabela 1 corresponde às características gerais das bases de dados, em que se encontram os registros de todos os cursos ofertados no *campus*. A data mínima de ingresso, se difere entre as bases, pois, alguns dados foram salvos em arquivos físicos, nos quais este trabalho não considerou. Este aspecto e a diferença entre as datas de implementação dos diferentes cursos no ICEA são tratados na etapa de Limpeza (Seção 3.4).

Tabela 1 – Características Gerais.

Base de Dados	**Quantidade de Matrículas**	**Quantidade de Variáveis**	**Data Mínima de Ingresso**
registros	4.991	52	28/03/2005
notas	4.920	20	12/02/2009

Fonte: Autora do Trabalho

3.3.3 Tipo das Variáveis

As variáveis contidas na base são definidas entre categóricas e numéricas. As variáveis categóricas são do tipo *string*, as quais representam os valores de maneira qualitativa. Já as variáveis numéricas, pertencem ao tipo *float* ou *int*, e possuem origem quantitativa. O tipo de dado *datetime*, é utilizado para registros que contém valores referente a uma data, ou há uma data e hora quando salvo o horário da transação. É apresentado na Tabela 2 e Tabela 3, uma representação das variáveis contidas na base de registros e na base de notas, respectivamente.

Tabela 2 – Variáveis de Registro

Variáveis	**Tipo**
data ingresso, data da evasão, primeiro horário, ultimo horário	*datetime*
aluno computado censo, cidade nascimento, código curso, código curso admissão, código enfase, código habilitação admissão, código modo admissão, código situação aluno, código evasão tipo, curso, descrição, descrição modo admissão, descrição situação aluno, destino, estado nascimento, feito por requerimento, modalidade concorrência, modalidade concorrência homologada, motivo, origem, pais nascimento, participou politica afirmativa, republica, sexo, situação aluno atual, tipo republica, turno, usou politica afirmativa	categórica
ano de admissão, ano diplomação, ano evasão, ano de nascimento, caixa arquivo, carga horaria cursada, carga horaria curso, classificação vestibular, código currículo, código currículo admissão, código habilitação, código polo, código polo admissão, CPF id, matrícula id, polo, pontuação vestibular, semestre de admissão, semestre diplomação, semestre evasão	numérica

Fonte: Autora do Trabalho

Tabela 3 – Variáveis de Nota

Variáveis	**Tipo**
gravação	*datetime*
caráter, código curso, código departamento, código disciplina, cor da pele, descrição, descrição modo admissão, sexo, situação, tipo escola	categórica
ano, ano nascimento, código turma, exame especial, faltas, matrícula id, media final, semestre	numérica

Fonte: Autora do Trabalho

3.4 Limpeza

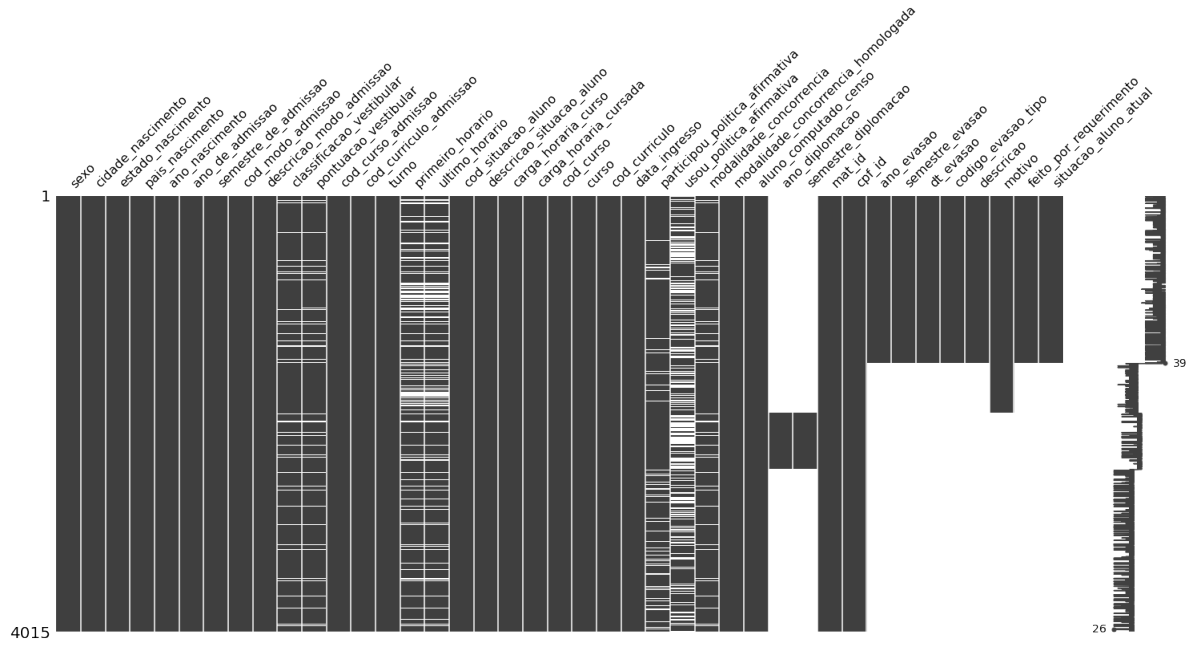
Nesta etapa é realizada a inspeção dos dados ausentes, das características irrelevantes presentes na base de dados, ou até mesmo as que não agregam nenhum tipo de informação. Nessa etapa também pode ser verificado os registros que são conhecidos como *outliers*, que fogem do padrão da base, sendo identificados como uma anomalia. Logo, essa etapa visa eliminar os dados inválidos, para gerar a informação e conhecimento de forma objetiva, sem interferências.

3.4.1 Dados Ausentes

Os valores ausentes em uma base de dados podem representar algum indício de erro como, por exemplo, anomalia na coleta dos dados, o não conhecimento de fato daquele valor ao preencher um formulário. Pode ocorrer também situações em que o dado está descrito por outra variável, ou simplesmente ser a consequência de aplicação da regra de negócio. A avaliação e a tratativa dessa situação são importantes para identificar *outliers* e deixar a base preparada para análise sem um possível viés advindo dos dados ausentes. A Figura 7 e a Figura 8 apresentam um *overview* dos dados ausentes da base de dados,

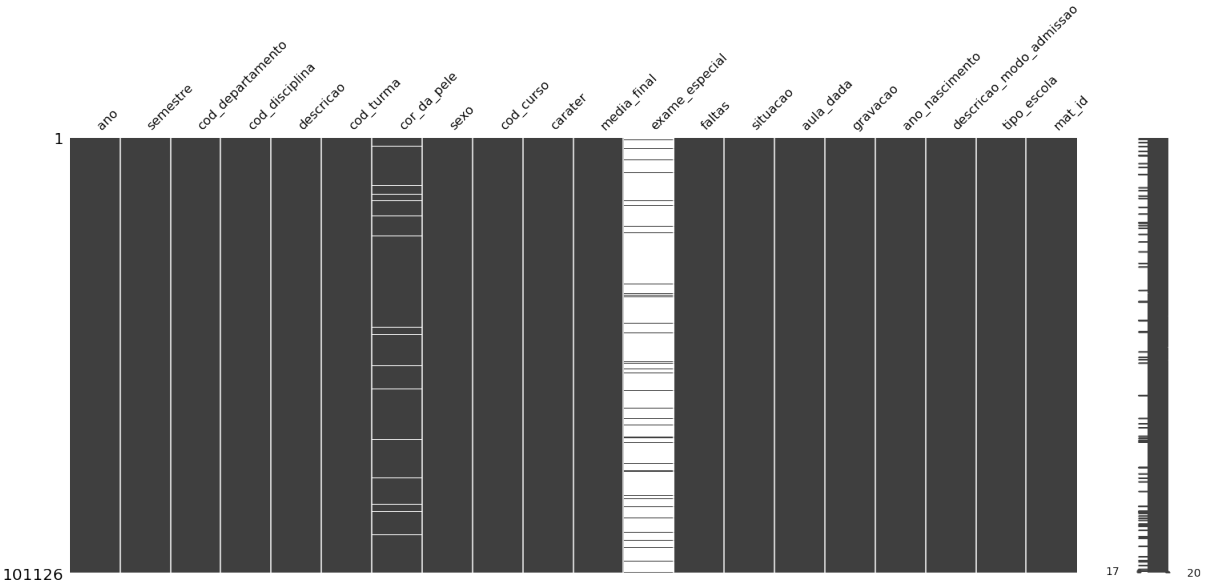
em que é possível visualizar a ausência desses valores na cor branca, presente em sua respectiva coluna.

Figura 7 – Dados ausentes: registros



Fonte: Autora do Trabalho

Figura 8 – Dados ausentes: notas



Fonte: Autora do Trabalho

Posto isso, a representatividade de alguns atributos que apresentam valores *null* superiores a 90% foram desconsiderados do trabalho. São eles: república, tipo de república, código de habilitação admissão, código do polo admissão, origem, código de habilitação, código de enfase, código de polo, polo, caixa de arquivo e destino, conforme apresentado na Figura 7. Ademais, a coluna exame especial, abordada na Figura 8, também apresentou anomalia em mais de 90% dos dados. Entretanto, para este caso foi investigado e constatado que os dados ausentes indicavam a ausência de realização de exame especial. Caso fosse excluído essa coluna, os valores referentes aos alunos que fizeram o exame especial, seriam perdidos. Por este fato, a coluna não foi excluída e sim tratada na etapa pertinente.

3.4.2 Definição do intervalo de tempo estudado

Algumas inconsistências, como a discrepância de data de início dos cursos, o uso do Exame Nacional do Ensino Médio (ENEM) e do vestibular como modo de admissão mencionado na Seção 3.3.2, são tratadas nesta etapa. Para atuar de forma igualitária tais aspectos, foi aplicado um filtro que deixa na base de dados somente os alunos que ingressaram a partir de tal adesão, ou seja, somente ingressantes a partir do primeiro período do ano de 2011. Essa escolha foi feita devido o fato de na tabela de registros conter características do desempenho obtido do aluno no ENEM, que complementa a base com características assertivas e relevantes do estado anterior do aluno em relação ao ingresso na Universidade.

Os dados foram coletados no primeiro semestre de 2021, entretanto, eles não refletem esse período. A incoerência entre o calendário acadêmico e o calendário corrente por consequência da COVID-19, dito isso, o último semestre letivo referente à base de dados foi o do segundo semestre de 2020. Assim sendo, o segundo semestre de 2020 foi considerado como limite superior do intervalo de tempo, e aplicado como filtro em ambas as bases de dados.

3.4.3 Desbalanceamento das classes

Neste trabalho, uma classe ou variável *target* é definida como a categoria que um aluno pertence. Essas categorias podem assumir os respectivos valores:

- Matriculado: caracteriza o aluno que está regularmente matriculado no curso.
- Diplomado: caracteriza o aluno que já concluiu o respectivo curso.
- Evadido: caracteriza o aluno que evadiu do curso.
- Trancado: caracteriza o aluno que está com o período do curso trancado.
- Afastado: caracteriza o aluno que está afastado do curso.

- Mobilidade: caracteriza o aluno que está em processo de mobilidade acadêmica.

A Tabela 4 apresenta uma sumarização dessas categorias. O não balanceamento das classes em AI pode prejudicar a identificação do padrão de uma determinada classe minoritária (CASTRO; BRAGA, 2011). Logo, é verificado esse desbalanceamento, ou seja, a proporção das classes não estão distribuídas de forma similar. Existem alguns métodos para fazer tratativas em classes desbalanceadas, como *Undersampling*, *Oversampling* e *Synthetic Minority Oversampling Technique* (SMOTE). Essas técnicas, embora eficazes, resultam na perda de informação e de explicabilidade do problema real, uma vez que ambas fazem a introdução de dados sintéticos. Todavia, para aproximar do contexto aplicado e por se tratar de uma representatividade inferior a 1 %, não foram considerados neste trabalho as situações de *trancado*, *afastado* e de *mobilidade*.

Tabela 4 – Proporção da Classes do Problema.

Classe	**Quantidade**	**Porcentagem**
evadido	2002	49.86 %
matriculado	1463	36.43 %
diplomado	518	12.90 %
trancado	1463	0.57 %
afastado	7	0.17 %
mobilidade	2	0.04 %

Fonte: Autora do Trabalho

3.4.4 Resultante da Etapa de Limpeza dos Dados

Na etapa de Limpeza dos Dados foram realizados procedimentos de limpeza em escala geral, no qual foi retirado os dados que poderiam gerar algum conflito nas próximas etapas e que não agregariam valor à análise de dados e ao Modelo. Após as principais operações de remoção de colunas com maior proporção de dados ausentes, definição do tempo de coleta dos dados e a escolha das classes que serão representadas, são apresentadas na Tabela 5 e na Tabela 6 sínteses da organização dos dados para a próxima etapa do processo.

Tabela 5 – Características Gerais Após Limpeza.

Base de Dados	**Quantidade de Matrículas**	**Quantidade de Variáveis**	**Data Mínima de Ingresso**
registro	3.654	41	03/01/2011
nota	3.654	20	03/01/2011

Fonte: Autora do Trabalho

Tabela 6 – Proporção da Classes do Problema Após Limpeza.

Classe	**Quantidade**	**Porcentagem**
evadido	1.877	51.36 %
matriculado	1.259	34.45 %
diplomado	518	14.17 %

Fonte: Autora do Trabalho

3.5 *Feature Engineering*

Segundo Duboue (2020), na etapa de *Feature Engineering* os dados são explorados de forma bruta, de modo a extrair o máximo de conhecimento possível, com o objetivo de servir de entrada para os algoritmos de ML. Este estágio é conhecido por sua complexidade e extensão de tempo, dedicados ao estudo das variáveis e a contextualização do problema.

Uma das partes deste processo é a criação e extração de novas características da base de dados, no qual se realizam operações entre os dados já conhecidos, ou em conjunto a uma, ou mais base de dados externa. Este processo é similar ao *Knowledge-Discovery in Databases* (KDD) que busca extrair conhecimento em uma base de dados. Todavia os processos diferem-se, ao fato que a *Feature Engineering* possui foco não só na descoberta, mas igualmente na modelagem gerando novos atributos, além de contemplar o pré-processamento dos dados.

3.5.1 Tratamento de Dados Ausentes

Após realizar a etapa de Dados Ausentes 3.4, cujo objetivo foi fazer um filtro inicial para reduzir o problema dos dados ausentes, ainda assim, algumas *features* tiveram a necessidade de se aplicar um processo transformação para não deixar esses valores desconhecidos. Após análise dos dados em questão, chegou-se na seguinte solução para cada problema, conforme apresentado na Tabela 7.

