UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

VICTOR AUGUSTO SOUZA DE OLIVEIRA

APRENDIZADO DE MÁQUINA APLICADO A EVASÃO NO ENSINO SUPERIOR

RIO DE JANEIRO
2023

VICTOR AUGUSTO SOUZA DE OLIVEIRA

APRENDIZADO DE MÁQUINA APLICADO A EVASÃO NO ENSINO SUPERIOR

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. João Carlos Pereira da Silva

RIO DE JANEIRO

2023

CIP - Catalogação na Publicação

O48a

Oliveira, Victor Augusto Souza de
Aprendizado de máquina aplicado a evasão no ensino superior / Victor Augusto Souza de Oliveira. -- Rio de Janeiro, 2023.
54 f.

Orientador: João Carlos Pereira da Silva.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Computação, Bacharel em Ciência da Computação, 2023.

1. Inteligência artificial. 2. Aprendizado de máquina. 3. Mineração de dados. 4. Evasão. 5. UFRJ. I. Silva, João Carlos Pereira da, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

VICTOR AUGUSTO SOUZA DE OLIVEIRA

APRENDIZADO DE MÁQUINA APLICADO A EVASÃO NO ENSINO SUPERIOR

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 28 de abril de 2023

BANCA EXAMINADORA:

Prof. João Carlos Pereira da Silva
D.Sc. (UFRJ)

Profa. Juliana Vianna Valerio
D.Sc. (PUC-Rio)

Profa. Silvana Rossetto
D.Sc. (PUC-Rio)

Dedicatória: Dedico esse trabalho de conclusão de curso primeiramente à Deus, que me conduziu por caminhos improváveis até onde me encontro hoje e me levou até pessoas especiais que possibilitaram e incentivaram a finalização desse trabalho. Obrigado à minha esposa maravilhosa, Maura, que foi a maior incentivadora para a conclusão deste trabalho. Mesmo quando eu me desanimei ela me estimulou. Agradeço aos meus pais, irmão e avó quem me aturaram, inventivaram, apoiaram psicologicamente e financeiramente durante tantos anos. E por fim, obrigados aos meus sogros pelo constante encorajamento!

AGRADECIMENTOS

Agradeço à Universidade Federal do Rio de Janeiro (UFRJ) e seus professores por todo conhecimento transmitidos. Agradeço também à Empresa Júnior de Computação(EJCM) pela experiência profissional propiciada. Ambas as intituições impactaram profundamente minha carreira e vida.

"If you think education is expensive, try ignorance."

Jeff Rich

RESUMO

A evasão no ensino superior é um problema que afeta tanto universidades públicas quanto privadas. Essa evasão representa infraestrutura, professores e funcionários subutilizados e, por isso, é interesse das universidades desenvolver programas e técnicas voltadas a diminuição desse índice. Este trabalho tem por objetivo utilizar a técnica de aprendizado de máquina conhecida como árvore de decisão para ajudar na identificação de estudantes do Bacharelado em Ciência da Computação da UFRJ com maior chance de evadir e traçar um perfil dos mesmos. Foram gerados quatro conjuntos de dados a partir das informações disponíveis em quatro períodos diferentes de tempo, com o objetivo de verificar se as características dos estudantes que evadem mudam dependendo do momento que a evasão ocorre. As árvores de decisão resultantes tiveram bom desempenho, identificando de 70 a 85% dos alunos evadidos dependendo do conjunto utilizado. O perfil traçado identificou os principais atributos dos alunos evasores possibilitando que orientadores e professores possam atuar antes que a evasão de fato ocorra.

Palavras-chave: inteligência artificial; aprendizado de máquina; mineração de dados; evasão no ensino superior; evasão; UFRJ.

ABSTRACT

Dropout rates in higher education are a problem that affects both public and private universities. It represents underutilized infrastructure, teachers, and staff, and therefore it is in the interest of universities to develop programs and techniques aimed at reducing this index. This work aims to use the machine learning technique known as decision tree to help identify Computer Science students at UFRJ with a higher chance of dropping out and to profile them. Four data sets were generated from the information available at four different periods of time, with the aim of verifying whether the characteristics of students who drop out change depending on when the dropout occurs. The resulting decision trees performed well, identifying 70 to 85% of the dropout students depending on the dataset used. The profile identified the main traits of dropout students, enabling advisors and teachers to act before the dropout actually occurs.