Tabela 7 – *Map* dos dados ausentes.

Base de Dados	***Feature***	**Motivação**	**Decisão**
registro	classificação vestibular, pontuação vestibular	os dados ausentes indicam que o aluno não ingressou por meio do ENEM, ele veio de transferência, ou decisão judicial	*input* com o valor zero para os ausentes
registro	cidade de nascimento	alunos que estão sem local de nascimento mas com endereço de residência em João Monlevade	assume-se que a cidade de nascimento possui o valor de "João Monlevade"
registro	tipo de escola	não foi possível identificar o porquê do *null*	*input* 0, se aluno era ingressante de escola pública, 1 se era ingressante de escola particular, se não 2

Fonte: Autora do Trabalho

3.5.2 Novas Características

A produção de novas características a partir da base de dados tem como objetivo de expandir o conhecimento originário dos dados brutos e aumentar a diversidade de dados, que são insumos para os algoritmos de ML. Assim, é possível aprimorar a qualidade e desempenho da predição.

A partir dessa etapa, as duas bases tanto a de registro quanto à de notas, são relacionas através de um *join* para a realização de alguns cálculos. Os atributos gerados a partir das características apresentadas na Tabela 2 e na Tabela 3 sofrem algumas alterações. A Tabela 8 e a Tabela 9 apresentam as descrições das operações realizadas. As manipulações dessas características foram baseadas na prévia análise dos dados e de acordo com a viabilidade das regras de negócio.

Tabela 8 – Geração de Novos Atributos.

***Feature* Gerada**	**Descrição**
idade	idade do aluno no momento da matrícula
porcentagem cursado	porcentagem que o aluno cursou até o presente momento
diferença de nota	diferença da nota do aluno por disciplina, e o necessário para passar, se o resultado for negativo, significa que faltou nota, se for positivo, significa que sobrou nota
nota final	se o aluno fez o exame especial, a nota final é a do exame especial, se não, é o valor refente à média final
período que cursou	refere ao período que o aluno cursou a disciplina, se foi no primeiro período dele na faculdade, no segundo...
flag situação	se o aluno está no tempo de 10 ou 8 períodos, ou não
período máximo	o período máximo que o aluno está ou esteve matriculado
situação agrupada	situação do aluno na disciplina, foram agregados todas os valores em que houve reprovação, para uma categoria de reprovado, as situações de aprovado, cancelado e trancado se mantiveram as mesmas.
coeficiente semestral	é a média de nota do aluno por semestre, sem considerar as situações de trancado e cancelado
coeficiente semestral acumulado	a média do coeficiente semestral atual com o semestre anterior
quantidade de disciplinas aprovadas	quantidade de disciplinas aprovadas no período
quantidade de disciplinas reprovadas	quantidade de disciplinas reprovadas no período
quantidade de disciplinas canceladas	quantidade de disciplinas canceladas no período
quantidade de disciplinas trancadas	quantidade de disciplinas trancadas no período
quantidade de disciplinas do DEELT	quantidade de disciplinas matriculadas do Departamento de Engenharia Elétrica (DEELT) no período
quantidade de disciplinas do DECEA	quantidade de disciplinas matriculadas do Departamento de Ciências Exatas e Aplicadas (DECEA) no período
quantidade de disciplinas do DECSI	quantidade de disciplinas matriculadas do Departamento de Computação e Sistemas (DECSI) no período
quantidade de disciplinas do DEENP	quantidade de disciplinas matriculadas do Departamento de Engenharia de Produção (DEENP) no período
quantidade de disciplinas eletivas	quantidade de disciplinas eletivas matriculadas no período
quantidade de disciplinas obrigatórias	quantidade de disciplinas obrigatórias matriculadas no período
quantidade de disciplinas facultativas	quantidade de disciplinas facultativas matriculadas no período
desempenho nos semestres iniciais	média da nota dos alnos nos anos 4 períodos iniciais
tag evasão	caracterizando o motivo que dos alunos que evadiram
região próxima	caracteriza de o aluno reside em algum município próximo à cidade de João Monlevade
curso sjm	1, se o curso for Sistemas de Informação, se não 0
curso cjm	1, se o curso for Engenharia da Computação, se não 0
curso ejm	1, se o curso for Engenharia Elétrica, se não 0
curso pjm	1, se o curso for Engenharia de Produção, se não 0

Fonte: Autora do Trabalho

Tabela 9 – Tratamento das *Features*.

Feature	**Tratamento**
exame especial	os valores de exame especial eram representados pela nota do aluno que realizou o exame especial, então essa variável foi transformada para a forma binária, onde o valor zero refere-se ao estado de não ter realizado o exame especial, e o valor de um, refere-se à realização do exame especial

Fonte: Autora do Trabalho

3.5.3 Pré-Processamento

Alguns algoritmos são capazes de trabalhar com variáveis categóricas, entretanto, a maioria dos algoritmos utiliza variáveis numéricas para a realização de cálculos e para melhorar o desempenho. Assim sendo, as variáveis categóricas deste problema foram transformadas para representação numérica, conforme descrito na Tabela 10 e na Tabela 11.

Tabela 10 – *One-Hot Encoding*.

Feature	**Codificação**
tag online	1, se a matéria foi realizada online, se não, 0
turno	1, se a matéria foi realizada no turno vespertino, 0 se a matéria foi realizada no turno noturno
flag situação	1, se o aluno está no período referente à correspondida ao curso, se não, 0
turno	1, se a matéria foi realizada no turno vespertino, 0 se a matéria foi realizada no turno noturno
turno	1, se a matéria foi realizada no turno vespertino, 0 se a matéria foi realizada no turno noturno
modo admissão duplo diploma	1, se o aluno possui competência de realizar a dupla diplomação conveniada com o exterior, se não 0
modo admissão programa PEC-G	1, se o aluno veio de outro país por meio do convênio, se não 0
modo admissão decisão judicial	1 se o aluno ingressou por meio de ação judicial, se não 0
modo admissão portador de diploma de graduação	1, se o aluno já possui um diploma de graduação, se não 0
modo admissão transferência externa	1, se o aluno ingressou via transferência, se não 0

Fonte: Autora do Trabalho

Tabela 11 – *One Hot Encoding* - Modo de Admissão.

Feature	**Codificação**
modalidade concorrência AC	1, se o aluno entrou via vestibular usando a categoria ampla concorrência, se não 0
modalidade concorrência L1	1, se o aluno entrou via vestibular usando a categoria L1: "Candidatos com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo que tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência L2	1, se o aluno entrou via vestibular usando a categoria L2: "Candidatos autodeclarados negros (pretos ou pardos) ou indígenas, com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência L5	1, se o aluno entrou via vestibular usando a categoria L5: "Candidatos que, independentemente da renda, tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência L6	1, se o aluno entrou via vestibular usando a categoria L6: "Candidatos autodeclarados negros(pretos ou pardos) ou indígenas que, independentemente da renda, tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência L9	1, se o aluno entrou via vestibular usando a categoria L9: "Candidatos com deficiência que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência L10	1, se o aluno entrou via vestibular usando a categoria L10: "Candidatos com deficiência autodeclarados negros (pretos ou pardos) ou indígenas, que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência L14	1, se o aluno entrou via vestibular usando a categoria L14: "Candidatos com deficiência autodeclarados negros (pretos ou pardos) ou indígenas que, independentemente da renda tenham cursado integralmente o ensino médio em escolas públicas." (UFOP, 2022) , se não 0
modalidade de concorrência PAA	1, se o aluno entrou via vestibular usando uma política de ação afirmativa, mas que não foi especificada, se não 0

Fonte: Autora do Trabalho

3.5.4 Modelagem

Definir a modelagem do problema é fazer a escolha de quais atributos estão presentes como dados de *input* para os algoritmos, bem como será a agregação desses valores. Para este trabalho, a modelagem foi desenvolvida de modo a agregar os valores e o conhecimento adquirido até o momento na visão de matrícula. Ou seja, cada matrícula terá apenas uma entrada de representação nos algoritmos considerando as métricas da sua performance acadêmica. Essas métricas foram desenvolvidas a partir da associação entre matrícula e período. Dessa maneira, as informações foram substituídas pelo valor médio referente à cada variável do tipo numérico. Os valores que não possuíam a necessidade dessa agregação, os atributos categóricos, foram mantidos conforme o pré-processamento. Veja a seguir, na Tabela 12, Tabela 13 e Tabela 14 um exemplo fictício de como foram realizadas essas associações.

Tabela 12 – Dados sem Agregação

Matrícula	**Período que fez a Disciplina**	**Nota**	**Situação**	**Quantidade de Disciplinas Aprovadas**	**Quantidade de Disciplinas Reprovadas**	**período máximo**
A	1	60	aprovado	1	0	3
A	1	78	aprovado	1	0	3
A	1	54	reprovado	0	1	3
A	2	32	reprovado	0	1	3
A	2	60	aprovado	1	0	3
A	3	62	aprovado	1	0	3

Fonte: Autora do Trabalho

Tabela 13 – Agregação por Período e Matrícula

Matrícula	**Período que fez a Disciplina**	**Média de Notas**	**Média de Disciplinas Aprovadas**	**Média de Disciplinas Reprovadas**	**Período Máximo**
A	1	64	0,66	0,33	3
A	2	46	0,5	0,5	3
A	3	62	1	0	3

Fonte: Autora do Trabalho

Tabela 14 – Agregação por Matrícula.

Matrícula	**Média de Notas**	**Média de Disciplinas Aprovadas**	**Média de Disciplinas Reprovadas**	**período máximo**
A	57,33	0,72	0,27	3

Fonte: Autora do Trabalho

Como pode ser observado no exemplo, algumas *features* serviram apenas de base para as agregações, ou para realização de alguma operação, por isso, ao final desse processo

elas foram removidas. Ademais, as características de sexo e de cor da pele, foram removidas para não enviesar o Modelo de predição a respeito dessas características sensíveis. Na Tabela 15, é apresentado o resultado da composição das variáveis.

Tabela 15 – Resultado da Modelagem dos Dados.

***Features* Excluídas**	***Features* para o Modelo**
aluno computado censo, ano, ano de admissão, ano diplomacao, ano evasao, ano nascimento, caráter, cidade nascimento, cod curriculo, cod curriculo admissão, cod curso, cod curso admissão, cod departamento, cod disciplina, cod modo admissão, cod situação aluno, cod turma, codigo evasao tipo, cor da pele, cpf id, curso, data, ingresso, descrição, dt evasao, estado nascimento, exame especial, feito por requerimento, gravacao, modalidade concorrência, modalidade concorrência homologada, motivo, n semestres total curso, pais nascimento,participou, politica afirmativa, período que cursou, período que entrou, período que fez a disciplina, primeiro horario, semestre, semestre de admissão, semestre diplomacao, semestre evasao, sexo, situação, situação agrupada, situação aluno atual, ultimo horario, usou politica afirmativa	curso cjm, curso ejm, curso pjm, curso sjm, classificação vestibular, descrição modo admissão convênio duplo diploma, descrição modo admissão convênio programa pec-g, descrição modo admissão decisão judicial, descrição modo admissão portador de diploma de graduação, descrição modo admissão transferência externa, descrição situação aluno, desempenho primeiros períodos, flag situação, idade que entrou, max n períodos, média caráter eletiva, média caráter facultativa, média caráter obrigatória, média coeficiente acumulado, média coeficiente semestral, média de disciplinas ap no período, média de disciplinas ca no período, média de disciplinas rp no período, média de disciplinas tr no período, média departamento decea, média departamento decsi, média departamento deelt, média departamento deenp, média diff nota, média exame especial, média faltas, média nota final, modalidade concorrência ac, modalidade concorrência l1, modalidade concorrência l2, modalidade concorrência l5, modalidade concorrência l6, modalidade concorrência paa, pct cursado, pontuação vestibular, região proxima, tag motivo evasão aprovacao outra instituição, tag motivo evasão defasagem ensino, tag motivo evasão deseja outro curso, tag motivo evasão familiares e pessoais, tag motivo evasão interno ufop, tag motivo evasão localidade, tag motivo evasão outro, tag motivo evasão saúde, tag motivo evasão trabalho, tipo escola, turno

Fonte: Autora do Trabalho

3.5.5 Normalização dos Dados

Para complementar os processos anteriores, foi realizado a etapa de normalização dos dados. Este processo consiste em transformar os valores dos registros em um intervalo de zero e um. Essa técnica tem como objetivo evitar que a grandeza de alguns atributos enviesem o Modelo de ML, devido à disposição dos dados no espaço. Desse modo, é possível reduzir o erro do Modelo e acelerar a fase de treinamento. Esses aspectos são explicados de forma detalhada nas próximas seções.

3.6 Aprendizado de Máquina

O termo *Machine Learning* (ML) é caracterizado ao longo da história por ser um processo que utiliza algoritmos capazes de reconhecer padrões em um conjunto de dados, sem ter sido programado para isso. O Aprendizado de Máquina se divide em três principais

áreas: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço (KONAR, 1999).

O aprendizado supervisionado é aquele no qual o sistema alimentado já possui uma saída conhecida. Essa saída é chamada de variável alvo ou *target*. Por exemplo, em um contexto de detecção de doenças, a variável de saída é se o paciente está doente ou não. Esses algoritmos, quando é apresentado a um novo conjunto de dados, tenta prever o rótulo de saída, com base nas caraterísticas presentes nos grupos previamente conhecidos. Como exemplo de aplicação, existe a previsão dos valores de ações, detecção de doenças, classificação de imagens, dentre outros.

Para os problemas de aprendizado não supervisionado, o sistema não conhece os valores da variável alvo, e dessa maneira, através das características do conjunto de dados, é possível realizar uma tarefa de agrupamento, ou de associação baseado somente nas características em comum dos possíveis grupos. Essa técnica é frequentemente utilizada para a recomendação de séries, filmes, disposição de produtos em prateleiras, entre outros. Já o aprendizado por reforço, amplamente utilizado em jogos e robótica, parte de um agente em um determinado ambiente, que é caracterizado por seus atributos e ações, cujo foco é atingir um determinado objetivo, como, por exemplo, ganhar algum jogo. Para desenvolver a atividade e mensurar o quão importante sua ação o deixou próximo do seu objetivo, ele trabalha por pontuação e penalização, sua recompensa será positiva caso ele obteve sucesso com a ação, e negativa caso tenha cometido um erro. Dessa maneira, ele pode reformular suas ações com o propósito de atingir a pontuação mais alta possível (SARKER, 2021).

Este trabalho utiliza as técnicas do aprendizado supervisionado, com ênfase nos algoritmos de classificação. Os algoritmos de classificação tem como objetivo classificar os dados de acordo com uma classe já determinada anteriormente. Neste trabalho, as classes conhecidas são decorrentes da variável alvo: situação do aluno. Essa variável indica a situação atual da matrícula do aluno, os possíveis valores assumidos são: matriculado, evadido ou diplomado.

3.6.1 Divisão dos Dados

O desenvolvimento de um Modelo de AI tem como seu insumo primário os dados já pré-processados, modelados e normalizados. Além disso, é necessário a divisão desses elementos em um conjunto de treino, que são utilizados para treinar os Modelos, ou seja, para construir a parte matemática e de reconhecimento do padrão em si, e um conjunto de teste. Esse conjunto é utilizado para testar o Modelo de fato, com dados que não foram apresentados para ele anteriormente. Além disso, pode ser utilizado um terceiro conjunto, chamado de validação, que contém um grupo de dados que será aplicado como uma segunda etapa de teste do Modelo, validando o desempenho do Modelo. O resultado

da divisão desses subconjuntos, é apresentado conforme a Tabela 16.

Tabela 16 – Divisão Treino, Teste e Validação.

Base	**Instâncias**	**% do Total**
Treino	2630	70 %
Teste	658	20 %
Validação	366	10 %

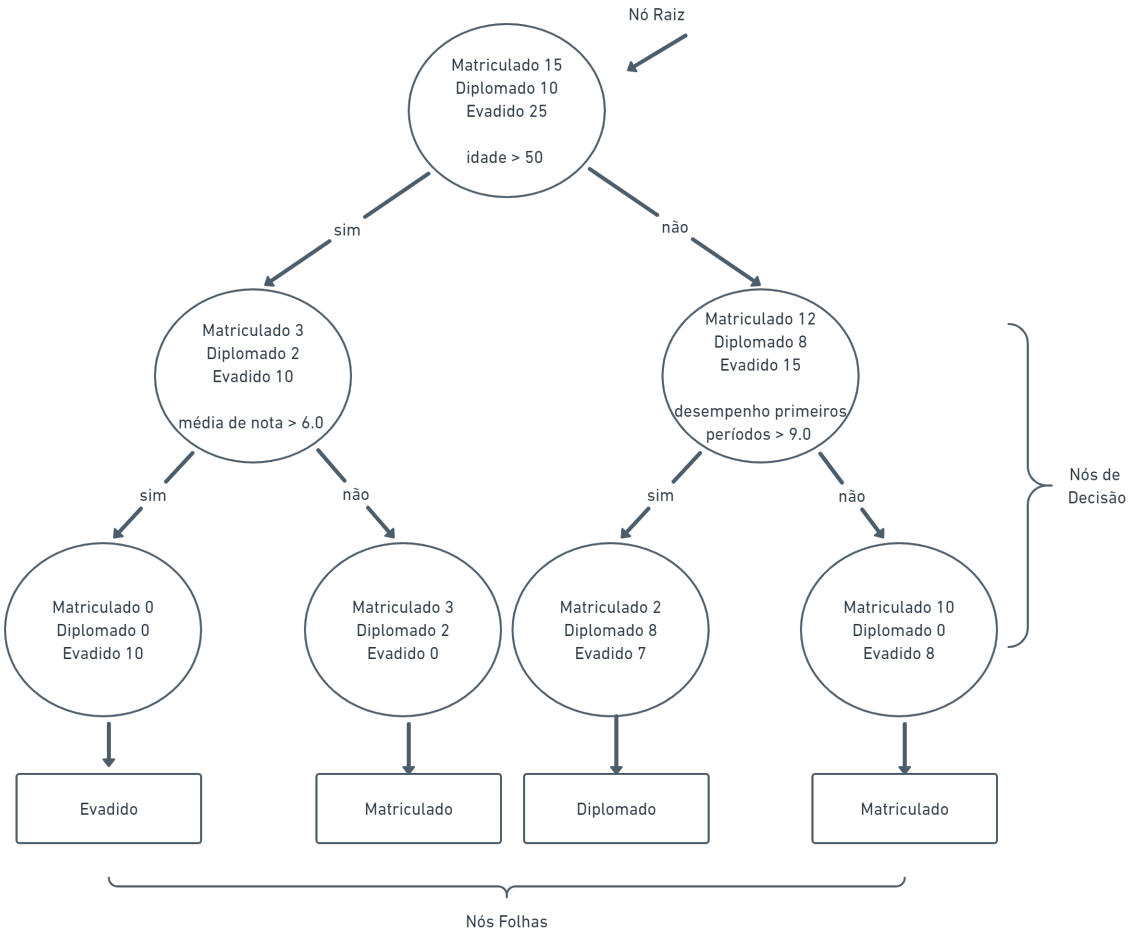
Fonte: Autora do Trabalho

3.6.2 *Random Forest*

Os algoritmos baseados em Árvores de Decisão são bastantes intuitivos e eficazes na resolução de problemas com mais de uma saída, ou seja, os que possuem mais de uma classe, além de sua facilidade de compreensão e execução. Uma Árvore de Decisão é formada pelo nó raiz, os ramos ou arestas, os nós de decisão, e os nós folhas. No nó raiz é concentrado a primeira decisão do problema, as arestas representam de fato a decisão tomada. A partir disso, é definido o caminho seguirá o fluxo dos dados que leva a um segundo nó, que pode ser de decisão, onde é tomada uma nova decisão, ou a um nó folha que representa a saída do problema, o resultado alvo (BREIMAN, 1987). A Figura 9 apresenta um exemplo de uma árvore de decisão.

No exemplo são apresentadas algumas características fictícias dos alunos, e definido pontos de corte para cada atributo. A partir disso, é tomada a decisão de qual ramo seguir no fluxo das informações. Por conseguinte, ao final, tem-se a classe de predição. A criação dos pontos de decisão e quais as características que serão utilizadas para a predição, bem como os pontos de corte, são definidos a partir do treinamento do Modelo. Treinar um Modelo de ML é apresentar os dados do conjunto de treino como entrada para tal algoritmo. No caso de algoritmos de classificação, o treinamento é realizado informando a classe alvo e, a partir disso, são realizados cálculos matemáticos, capazes de identificar um padrão no conjunto de dados, apto a maximizar a separação dos alunos entre os três grupos apresentados. Como, por exemplo, se seguido a partir do nó raiz, a seguinte sequência de ações: `sim -> sim`, tem-se que os atributos, idade, média de nota, define claramente a divisão dos alunos evadidos em relação aos matriculados e diplomados.

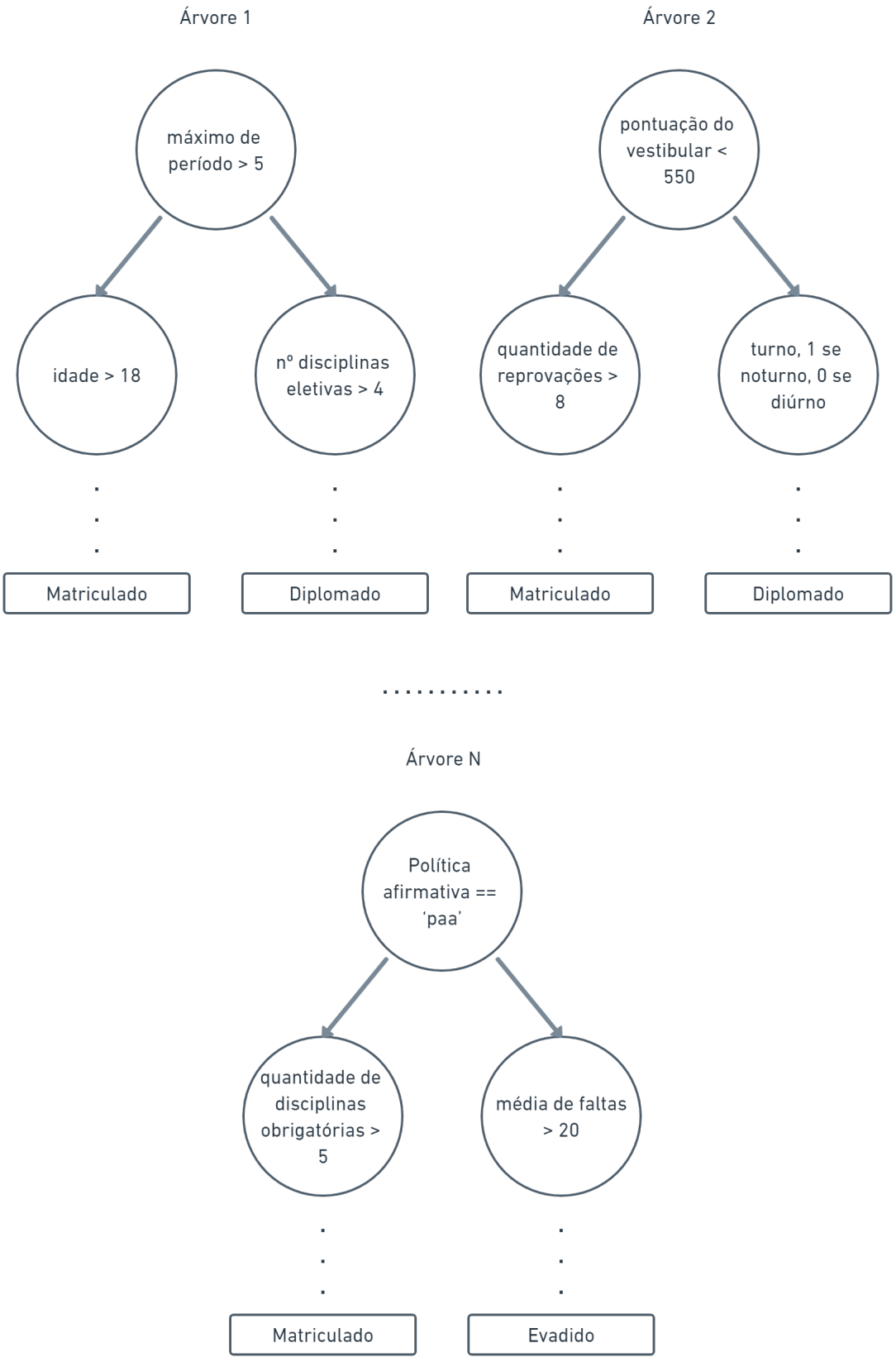
Figura 9 – Exemplo de Árvore de Decisão



Fonte: Autora do Trabalho

O *Random Forest* é um algoritmo que apresenta uma coleção de Árvores de Decisão, onde para cada árvore é destinado características distintas como nó raiz de forma aleatória, em sequência para os nós de decisão, até os resultados (BREIMAN, 2001). Desse modo, há uma maior variabilidade de características e de fluxo de dados apresentados, com diferentes saídas, como apresentado na Figura Figura 10.

Figura 10 – Exemplo de *Random Forest*



Fonte: Autora do Trabalho

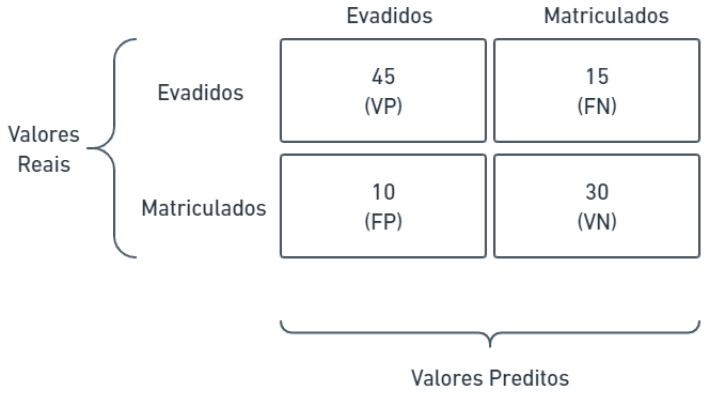
Dessa maneira, é possível estabelecer inúmeras regras para a tomada de decisão. Esses valores são combinados e apresentados os resultados com maior desempenho, isto é, os que conseguem separar melhor as classes alvo. Geralmente, a *Random Forest* possui uma previsão melhor, devido essa diversidade de resultados. Entretanto, devido à pluralidade, pode haver um sobre ajuste aos dados de treinamento, um estado chamado de *overfitting*, (BREIMAN, 2001). Este estado ocorre quando o algoritmo prevê muito bem as classes alvo do conjunto de treinamento, e quando apresentado aos dados ainda não vistos, os de teste e de validação, não consegue manter a eficiência fora do cenário controlado. Alguns parâmetros do próprio algoritmo podem ser regulados, como, por exemplo, o número de árvores geradas, grau de profundidade da árvore, ou seja, o quão profunda será o nível dos nós de decisão, entre outros. A identificação desses problemas e a gerência dos Modelos de ML se dá a partir das métricas de avaliação, que quantificam a eficiência dos Modelos. Esses aspectos são relatados na seção seguinte.

3.6.3 Métricas de Avaliação

As métricas de avaliação são utilizadas para medir o desempenho dos Modelos de ML. Para detalhar quais métricas são utilizadas no contexto de classificação, é essencial que seja apresentado os conceitos que circundam uma matriz de confusão (STEHMAN, 1997). A matriz de confusão é uma tabela que representa a quantidade de acertos e erros do Modelo.

Suponha uma base de dados com 100 alunos, 60 estão em situação de evasão, e 40 em situação de matriculado. A partir disso, é apresentado a um algoritmo de origem supervisionado, com o objetivo de prever a classe dos alunos evadidos. Desse modo, obtém-se o seguinte resultado:

Figura 11 – Exemplo de Matriz de Confusão



Fonte: Autora do Trabalho

As linhas da matriz de confusão são compostas pelos valores reais que cada classe apresenta originalmente, enquanto as colunas apresentam as classes previstas pelo algoritmo. Posto isso, têm-se as possíveis situações:

- Verdadeiro Positivo: ocorre quando o Modelo prevê que as classes que eram positivas (evadido), realmente foram previstas como positiva (evadido), isto é, 45 casos.
- Falso Negativo: ocorre quando o Modelo prevê que a classe é negativa (matriculado), e erra, pois, nos dados reais, ela é negativa (matriculado), isto é, 15 casos.
- Falso Positivo: ocorre quando o Modelo prevê que a classe é positiva (evadido), quando, na verdade, ela é negativa (matriculado), isto é, 10 casos.
- Verdadeiro Negativo: ocorre quando no conjunto real a classe negativa (matriculado), é prevista corretamente pelo Modelo, isto é, 30 casos.