Keywords: artificial intelligence; evasion; data mining;drop out in higher education; machine learning; UFRJ.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

UFRJ	Universidade Federal do Rio de Janeiro
IFSC	Instituto Federal de Santa Catarina
IES	Instituições de Ensino Superior
EJCM	Empresa Júnior de Computação
IFES	Institutos Federais de Ensino Superior
TCU	Tribunal de Contas da União
TSG	Taxa de Sucesso na Graduação
MEC	Ministério da Educação
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

SUMÁRIO

1 INTRODUÇÃO

Este projeto tem como foco a evasão escolar no contexto do ensino superior, que é um problema não só brasileiro como internacional. Segundo Lobo (2017) a taxa de evasão anual no Brasil se manteve constante no periodo de 2002 à 2017 em torno de 22%, sendo maior para o ensino privado do que para o público. Ela é encontrada dividindo o número de matrículas que foram efetivadas por estudantes já matriculados no ano anterior pelo número de estudantes que poderiam ter se matriculado.

Conforme Filho et al. (2007) explica, a evasão é um fenômeno de causas variadas, as principais sendo: questões de ordem acadêmica, a integração do estudante na instituição de ensino e as expectativas do mesmo sobre a sua formação.

Além de consequências para o próprio aluno e para a sociedade, a evasão tem impacto financeiro grande nas intituições de ensino superior, gerando ociosidade de salas, funcionários e infraestrutura. No setor público isso representa investimento público que não trouxe retorno. Por exemplo, a Universidade de Brasília teve prejuízo financeiro devido à evasão de cerca de 95,6 milhões de reais em 2015 (PINHEIRO, 2015). No setor privado a evasão representa receitas perdidas, e como observado por Pereira (2014), se a mensalidade for de 500 reais e a evasão for de 25%, a perda anual de receitas para cada mil alunos é de 375 mil reais, chegando a 7,5 milhões para uma instituição com 20 mil alunos. Outro aspecto a ser considerado é que a evasão pode variar de acordo com a cidade, região ou IES e que devido a essas particularidades é necessário adaptar os modelos e pesquisas para cada contexto (SANTOS, 2021).

A relevância e impacto financeiro deste tema têm levado alguns Institutos de Educação Superior (IES) a buscar novas formas de acompanhar os alunos, um delas sendo a utilização de sistemas inteligentes que permitem identificar necessidades e carências dos estudantes e separá-los em grupo de risco de evasão.

No trabalho de conclusão de curso, Santos (2021) teve como objetivo predizer alunos com potencial de evasão nos cursos de graduação no Instituto Federal de Santa Catarina (IFSC), Campus Caçador. Para tal foram utilizadas as técnicas de aprendizado de máquina de árvore de decisão e redes neurais. Existem apenas dois cursos no campus Caçador, Sistemas de Informação e Engenharia de Produção, e ambos foram analisados, abrangendo dados de 380 alunos analisados. O autor não detalhou o período analisado. Foi utilizado como entrada para o aprendizado atributos socioeconômicos tais como idade, raça e sexo, e também atributos acadêmicos, que são aqueles baseados no desempenho do aluno durante a universidade. O aluno era considerado "evadido"caso sua matrícula estivesse trancada, desativada ou cancelada. O resultado encontrado tanto em acurácia quanto em precisão foi ao redor de 80%, com desempenho ligeiramente superior para a árvore de decisão.

Dutra (2015) tem objetivo semelhante ao trabalho anterior ao criar um método para previsão precoce da evasão de um aluno atráves de técnicas de redes neurais. O foco desse trabalho é exclusivo nos alunos de graduação da Rede Doctum de ensino, uma rede privada. Não foram listados quais cursos de graduação abrangidos pelos dados, porém foi apontado que foram utilizados dados a partir de 2006 totalizando 89000 alunos analisados.

O autor utilizou como entrada nos algoritmos dados semelhantes a Santos (2021) com a adição de dados financeiros como, por exemplo, a taxa de inadimplência e o valor da matrícula na universidade . O aluno foi considerado evadido quando sua matrícula estava trancada, cancelada ou quando pediu transferência, caso contrário é considerado como não evadido. O resultado encontrado foi uma taxa de acurácia de cerca de 90%, porém a taxa de acertos na classe dos alunos que evadiram foi baixa, próxima aos 50%. Além disso, observou-se que os atributos com mais peso ao predizer se o aluno irá evadir ou concluir o curso foram, em ordem descrescente: curso do discente, valor da mensalidade, nota alcançada na primeira etapa, média semestral e período.

Em Junior (2018) foram utilizados dados de 196 alunos do curso de Sistemas de Informação da Universidade Federal do Rio Grande do Norte para analisar quais são os atributos relevantes para evasão, identificar se as disciplinas iniciais do curso tem influência na evasão e verificar perfis de alunos mais provavéis de evadir. Os dados se referem aos alunos concluintes e evadidos entre 2011 e 2018, não havendo dados de alunos ativos no curso ou com matrícula trancada. O autor utilizou o algoritmo de árvore de decisão e concluiu que o local de moradia, o recebimento de algum tipo de auxílio e a quantidade de matérias cursadas por período não influenciam a evasão. Porém a idade, a participação em projeto de pesquisa ou extensão e reprovação nas disciplinas que o autor considera base do curso estão sim relacionados a taxa de evasão. O perfil dos alunos que evadem do curso é composto pelas seguintes características:

1. Os que reprovaram nas 4 disciplinas básicas do curso
2. Aqueles com mais de 26 anos
3. Os que excederam os 8 semestres normais do curso e não estão matriculados em nenhum projeto
4. Quem reprovou a disciplina de *introdução e informática* seguida pela reprovação na disciplina de algoritmos.

Como objetivo geral deste trabalho, pretendemos verificar a possibilidade de utilizar a técnica de aprendizado de máquina chamada de árvore de decisão para identificar características que possam indicar alunos com maior propensão à evasão dentro do curso de Ciências da Computação da Universidade Federal do Rio de Janeiro(UFRJ). As decisões de modelagem do presente trabalho diferem em alguns aspectos das tomadas pelos autores dos demais trabalhos apresentados. Optou-se por utilizar apenas o algoritmo de

árvore de decisão de forma a poder investigar quais atributos impactam mais a evasão. Foram excluídos dados de alunos com matrícula ainda ativa ou trancada porque não se sabe se esses alunos vão evadir ou concluir o curso, mantendo-se apenas alunos que de fato evadiram ou concluiram o curso. Os dados utilizados foram socioeconômicos, tais como sexo e curso de ingresso, e acadêmicos tais como o coeficiente de rendimento, matérias cursadas e notas obtidas. Além disso, foram gerados quatro conjuntos de dados a partir das informações disponíveis em quatro períodos diferentes de tempo, com o objetivo de verificar se as características dos estudantes que evadem mudam dependendo do momento que a evasão ocorre.

As árvores de decisão resultantes tiveram bom desempenho, identificando de 70 a 85% dos alunos evadidos dependendo do conjunto utilizado. O perfil obtido identificou os principais traços dos alunos evasores possibilitando que orientadores e professores possam atuar antes que a evasão de fato ocorra.

O trabalho está organizado da seguinte forma: no capítulo 2 são apresentados os fundamentos teóricos para evasão e retenção e no capítulo 3 é mostrada a fundamentação teórica da mineração de dados. A metodologia está descrita no capítulo 4 e no capítulo 5 apresentamos os experimentos realizados e resultados obtidos. Por fim, no capítulo 6 temos a conclusão.

2 FUNDAMENTAÇÃO TEÓRICA: EVASÃO NO ENSINO SUPERIOR

Conforme Lobo (2012) informa, a evasão é um dos maiores problemas em qualquer nível de ensino no Brasil, incluindo o ensino superior seja ele público ou privado. O fato de alunos não concluírem seus estudos representa uma perda social, de recursos e de tempo para todos os participantes do processo educacional.

No setor público e no setor privado a evasão é uma considerável fonte de ociosidade de professores, funcionários, equipamentos e espaço físico. A falta de indivíduos com formação completa representa uma perda tanto para a sociedade quanto para o país, visto que o número de pessoas com formação completa será menor do que poderia ter sido alcançado. Apesar disso, Lobo (2017) explica que são raros os programas institucionais profissionalizados para o combate a evasão nas IES brasileiras e que a disparidade entre o gasto com marketing no setor privado, entre 2% e 6% da receita, e o gasto para manter os estudantes matriculados é grande.