Entendido isso, tem-se que o melhor resultado consiste em situações do tipo Verdadeiro Positivo e Verdadeiro Negativo, no qual os casos reais das duas classes são previstos de forma correta. Já os Falsos Negativos e os Falsos Positivos representam os casos de erro do Modelo. Uma das métricas mais utilizadas para validar os algoritmos de classificação é a Acurácia. Ela representa a quantidade de Verdadeiro Positivo mais o Verdadeiro Negativo, dividido pela quantidade total de casos, é dada pela seguinte expressão:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.1}$$

Para o exemplo acima, tem-se uma Acurácia de 75%, isto significa que o algoritmo é capaz de acertar aproximadamente 75% de todos os casos. A métrica *Precision* define o quão bom o algoritmo é para acertar os exemplos do tipo Verdadeiro Positivo. Essa métrica é definida pelo resultado que ela atingiu na classe alvo, Verdadeiro Positivo, dividido pela quantidade total de casos daquela classe específica, Verdadeiro Positivo mais Falso Positivo, dado pela seguinte expressão:

$$Precision = \frac{VP}{VP + FP} \tag{3.2}$$

Essa métrica busca responder à seguinte pergunta: qual a proporção dos dados classificados como evadidos que realmente eram evadidos? Seguindo o exemplo aplicado à matriz de confusão, tem-se uma precisão de 81.81% para a classe de evadido. Tem-se também a métrica de *recall*, que visa detectar a relação das classes corretas, com todas as previsões que realmente são positivas. É dada pelos valores de Verdadeiro Positivo dividido pela soma de casos Verdadeiro Positivo e o Falso Negativo, conforme a seguinte expressão:

$$Recall = \frac{VP}{VP + FN} \tag{3.3}$$

Essa métrica responde à seguinte questão: de todos os casos de alunos que realmente são evadidos, qual é o percentual identificado corretamente pelo Modelo? Posto isso, tem-se o valor de *recall* para a classe de evadido de 75%. Para findar, tem-se a *F1-score* que é uma média harmônica entre a *precision* e o *recall*, dessa maneira é possível analisar uma métrica ao invés de duas, de forma a balancear os valores do *precision* e *recall*. É dada pela seguinte expressão:

$$F1\text{-}score = 2 * \frac{precision + recall}{precision * recall} \tag{3.4}$$

3.7 Seleção de Variáveis

Ao desenvolver Modelos é importante se atentar à complexidade de características que são utilizadas para resolução de um problema. Quanto maior a dimensão de *features*, maior será o esforço de interpretabilidade do Modelo, tempo de execução e *hardware* específicos para o desenvolvimento (RAMASWAMI; BHASKARAN, 2009). Por isso, conhecer as características que mais impactam nos algoritmos de AI, permite reduzir a complexidade dos cálculos, além de utilizar somente os atributos que maximizam a separação das classes, assim é possível eliminar características irrelevantes que não melhoram o desempenho de um Modelo. Além disso, conhecer as *features* mais importantes para o Modelo, viabiliza a descoberta de *insights* ainda não observados no contexto de estudo, bem como confirmar alguma hipótese a respeito do problema. Existem algumas metodologias para a identificação e seleção dessas características, as quais são abordadas a seguir.

3.7.1 Correlação

A correlação entre as variáveis é caracterizada pela dependência ou associação entre duas variáveis sendo elas causal, na qual uma implica à outra, ou não causal, quando não ocorre essa implicação (MORETTIN; BUSSAB, 2004). Uma forma de calcular a correlação é através do coeficiente de *pearson*, uma função matemática que retorna valores ente 1 e -1. Nesse retorno, 1 significa uma dependência forte e positiva entre elas, ou seja, se uma possui seus valores crescentes, a outra também possui seus valores crescentes. Se o valor do coeficiente for 0, significa que as variáveis não possuem nenhuma dependência. E se o valor for de -1, significa uma dependência forte, entretanto, inversa negativa, ou seja, à medida que uma variável vai aumentando seu valor, a outra tende a diminuir.

O cálculo da correlação é importante para a construção de Modelos de ML, para remover algum viés e ruídos proveniente delas. Uma variável que possui uma dependência muito forte em relação à variável alvo, por sua vez, explica muito bem a separação das classes. Com isso, os outros atributos se tornam desnecessários ou pouco influentes. Por sua vez, dependendo das características do problema, pode inviabilizar a utilização de técnicas de AI. Outro aspecto discutido é a correlação entre as variáveis do problema. Em situações de correlação muito forte, isso indica que ambas as variáveis podem ser sobrepostas e capazes de causar o mesmo efeito no algoritmo (XU; MOON; SCHAAR, 2017). Neste caso, uma variável pode ser substituída pela outra, mantendo-se na base de dados apenas uma. Para este trabalho, adota-se um *treshold* de 0.7 para detectar relações fortes positivas, e -0.7 para detectar relações fortes negativas.

3.7.2 *Select k-best*

Uma forma de aplicar a seleção de *features* é através de testes estatísticos univariados, que permite a análise de cada caraterística de forma única. Dessa maneira, cada atributo da base de dados é analisado e comparado à distribuição dos dados em relação aos grupos da classe alvo. A função *Select k-best* da biblioteca *Sklearn* realiza o teste estatístico ANOVA para cada variável e retorna a importância dessas características na separação das classes. Entretanto, é necessário a estimação do parâmetro k, para indicar a quantidade de características importantes que devem ser retornadas pelo método.

3.7.3 *Recursive Feature Elimination*

Uma segunda maneira de aplicar a seleção de *features* é por meio da *Recursive Feature Elimination* (RFE). Este princípio parte de um Modelo de ML qualquer, desenvolvido a partir dos dados de treinamento, que possui como propriedade o cálculo de importância das características no Modelo. Desse modo, a cada iteração do RFE, ele realiza a eliminação das características menos importantes recursivamente. Esses passos são repetidos até que seja alcançado um número específico de *features*, caso este número seja definido como um parâmetro da função. Caso contrário, retorna a metade da quantidade de *features* de entrada.

3.7.4 *Shapley Additive Explanations*

O *Shapley Aditive Explanations* (SHAP) é um método que busca explicar as decisões de um Modelo de ML de forma intuitiva, ele é utilizado após o desenvolvimento do Modelo para analisar não somente as variáveis mais importantes, mas também como cada valor em específico do registro impacta nas decisões do Modelo. Essa metodologia tem sua formulação baseada na teoria dos jogos, entretanto, para a utilização desse método de explicação dos Modelos de ML, não é necessário o entendimento a fundo de todo o cálculo

matemático explicado na teoria, uma vez o resultado é apresentado de forma gráfica e intuitiva. Sucintamente, após o desenvolvimento de um Modelo, é possível calcular o valor SHAP e obter como saída o resultado de como o Modelo se comportou de acordo com os dados de entrada. Isto é, dado uma variável X, o impacto dela foi positivo ou negativo para a decisão do Modelo, de acordo com seus respectivos valores assumidos.

3.8 Considerações Finais

Utilizando técnicas e algoritmos de AI tem-se como objetivo desenvolver um Modelo capaz de identificar um padrão nos dados curriculares dos alunos, com o intuito de maximizar a predição de alunos evadidos. São inúmeros os algoritmos para realização da tarefa de classificação. Entretanto, para que os cenários de testes fossem igualmente comparáveis, utilizou-se então o algoritmo *Random Forest*. Além disso, como o conjunto de dados dispõe de uma quantidade considerável *features*, que podem ou não impactar na resolução do problema, foram utilizadas as técnicas de seleção de variáveis apresentadas anteriormente, para aperfeiçoar e compreender melhor o resultado. O capítulo seguinte apresenta os resultados em diferentes cenários e abordagens.

4 Análise dos Resultados

Este capítulo é destinado para abordar e discutir os cenários em que o algoritmo *Random Forest* foi aplicado, e os Modelos de ML gerados para identificar o padrão curricular dos alunos do ICEA. Os testes dirigiram-se para um resultado que pudesse maximizar a classe de alunos evadidos, tal como distinguir as características que mais impactam na separação das classes alvo, são elas: os alunos em situação de evadido, matriculado e diplomado. Dessa maneira, é possível fundamentar e fomentar a tomada de decisão acerca da evasão precoce no instituto, bem como conhecer as características que desassociam esses grupos.

Para manter o padrão dos testes uma biblioteca de Auto ML, a *PyCaret* foi utilizada para a realização de todos os cenários aplicados. A seguir na Tabela 17 é apresentada a parametrização do algoritmo *Random Forest*. Além disso, a biblioteca dispões da realização do treinamento do algoritmo utilizando a validação cruzada, com o valor *default* de k = 10.

Tabela 17 – Parametrização *Random Forest*

Parâmetro	**Referência**
bootstrap	True
ccp_alpha	0.0
class_weight	None
criterion	gini
max_deph	None
max_features	auto
max_leaf_node	None
max_samples	None
min_impurity_decrease	0.0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0.0
n_estimators	100
n_jobs	−1
oob_score	False
random_state	100
verbose	0
warm_start	False

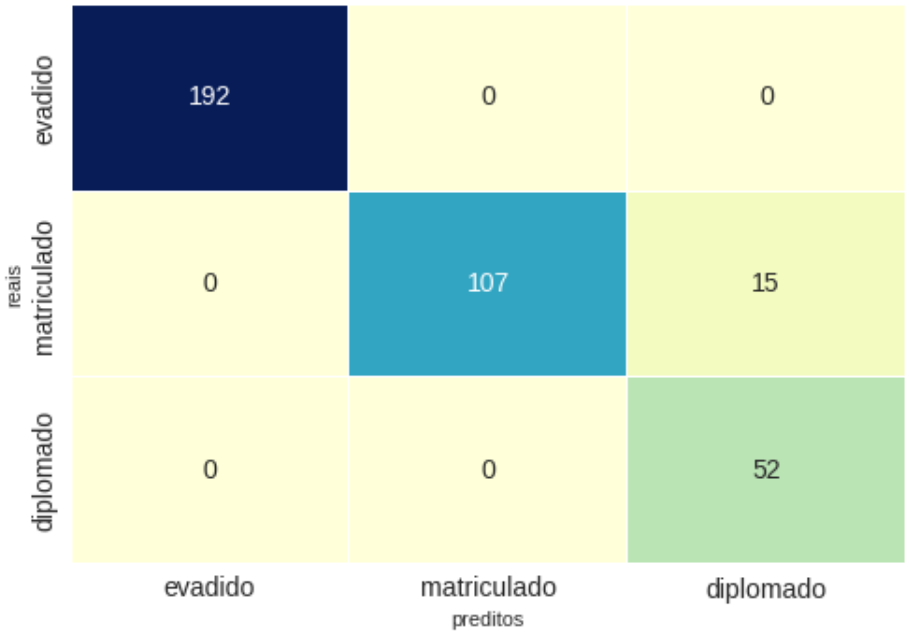
Fonte: Autora do Trabalho

Todos os cenários criados apresentam uma breve descrição de sua aplicação. Logo em seguida temos os resultados que o Modelo foi capaz de atingir. O conjunto de resultado é comporto pela acurácia, *f1-score*, taxa de acetos da classe de alunos evadidos e as variáveis mais importantes que foram consideradas pela própria *Random Forest*.

4.1 Cenário 1: Todas as Variáveis

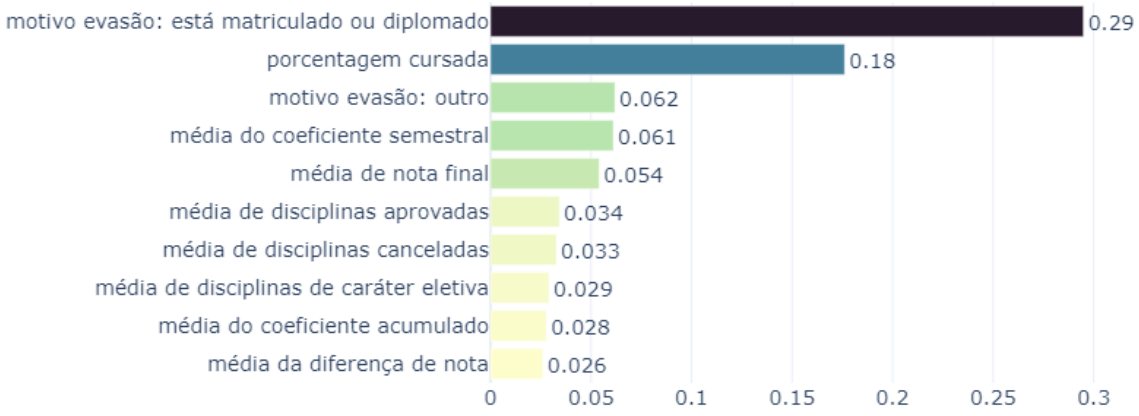
Foi realizado um experimento no qual se integrou todas as variáveis disponibilizadas para o estudo e as criadas no processo de *Feature Engineering* (Seção 3.5). A partir disso, foi proposto um Modelo que contemplou todas essas variáveis. O Modelo proposto atingiu 95.90% de acurácia e o *f1-score* de 96%. Além disso, o Modelo previu todas as classes da situação de evasão nos dados de validação. É apresentado a matriz de confusão na Figura 12, e as variáveis mais importantes do problema na Figura 13.

Figura 12 – Matriz de Confusão Cenário 1.



Fonte: Autora do trabalho

Figura 13 – *Feature Importance* Cenário 1.



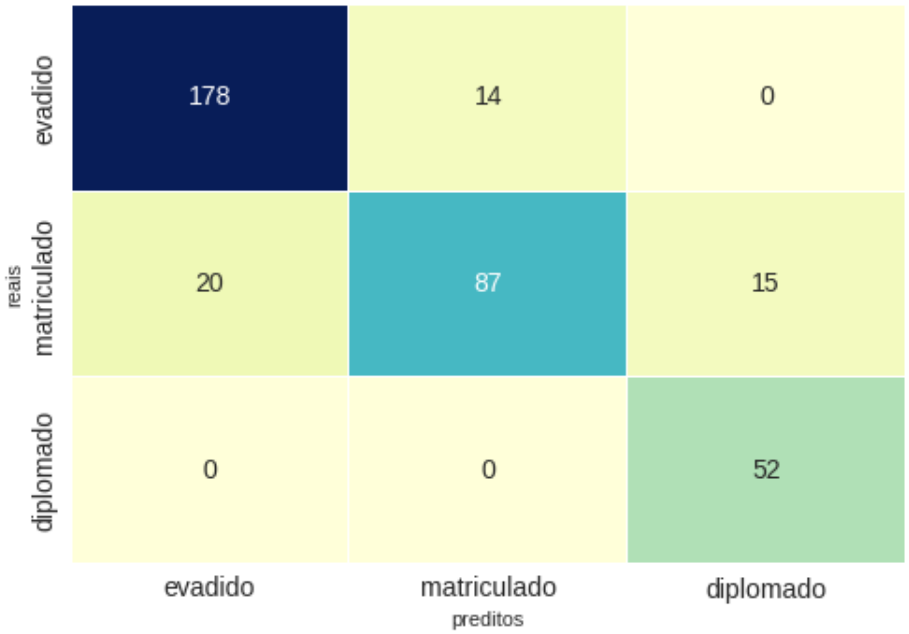
Fonte: Autora do trabalho

Tem-se que a variável que representa o motivo de evasão, matriculado ou diplomado, separou muito bem as classes de forma discrepante das outras variáveis. Este Modelo atingiu uma excelente desempenho e foi possível identificar todos os alunos evadidos. Entretanto, a *feature* "motivo evasão: está matriculado ou diplomado", foi criada para diferenciar os alunos que tinham algum motivo de evasão dos alunos que estavam em situação de matriculado ou diplomado. Por conseguinte, foi possível prever perfeitamente este cenário, pois houve um vazamento de informação das classes reais. Como o objetivo do trabalho não implica em ter o conhecimento da motivação prévia destes alunos, para os próximos experimentos, todas as variáveis que referenciavam algum motivo de evasão foram removidas.

4.2 Cenário 2: Remoção *Tags* de Evasão

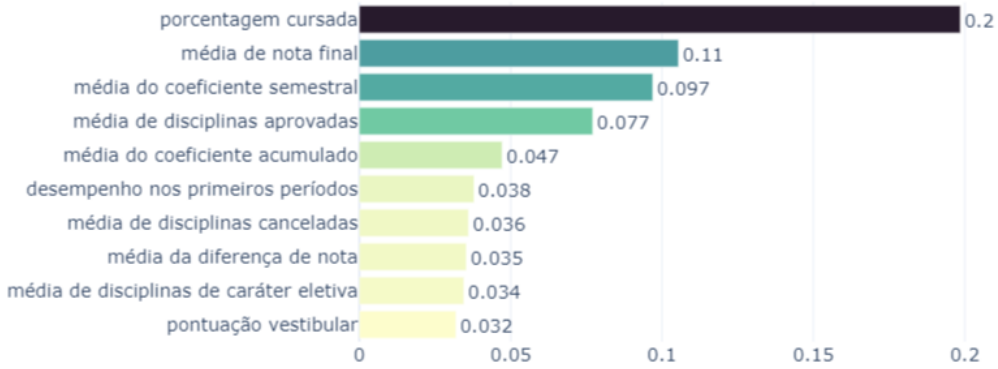
Após remover as variáveis que continham a motivação de evasão, foi gerado um novo Modelo. Esse modelo foi capaz de atingir 86.61% de acurácia, com o *f1-score* de 86%. Este Modelo foi capaz de acertar 92% das classes de evasão. A matriz de confusão é apresentada na Figura 14 seguido das variáveis mais importantes na Figura 15.

Figura 14 – Matriz de Confusão Cenário 2.



Fonte: Autora do trabalho

Figura 15 – *Feature Importance* Cenário 2.



Fonte: Autora do trabalho

As características mais importantes neste cenário, em sua maioria, são as características que possuem correlação forte com a variável alvo, conforme apresentado na Figura 16.

Figura 16 – Correlações Fortes da Base de Dados.

	variável alvo	porcentagem cursada	média de nota final	média de coeficiente semestral	média de disciplinas aprovadas
variável alvo	1,0000	0,8187	0,7225	0,7233	0,7340
porcentagem cursada	0,8187	1,0000	0,7341	0,7322	0,7539
média de nota final	0,7225	0,7341	1,0000	0,9967	0,9557
média de coeficiente semestral	0,7233	0,7322	0,9967	1,0000	0,9557
média de disciplinas aprovadas	0,7340	0,7539	0,9547	0,9557	1,0000

Fonte: Autora do trabalho

Conforme abordado na Seção 3.7.1, essas variáveis impactam negativamente na construção de um Modelo. Neste cenário pode-se observar o impacto delas, de modo a apresentar um viés no Modelo. Diante disso, essas características não foram levadas para os próximos cenários.

4.3 Cenário 3: Remoção da Multicolinearidade

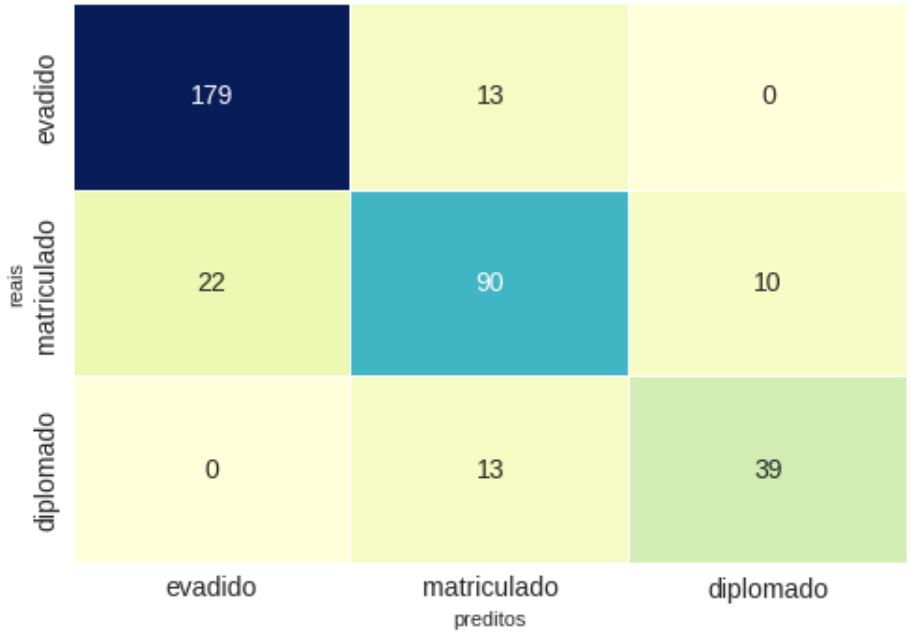
A multicolinearidade ocorre quando as variáveis estão muito correlacionadas entre si, conforme explicado na Seção 3.7.1, foi definido um *threshold* de -0.7 e 0.7 para definir essa relação.Diante disso, o Modelo desenvolvido neste cenário tem como objetivo não utilizar as variáveis que estão fortemente correlacionadas com a variável alvo e as que estão fortemente correlacionadas entre si. O propósito é retirar o viés causado pelo problema da multicolinearidade. Logo, foram removidas as seguintes variáveis fortemente correlacionadas com a variável alvo: porcentagem cursada, media de nota final, média de coeficiente semestral e a média de disciplinas aprovadas.

A variável média de coeficiente acumulado e a variável de desempenho nos primeiros períodos estavam fortemente correlacionadas entre si. Devido a esse fato, uma pode ser interpretada pela outra. A variável de desempenho nos primeiros períodos foi eleita para permanecer no Modelo, devido à literatura apresentar casos de sucesso com essa variável, (LEHMAN, 2005). Portanto, a variável que representa a média de coeficiente acumulado foi removida.

Assim sendo, desenvolvido o Modelo com tais aspectos, foi possível obter uma

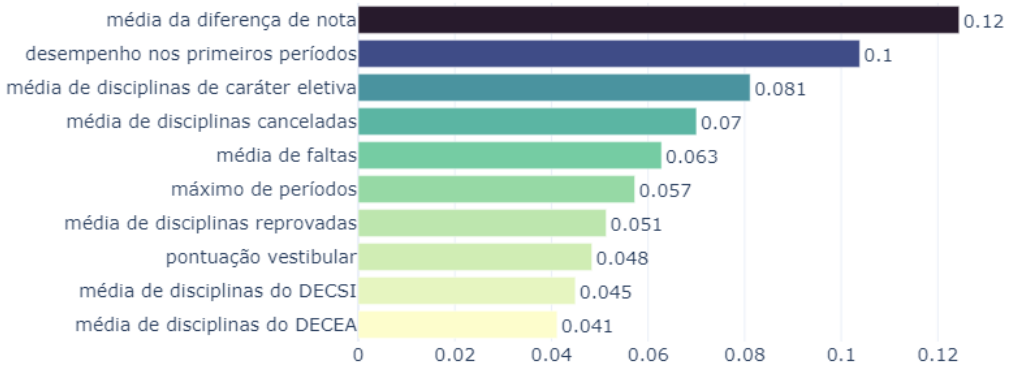
acurácia de 84.15% e um *f1-score* de 84%. Para este caso, foi possível prever cerca de 93% das ocorrências das classes de evasão. Veja na Figura 17 e Figura 18 a matriz de confusão e o *ranking* da importância das características.

Figura 17 – Matriz de Confusão Cenário 3.



Fonte: Autora do trabalho

Figura 18 – *Feature Importance* Cenário 3.



Fonte: Autora do trabalho

Após retirar as características de viés, o Modelo continuou eficiente para a predição de alunos evadidos e apresentou como características relevantes. Alguns desses fatos

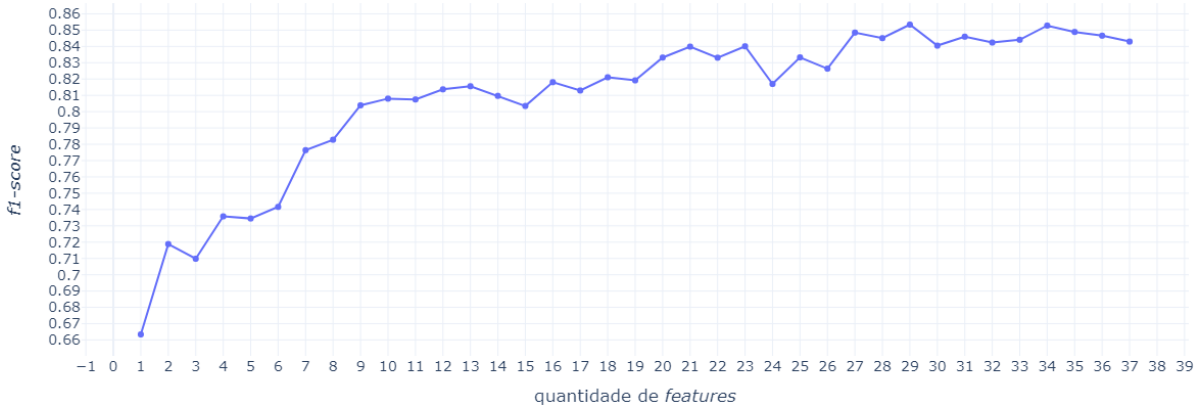
foram observados anteriormente na análise de dados e na literatura, como, por exemplo o desempenho nos primeiros períodos, a média de faltas e de disciplinas reprovadas (FLORES; HERAS; JULIAN, 2022) e (LEHMAN, 2005). Não obstante, este Modelo é composto por 37 *features*, sendo que esse número pode ser aperfeiçoado pela técnica de seleção de *features*, cujo objetivo é trazer ganho nas métricas e na performance do Modelo.

4.4 Cenário 4: Aplicação Seleção de *Features*: *Select k-best*

A seleção de *features* é uma técnica aplicada para melhorar o desempenho do Modelo e reduzir sua complexidade. Modelos com muitas dimensões podem ser custosos em relação ao tempo de execução, memória e eficiência. Dito isso, pode existir um subconjunto de caraterísticas que por si só contemple uma alta precisão do Modelo e uma boa acurácia, a baixo custo de recursos computacionais.

Com a finalidade de avaliar o desempenho do Modelo *Random Forest*, foi realizado um experimento utilizando a técnica *Select k-best*, apresentada na Seção 3.7.2 variando a quantidade de características do problema de 01 a 37. A métrica de avaliação abordada como medida de desempenho será o *F1-score*. Desse modo, é apresentado na Figura 19 a evolução da métrica de acordo com a quantidade de características utilizada para desenvolvimento do Modelo.

Figura 19 – Seleção de Características - Dados de Treino - Cenário 4.



Fonte: Autora do trabalho

É possível compreender que os Modelos que possuem menos de 8 variáveis obtém um *F1-score* inferior a 80%, e A partir de 16 características tende a obter um resultado melhor. O maior *F1-score* encontrado foi o de 85.34%, o qual utiliza 29 *features*. A partir disso, foi realizado um novo experimento que contemplou as 29 variáveis, sendo elas apresentada na Tabela 18.

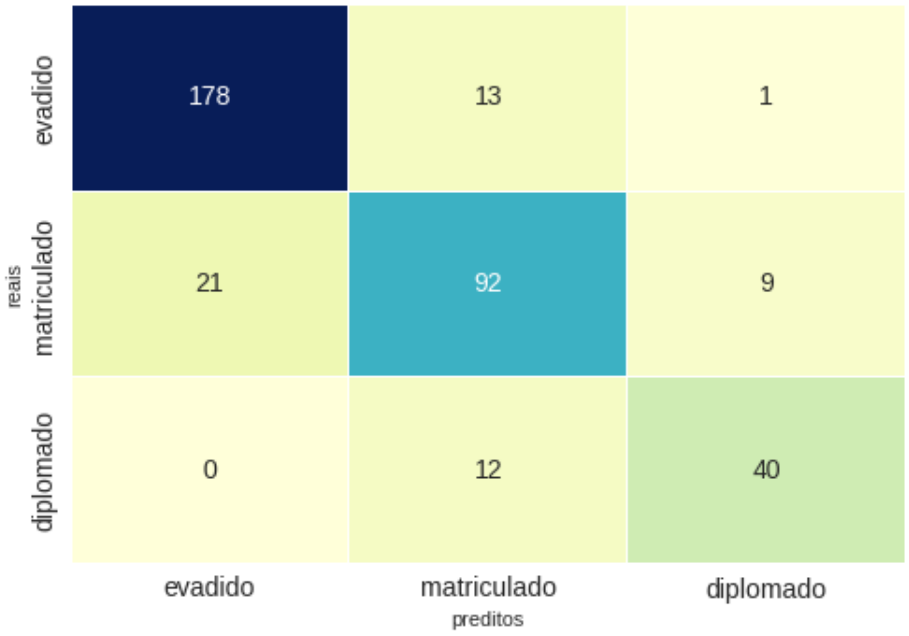
Tabela 18 – Variáveis *Select k-best*.

Variáveis
Classificação vestibular
Curso EC
Curso EP
Desempenho nos primeiros períodos
Idade
Máximo de períodos
Média de diferença de notas
Média de disciplinas canceladas
Média de disciplinas eletivas
Média de disciplinas facultativas
Média de disciplinas cursadas do DECEA
Média de disciplinas cursadas do DECSI
Média de disciplinas cursadas do DEELT
Média de disciplinas cursadas do DEENP
Média de disciplinas reprovadas
Média de disciplinas trancadas
Média de exames especiais
Média de faltas
Modalidade de concorrência ac
Modalidade de concorrência L1
Modalidade de concorrência L2
Modalidade de concorrência L5
Modalidade de concorrência L6
Modalidade de concorrência paa
Modo de admissão duplo diploma
Modo de admissão portador de diploma de graduação
Pontuação no vestibular
Tempo de curso
Tipo de escola

Fonte: Autora do Trabalho

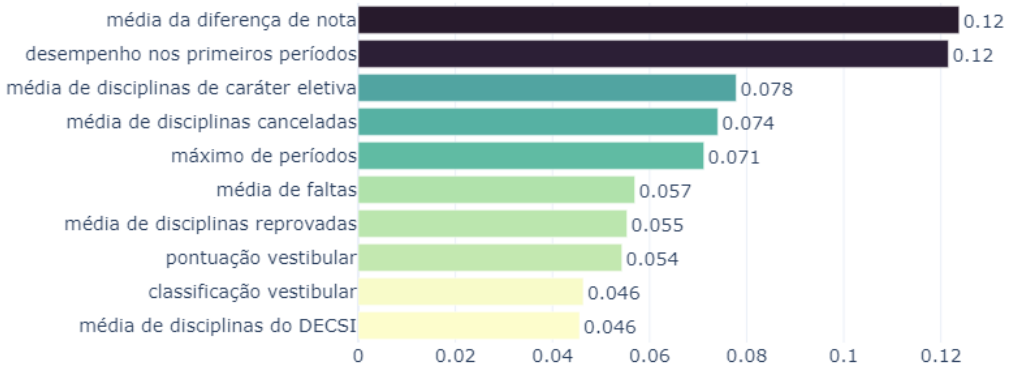
Na Figura 20 e na Figura 21 é possível visualizar a matriz de confusão e as características mais importantes segundo o Modelo. O valor de acurácia é de 84.70% e o *f1-score* atingiu 85%. Além disso, esse Modelo acertou cerca de 92% dos alunos da classe de evasão, presentes no conjunto de validação.

Figura 20 – Matriz de Confusão Cenário 4.



Fonte: Autora do trabalho

Figura 21 – *Feature Importance* Cenário 4.



Fonte: Autora do trabalho

Após aplicar a seleção de características, é possível identificar que o Modelo teve uma pequena melhora em 0.55% de acurácia. Entretanto, neste contexto não foi uma melhora tão significativa, e fez com que os verdadeiros positivos da evasão diminuísse em 1 caso. Uma segunda técnica de seleção de *features* foi aplicada, o que gera um novo cenário. Para fins de comparação de desempenho, esse assunto é apresentado na próxima seção.

4.5 Cenário 5: Aplicação Seleção de *Features*: RFE

A *Recursive Feature Elimination* é uma técnica utilizada para seleção de características e foi empregue neste trabalho, conforme abordado na Seção 3.7.3. Esta técnica carece de um Modelo desenvolvido previamente utilizando os dados de treinamento para que seja possível retirar as dimensões do problema de maneira recursiva. Desse modo, o Modelo remove as variáveis uma a uma e avalia o seu desempenho após a remoção da característica e avalia de essa ação foi capaz de piorar ou melhorar as métricas. Posto isso, foi criada uma *Random Forest* simples com os seguintes parâmetros: número de estimadores igual à 10, e o máximo de profundidade igual à 3. Após a etapa de construção de um Modelo qualquer, é utilizado uma função do *sklearn* que realiza a implementação do RFE. Desse modo, a função retorna as características mais significantes no Modelo criado, neste caso 18 variáveis que são apresentadas a seguir na Tabela 19:

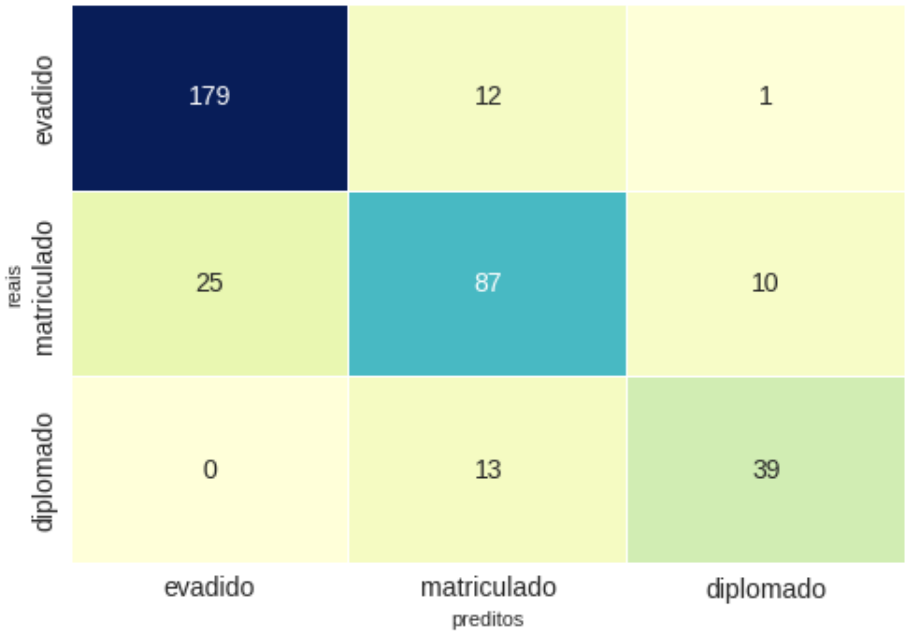
Tabela 19 – Variáveis RFE.

Variáveis
Curso EP
Desempenho nos primeiros períodos
Máximo de períodos
Média de diferença de notas
Média de disciplinas canceladas
Média de disciplinas cursadas de caráter eletiva
Média de disciplinas cursadas de caráter obrigatória
Média de disciplinas cursadas do DECEA
Média de disciplinas cursadas do DECSI
Média de disciplinas cursadas do DEELT
Média de disciplinas cursadas do DEENP
Média de disciplinas trancadas
Média de exames especiais feitos
Média de faltas
Média disciplinas reprovadas
Modalidade de concorrência L1
Modalidade de concorrência L2
Modalidade de concorrência L5

Fonte: Autora do Trabalho

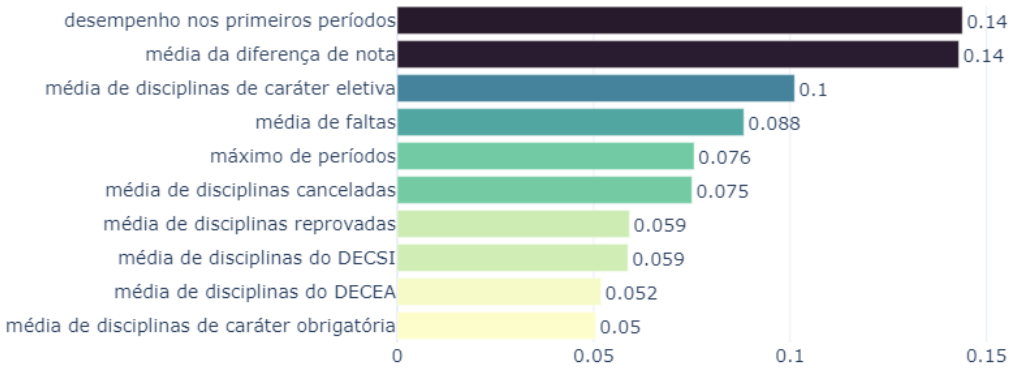
Este cenário gerou um Modelo com 83.36% de acurácia e o *f1-score* de 83%. Ademais, houve um acerto referente à classe de evadidos de, aproximadamente, 93%. Na Figura 22 é apresentada a matriz de confusão dos dados de validação, bem como na Figura 23, a precedência das características mais impactantes neste cenário.

Figura 22 – Matriz de Confusão Cenário 5.



Fonte: Autora do trabalho

Figura 23 – *Feature Importance* Cenário 5.



Fonte: Autora do trabalho

Embora a acurácia e o *f1-score* do Modelo tenha sido inferior ao do Cenário 4, 4.4, aumentou em 1 caso a exatidão da previsão dos alunos evadidos. O maior diferencial deste Modelo com os demais válidos, os sem viés presentes a partir do Cenário 3, é a quantidade de características utilizadas. Neste cenário, é utilizada uma menor quantidade de características, porém, de maneira assertiva.

Durante o *plot* das *feature importances*, é possível notar a presença de dois subtipos de variáveis no *ranking*. São as características provenientes de um estado anterior ao aluno ingressar na faculdade e uma posterior ao seu ingresso. Entretanto, neste cenário, apenas as características *Pós Universidade* são apresentadas com maior grau de importância. A partir disso, um novo experimento é criado para fazer uma referência entre esses dois aspectos e o padrão curricular dos discentes do Instituto, também citado no estudo de (VILORIA et al., 2020).

4.6 Cenário 6: Variáveis Pré Universidade

A base de dados é composta pelos atributos referentes ao desempenho do aluno no ENEM e no Sistema de Seleção Unificada (SISU) e dos dados gerados a partir do histórico do aluno no ICEA. Desse modo, é viável a análise do padrão dos alunos antes e depois de ingressar nos respectivos cursos. Para isso, foi desenvolvido um Modelo que inclui apenas as variáveis chamadas de *Pré Universidade*, ou seja, que contém somente as características referentes à jornada do aluno antes de seu vínculo com o ICEA. As variáveis são apresentadas a seguir na Tabela 20:

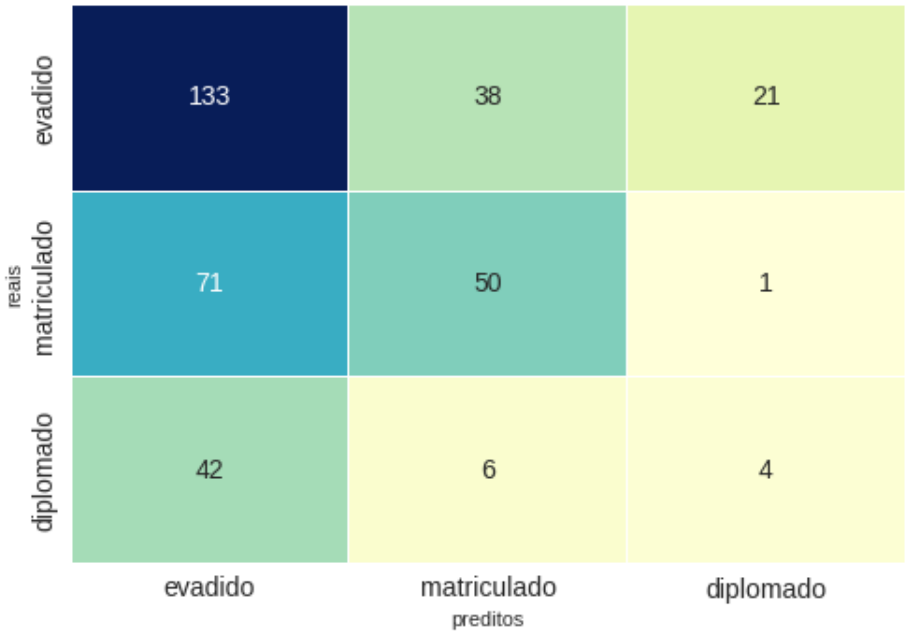
Tabela 20 – Variáveis Pré Universidade.

Variáveis
Classificação do Vestibular
Modo de admissão: convênio duplo diploma
Modo de admissão: convênio programa pec-g
Modo de admissão: decisão judicial
Modo de admissão: portador de diploma de graduação
Modo de admissão: transferência externa
Idade
Modalidade de Concorrência: AC
Modalidade de Concorrência: L1
Modalidade de Concorrência: L2
Modalidade de Concorrência: L5
Modalidade de Concorrência: L6
Modalidade de Concorrência: PAA
Pontuação do Vestibular
Região Próxima

Fonte: Autora do Trabalho

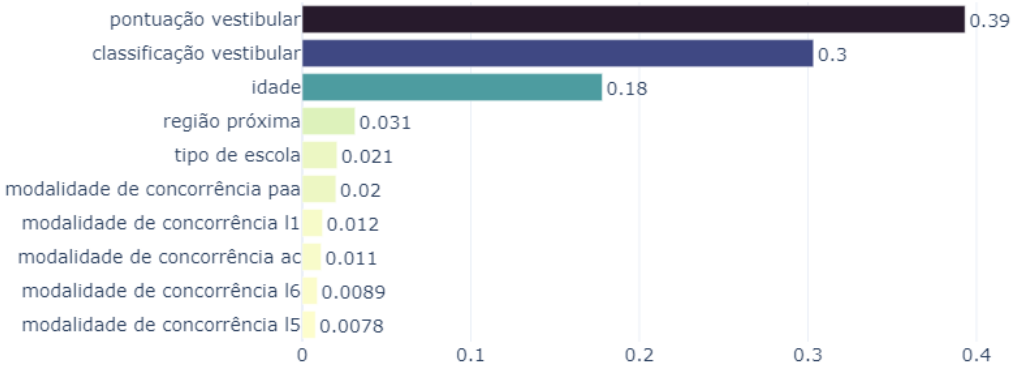
Posto isso, o Modelo desenvolvido obteve uma acurácia de 51.09% e um *f1-score* de 49%. Foi observado na matriz de confusão, que a taxa de acertos da classe de evadido diminuiu e o Modelo teve uma maior quantidade de erro. De maneira geral, este Modelo apresentou uma baixa performance em relação aos apresentados até o momento. Nas Figura 24 e na Figura 25 são apresentadas a matriz de confusão e a classificação da importância das características, respectivamente.

Figura 24 – Matriz de Confusão Cenário 6.



Fonte: Autora do trabalho

Figura 25 – *Feature Importance* Cenário 6.



Fonte: Autora do trabalho

O baixo desempenho do Modelo indica que o padrão curricular dos discentes, não pode ser definido apenas pelas características *Pré Universidade*. Além disso, quando abordado a representatividade das características importantes, as variáveis pontuação do vestibular e classificação no vestibular tiveram uma pontuação alta na decisão do Modelo também em outros cenários, conforme apresentado na Figura 18 e na Figura 21.

Foi possível indicar que o grau de importância das modalidades de concorrência presentes no SISU é muito baixa, tal fato aponta que esses fatores não são determinantes para contribuir com a taxa de evasão aplicada neste contexto.

4.7 Cenário 7: Variáveis Pós Universidade

O grupo de variáveis *Pós Universidade* é composto pelas características geradas quando aluno se matricula no curso, sendo elas apresentadas na Tabela 21:

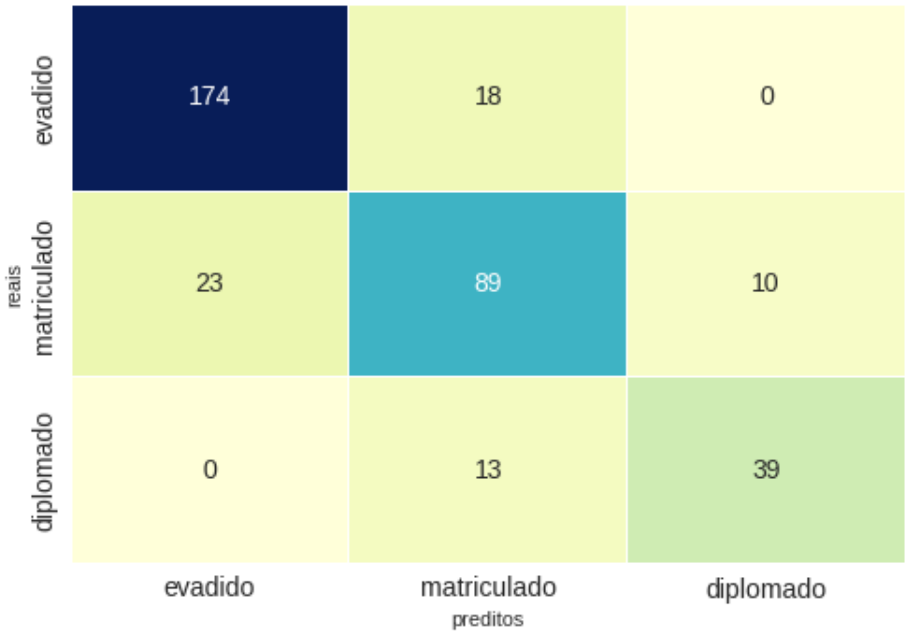
Tabela 21 – Variáveis Pós Universidade.

Variáveis
Curso EC
Curso EE
Curso EP
Curso SI
Desempenho nos primeiros períodos
Máximo de períodos
Média de disciplinas eletivas
Média de disciplinas facultativas
Média de disciplinas obrigatórias
Média de disciplinas canceladas
Média de disciplinas reprovadas
Média de disciplinas trancadas
Média de disciplinas cursadas do DECEA
Média de disciplinas cursadas do DECSI
Média de disciplinas cursadas do DEELT
Média de disciplinas cursadas do DEENP
Média da diferença de nota
Média de exame especial
Média de faltas
Tempo de curso
Turno

Fonte: Autora do Trabalho

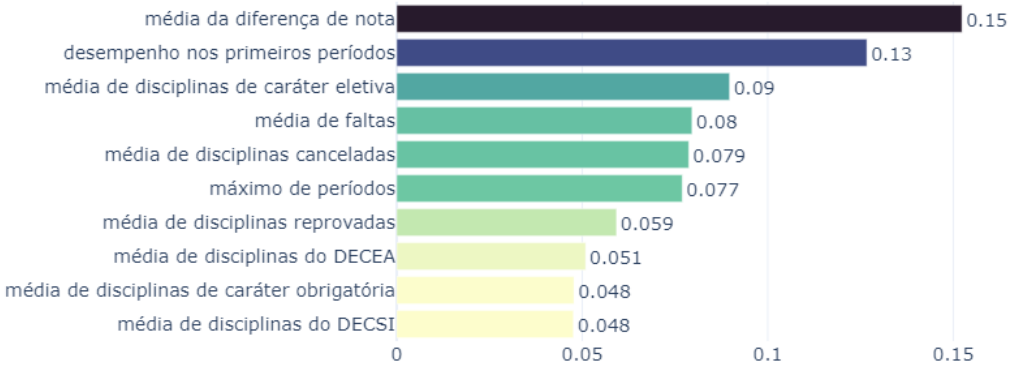
Posto isso, foi desenvolvido um Modelo a partir dessas variáveis. Ele obteve uma acurácia de 82.51%, junto ao *f1-score* de 82%. Ademais, a taxa de acerto foi a segunda mais baixa, com apenas 90%. A Figura 26 e a Figura 27 representam a matriz de confusão e os atributos mais importantes para este cenário, respectivamente.

Figura 26 – Matriz de Confusão Cenário 7.



Fonte: Autora do trabalho

Figura 27 – *Feature Importance* Cenário 7.



Fonte: Autora do trabalho

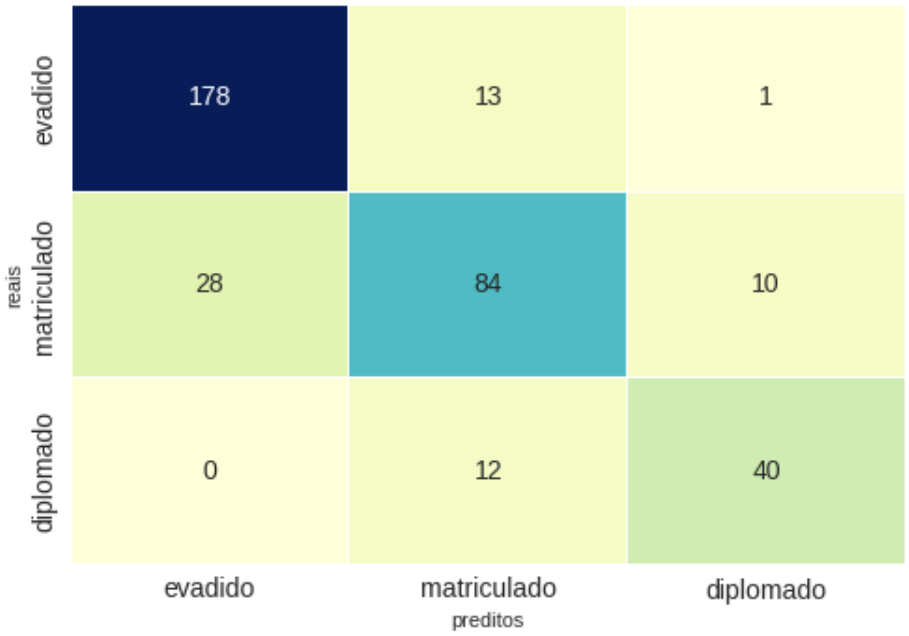
Foi observado que as variáveis *Pré* e *Pós Universidade*, quando trabalhadas de formas isoladas, reduzem o desempenho dos Modelos. Por essa razão, surge a ideia de trabalhar apenas com as variáveis destaques do Cenário 7 e do Cenário 8. Esse experimento é abordado na seção seguinte.

4.8 Cenário 8: Variáveis Destaques

Após avaliar as características do problema que mais impactam no desenvolvimento de um Modelo de ML em relação ao grupo de variáveis *Pré* e *Pós Universidade*, foi criado um Modelo somente com as características mais importantes dos respectivos cenários. Essas características estão apresentadas na Figura 25 e na Figura 27.

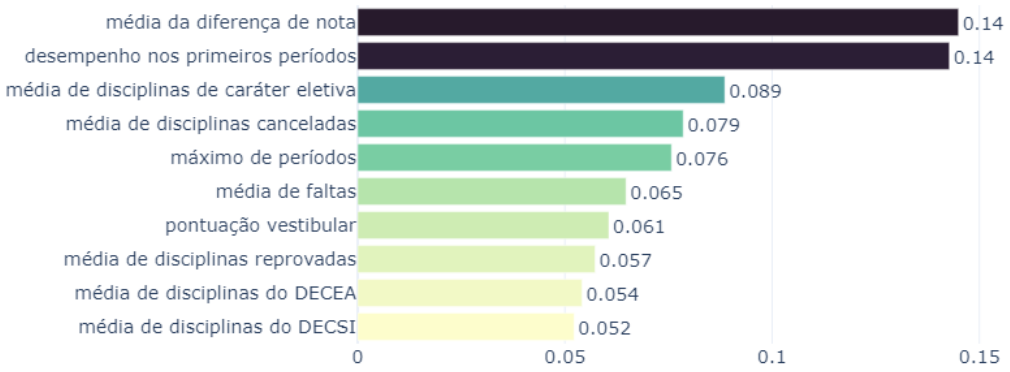
O Modelo desenvolvido apresentou uma acurácia de 82.51%, como o valor de *f1-score* igual à 82%. Esse Modelo conseguiu uma taxa de acerto de aproximadamente 92%. A seguir, na Figura 28 e na Figura 29 são apresentadas, respectivamente, a matriz de confusão e o *ranking* da representatividade dos atributos.

Figura 28 – Matriz de Confusão Cenário 8.



Fonte: Autora do trabalho

Figura 29 – *Feature Importance* Cenário 8.



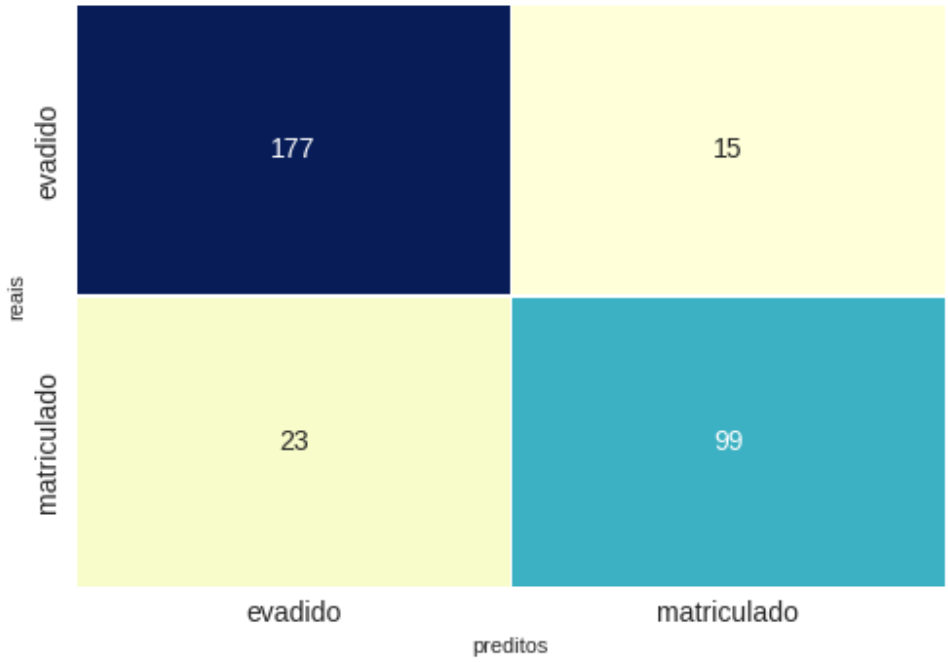
Fonte: Autora do trabalho

Tem-se que, mesmo utilizando somente as melhores características, o Modelo não obteve uma acurácia muito diferente das demais, sugerindo que as outras técnicas de seleção de variáveis se aplicam melhor ao contexto do ICEA.

4.9 Cenário 9: Classes Binárias

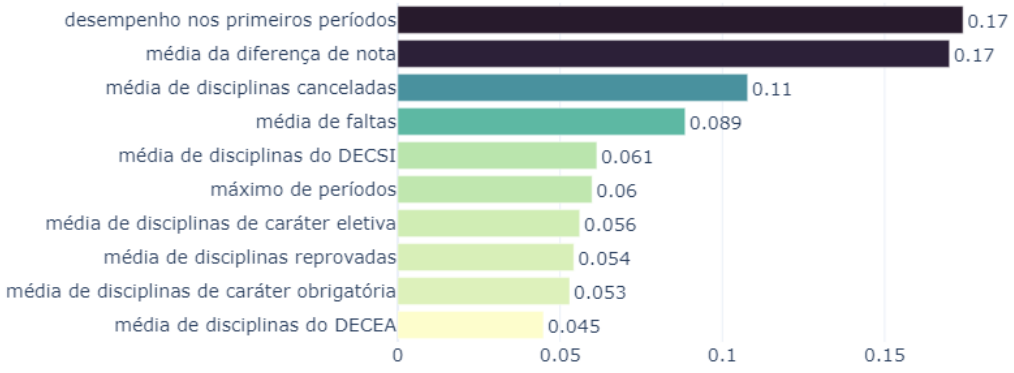
Segundo Anupama, Meenu e Supriya (2019), problemas que envolvem duas classes geralmente são mais performáticos em alguns algoritmos de ML. Como este trabalho está modelado com a representatividade de três classes, foi realizado um teste com a exclusão da classe de alunos diplomados. Os atributos utilizados para este teste foram os mesmos utilizados no cenário da Seção 4.5, o qual obteve uma boa performance. Desse modo, o Modelo gerado atingiu uma acurácia de 87.90%, com um *f1-score* de 88%. Além disso, teve uma taxa de acerto de 93% para os alunos evadidos. Com isso, esse Modelo passa a ser o melhor até o presente momento. A seguir, na Figura 30 e na Figura 31 é possível visualizar a matriz de confusão e as características mais importantes, respectivamente.

Figura 30 – Matriz de Confusão Cenário 9.



Fonte: Autora do trabalho

Figura 31 – *Feature Importance* Cenário 9.



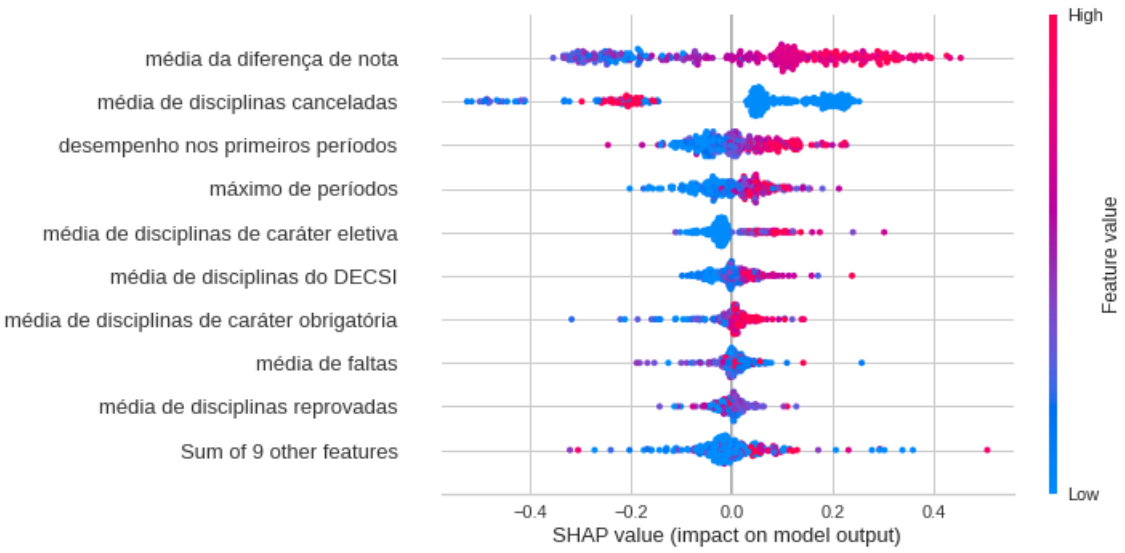
Fonte: Autora do trabalho

Este cenário não se difere expressivamente dos outros, como os que foram aplicados as técnicas de seleção de *features*, os da Seção 4.4 e da Seção 4.5, vinculados à precisão da classe de evasão. Porém, de acordo com a acurácia e *f1-score*, o Modelo é eficiente. Esse fato sugere que utilizando a seleção de *features* é possível atingir um resultado tão bom

quanto empregando apenas duas classes. É importante salientar que todos os Modelos desenvolvidos até o momento são válidos para realização de outros testes, de modo a excluir os que continham algum viés, os dos Cenário 1 (Seção 4.1) e Cenário 2 (Seção 4.2).

Como este cenário obteve um melhor resultado, foi então calculado o valor SHAP e gerado um gráfico para a visualização da explicabilidade do Modelo. O resultado é apresentado na Figura 32.

Figura 32 – Explicabilidade do Modelo - Valor SHAP.



Fonte: Autora do trabalho

Neste gráfico é possível analisar as variáveis que possuem impacto no Modelo, não de forma isolada, mas sim como os valores influenciam na tomada de decisão do Modelo. Cada ponto do gráfico é atribuído a uma instância do problema, com seus respectivos valores reais. É importante frisar que os dados estão normalizados, entretanto, este fator não interfere na análise dos resultados. Para este contexto, tem-se que a média de disciplinas canceladas impacta positivamente no Modelo, quando menor é o valor desse atributo. Ou seja, se a média de disciplinas canceladas for baixa, significa que o aluno não realiza o trancamento de disciplinas, com isso ele impacta de forma positiva no Modelo, maximizando a separação das classes e esse aluno tem menos chance de pertencer à classe de evadidos.

O mesmo raciocínio segue para as demais características, como, por exemplo, a média de disciplinas do DECSI. Quanto maior é a média de disciplinas realizadas deste departamento, o impacto tende a ser positivo para o Modelo e o aluno tem menos chance de pertencer à classe de evadidos. Sabe-se que esse departamento possui disciplinas comuns a todos os cursos. Algumas dessas disciplinas são oferecidas em períodos iniciais, como Programação de Computadores I e Algoritmo e Estruturas de Dados I. Ademais, essas

disciplinas possuem altas taxas de reprovação. Para entender de fato a tomada de decisão do Modelo, é viável uma investigação mais aprofundada a respeito das características de saída de um Modelo de ML.

4.10 Cenário 10: Outros Algoritmos

Os algoritmos utilizados para a tarefa de classificação não se limitam à *Random Forest*. Existem outros algoritmos que não foram abordados de forma explicativa neste trabalho. Esses algoritmos também foram implementados para fim de comparação com os resultados obtidos até o momento. Foram escolhidas as características do melhor Modelo criado neste trabalho, e também considerou-se como base as três classes presentes, o Modelo do Cenário 5 (Seção 4.5). Posto essas informações, foram desenvolvidos alguns algoritmos além da *Random Frest*. A seguir, na Tabela 22 é possível identificar os resultados como a acurácia e o *f1-score*.

Tabela 22 – Métricas de Diferentes Classificadores - Treinamento.

Classificador	**Acurácia**	***F1-score***
Gradient Boosting Classifier	80.19%	79%
Random Forest Classifier	79.96%	79%
Extra Trees Classifier	79.73%	79%
Linear Discriminant Analysis	76.92%	76%
K Neighbors Classifier	76.43%	76%
Ada Boost Classifier	74.45%	73%
Naive Bayes	74.30%	74%
Decision Tree Classifier	73%	72%

Fonte: Autora do Trabalho

Como observado na Tabela 22, o melhor Modelo de classificação aplicado ao conjunto de dados do problema em nível de acurácia foi o *Gradient Boosting*, seguido pela *Random Forest*, utilizada como base do desenvolvimento deste trabalho. Ambos se diferem apenas por 0.23% de acurácia, já em análise do *f1-score*, esses algoritmos se mantém equivalentes.

4.11 Considerações Finais

De acordo com os cenários aplicados, pode-se observar as mudanças de resultados dos Modelos desenvolvidos. Criar cenários com variáveis ainda desconhecidas, foi um equívoco não observado durante a etapa de Análise. Entretanto a flexibilidade do ciclo do projeto, permitiu fazer as alterações necessárias em tempo hábil e que entregasse melhores resultados. O viés da multicolinearidade apresentado no Cenário 4.2, caso não fosse observado, geraria uma diferença de 6% no resultado de acurácia do Modelo. A

divisão entre as variáveis Pré e Pós Universidade foi interessante pois possibilitou entender o impacto delas no rendimento dos alunos, e obteve sucesso quando aplicadas de forma conjunta. A utilização das técnicas de seleção de *features*, permitiu trabalhar com menos variáveis, entretanto com variáveis objetivas. O *pipeline* de tarefas para a construção de todas as variáveis originais do problema levava aproximadamente 5 minutos de execução. Com a utilização do RFE foi possível executar essas tarefas em 37 segundos, um ganho de aproximadamente 87%, sem reduzir o desempenho do Modelo.

5 Conclusão

Utilizar *Artificial Intelligence* na esfera acadêmica permite o tratamento de uma quantidade abundante de dados, bem como o ganho de velocidade na geração de informação para a tomada de decisão. No cenário em que as verbas para a educação são cada vez menores e a taxa de abandono dos cursos são cada vez maiores, é preciso criar políticas capazes de reduzir esse impacto negativo na sociedade. A partir da análise das entrelinhas do problema e de um estudo baseado nos dados educacionais, é possível criar ferramentas para apoiar as decisões diante a comunidade acadêmica. Por esse motivo, este trabalho atuou nessa frente, utilizando técnicas de AI e desenvolvendo diversos Modelos de *Machine Learning* capazes de compreender o padrão curricular dos discentes Instituto de Ciências Exatas e Aplicadas.

Este trabalho analisou os dados dos discentes referente à situação acadêmica Pré e Pós Universidade. Através de técnicas de *Educational Data Mining* e AI foi possível compreender as características que impactam no sucesso e o fracasso acadêmico, a partir de um Modelo de ML. Foram desenvolvidos Modelos baseados em *Random Forest* aptos a classificar um discente em situação de evadido, diplomado ou matriculado. O estudo verificou os impactos das variáveis disponíveis na base de dados, bem como as desenvolvidas no processo de *Feature engineering*. Devido à dimensionalidade de características apresentadas no problema, viabilizou-se o desenvolvimento de Modelos para diferentes cenários. Uma das frentes foi o impacto das variáveis *Pré* e *Pós Universidade*, de tal modo a avaliar como o padrão dos dados se comportaram utilizando os conjuntos de características de forma isolada. Técnicas de seleção de variáveis, como o *Select k-best* e *Recursive Feature Elimination*, foram utilizadas para otimizar os resultados, já o *Shapley Aditive Explanations* (SHAP) para facilitar a explicabilidade e entendimento dos Modelos desenvolvidos.

Os resultados alcançados mostram-se bastante promissores em relação às métricas de avaliação de Modelos de ML. O melhor Modelo, capaz de identificar os alunos com maior propensão a evadir, atingiu uma acurácia de 87.90%, utilizando um subconjunto de características *Pré* e *Pós Universidade*, que por sua vez, constatou-se uma melhora significativa quando trabalhadas em conjunto.

Ademais, este trabalho aponta uma forma de identificar de forma precoce os discentes que possuem uma maior probabilidade de evasão. Essa frente de atuação permite uma gerência desses alunos, possibilitando propor ações preventivas de forma coletiva e/ou individual. São exemplos de políticas que podem ser implantadas para esse grupo, o acompanhamento do discente durante o período, disponibilização de tutores ou equipe

multidisciplinar de apoio para entender e atender às necessidades dos discentes nesta situação. O uso de estudos orientados à dados, além de aperfeiçoar os recursos e dar suporte à tomada de decisão, fortalece a cultura *data driven* e gera transparência para toda a sociedade.

5.1 Propostas para trabalhos futuros

Como trabalhos futuros, outros Modelos podem ser desenvolvidos a partir de cada curso específico, pois, acredita-se que as características marcantes para cada Modelo serão diferentes, adequando-se ao perfil de cada curso. Outra sugestão é enriquecer a base de dados com outros atributos não utilizados anteriormente, como, por exemplo, a partir de formulários pessoais ou outras fontes que também poderão ser úteis para continuar este estudo. Ademais, incorporar esses resultados em um *dashboard* para acompanhar a evolução do comportamento do Modelo e das métricas. Esta ação é indicada devido à possibilidade de mudança no perfil dos alunos, para o contexto do ICEA.

Referências

ACHARYA, A.; SINHA, D. Early prediction of students performance using machine learning techniques. *International Journal of Computer Applications*, v. 107, n. 1, p. 37–43, 2014. Citado 2 vezes nas páginas 15 e 22.

ANDIFES, A.; ABRUEM, A.; SESU/MEC, S. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a andifes, abruem e sesu/mec pela comissão especial. *Avaliação: Revista da Avaliação da Educação Superior*, v. 1, n. 2, 1996. Disponível em: <http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/739>. Citado 4 vezes nas páginas 18, 20, 22 e 23.

ANUPAMA, J.; MEENU, D.; SUPRIYA, M. Comparison of binary class and multi-class classifier using different data mining classification techniques. *Proceedings of International Conference on Advancements in Computing & Management (ICACM)*, 10 2019. Citado na página 64.

BEHR, A. et al. Early prediction of university dropouts – a random forest approach. *Jahrbücher für Nationalökonomie und Statistik*, v. 240, 02 2020. Citado na página 20.

BREIMAN, L. Simplifying decision trees. *International Journal of Man-Machine Studies*, v. 27, p. 221–234, 1987. Citado na página 40.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. Citado 2 vezes nas páginas 41 e 43.

CALDEIRA, D. M. *Caracterização do problema de evasão de discentes nos cursos do ICEA mediante técnicas de mineração de dados*. 2021. Monografia (Graduação em Sistema de Informação) - Instituto de Ciências Exatas e Aplicadas, Universidade Federal de Ouro Preto, João Monlevade, Brasil. Citado 2 vezes nas páginas 16 e 23.

CASTRO, C. L.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatção*, v. 5, p. 22, 2011. Citado na página 31.

DAVID, J. L.; CLARE, B.; DOLECK, T. Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, v. 2, p. 100016, 2021. Citado na página 15.

DUBOUE, P. *The Art of Feature Engineering: Essentials for Machine Learning*. [S.l.: s.n.], 2020. ISBN 9781108709385. Citado na página 32.

FILHO, R. S. et al. A evasão no ensino superior brasileiro. *Cadernos De Pesquisa*, v. 37, 12 2007. Citado na página 20.

FLORES, V.; HERAS, S.; JULIAN, V. Comparison of predictive models with balanced classes using the smote method for the forecast of student dropout in higher education. *Electronics*, v. 11, n. 3, p. 457, 2022. Citado 3 vezes nas páginas 20, 23 e 54.

HALVERSON, R. et al. The new instruction leadership: Creating data-driven instructional systems in schools. *Journal of School Leadership*, v. 25, 2006. Citado na página 15.

HARTMANN, P. et al. Capturing value from big data: A taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, v. 36, n. 10, p. 1382–1406, 2016. Citado na página 15.

IQBAL, Z. et al. Machine learning based student grade prediction: A case study. *Arxiv*, v. 1, 2017. Citado na página 15.

IRAJI, m. et al. Students classification with adaptive neuro fuzzy. *International Journal of Modern Education and Computer Science*, v. 4, 07 2012. Citado na página 20.

JAIN, A. et al. Intellectual performance analysis of students by comparing various data mining techniques. p. 57–63, 2017. Citado na página 22.

KONAR, A. *Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain*. [S.l.: s.n.], 1999. Citado na página 39.

LEHMAN, Y. P. *Estudo sobre a evasão universitária: as mudanças de paradigma na educação e suas consequências*. Dissertação (Mestrado) — Universidade de São Paulo, 2005. Citado 3 vezes nas páginas 17, 52 e 54.

MOHAMAD, S.; TASIR, Z. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, v. 97, p. 320–324, 2013. The 9th International Conference on Cognitive Science. Citado na página 23.

MORETTIN, P. A.; BUSSAB, W. d. O. *Estatística básica*. [S.l.]: Saraiva, 2004. Citado na página 45.

MUSSLINER, B. O. et al. O problema da evasão universitária no sistema público de ensino superior: uma proposta de ação com base na atuação de uma equipe multidisciplinar. *Brazilian Journal of Development*, v. 7, n. 4, p. 42674–42692, 2021. Citado 5 vezes nas páginas 15, 16, 18, 20 e 24.

MUTTATHIL, A.; RAHMAN, Z. Model of tuned j48 classification and analysis of performance prediction in educational data mining. *International Journal of Applied Engineering Research*, v. 13, p. 14717–14727, 2018. Citado na página 16.

RAMASWAMI, G.; SUSNJAK, T.; MATHRANI, A. On developing generic models for predicting student outcomes in educational data mining. *Big Data and Cognitive Computing*, v. 6, n. 1, 2022. Citado na página 15.

RAMASWAMI, M.; BHASKARAN, R. A study on feature selection techniques in educational data mining. *Journal of Computing*, v. 1, 12 2009. Citado na página 45.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN: Computer Science*, v. 2, p. 160, 2021. Citado na página 39.

STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, v. 62, n. 1, p. 77–89, 1997. Citado na página 43.

UFOP, U. F. d. O. P. *Processos Seletivos*. 2022. Disponível em: <https:/vestibular.ufop.br/index.php?option=com_content&view=article&id=1065:documentos-para-confirmacao-da-matricula-sisu-2021-1&catid=195&Itemid=216>. Citado na página 36.

VILORIA, A. et al. Dropout-permanence analysis of university students using data mining. Springer International Publishing, Cham, p. 374–383, 2020. Citado 3 vezes nas páginas 20, 23 e 59.

WEN, Y.; WEN, Y. Consideration of the local correlation of learning behaviors to predict dropouts from moocs. *Tsinghua Science & Technology*, v. 25, p. 336–347, 2020. Citado 2 vezes nas páginas 18 e 23.

XU, J.; MOON, K. H.; SCHAAR, M. van der. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, v. 11, n. 5, p. 742–753, 2017. Citado 2 vezes nas páginas 22 e 46.