Existem diversos tipos de evasão. A *evasão da instituição de ensino* se trata da evasão na qual o aluno deixa uma IES e vai para outra, não abandonando o sistema de ensino superior. A *evasão de sistema* é aquela em que o aluno deixa de estudar e abandona o sistema de ensino superior, excluindo-se os alunos concluíntes. A *evasão de curso*, por sua vez, é aquela em que o aluno deixa um curso de ensino superior por qualquer razão, podendo ter como destino um outro curso na mesma na mesma IES, ir para outro curso em outra IES e até abandonar totalmente os estudos universitários. Pode se notar que toda evasão de instituição de ensino e de sistema é também uma evasão de curso. As distinções entre os tipos de evasão são importantes já que o abandono do curso sem o abandono de instituição ou do sistema pode ser considerado um caso de mobilidade acadêmica e não de evasão.

Sobre a evasão da instituição de ensino, Lobo (2012) informa que um dos maiores problemas é a tendência de colocar as questões financeiras do aluno como sendo a principal ou única responsável pela mudança de IES. Isso leva a ignorar outras possíveis causas como, por exemplo, problemas de atendimento ao aluno de origem pedagógica ou administrativa . O mesmo autor informa que a evasão de sistema é a que precisa de politicas públicas que vão além de questões institucionais e acadêmicas das IES.

Existem diversas formas de medir a evasão e é importante escolher as medidas mais adequadas para cada situação, considerando o contexto específico da instituição de ensino superior em questão. A discussão de algumas das formas de mensurar a evasão segue abaixo.

Conforme Lima et al. (2019) explica, nos Institutos Federais de Ensino Superior (IFES) estão em vigência os indicadores de desempenho de gestão propostos pelo Tribunal de Contas da União (TCU) em conjunto com a Secretaria Federal de Controle Interno e a

Secretaria de Educação Superior do Ministério da Educação (MEC), entre os quais se destaca Taxa de Sucesso na Graduação (TSG). Esse indicador (Equação 2.1) é definido como a razão entre o número de diplomados e o número de ingressantes , ajustados pelo ano em que esses alunos ingressaram na instituição e por um tempo de permanência k fixado pelo MEC:

$$TSG(n) = \frac{quantidade\ de\ diplomados\ no\ ano\ n}{quantidade\ de\ ingressantes\ no\ ano\ (n\ -\ k)} \tag{2.1}$$

em que n é o ano de exercício e $k = 4$, 5 ou 6 é o número de anos previsto de duração do curso. Segundo orientações da União [1] , para calcular o numerador da Equação 2.1 deve-se considerar a quantidade de alunos concluintes do curso no ano letivo correspondente ao exercício, sendo que o número de concluintes nos dois semestres do ano deve ser somado. Um aluno concluinte é aquele que completou os créditos mesmo que não tenha colado grau.

Para o cálculo do denominador da Equação 2.1 deve ser levado em conta o ano em que os estudantes que se graduam no exercício supostamente ingressaram na universidade, com base na duração padrão prevista para cada curso. Portanto, para cursos que duram 4 anos deve ser contabilizada a quantidade de ingressantes quatro anos letivos atrás e de forma análoga para cursos de 5 e 6 anos de duração.

Em cursos em que ocorre ingresso de novos alunos semestralmente o cálculo deve levar em conta os dois semestres de ingresso que correspondem aos dois semestres de graduação do exercício. Seja por exemplo o ano de exercício 2016 em um curso com duração de 4 anos. Os alunos que se formaram no primeiro semestre de 2016 supostamente entraram no segundo semestre de 2012 e os alunos que se formaram no segundo semestre de 2016 ingressaram no primeiro semestre de 2013, logo, devemos somar os ingressantes desses dois períodos. Para um curso de 5 anos deveriamos somar os ingressantes do segundo semestre de 2011 e primeiro semestre de 2012.

Utilizando os dados fornecidos pelo Sistema Integrado de Gestão Acadêmica (SIGA) foi possível calcular a TSG do exercício 2016 para o Bacharelado em Ciências da Computação da UFRJ. Como o curso tem duração esperada de 4,5 anos optou-se por calcular a TSG com os valores de k=4 e k=5. O número de diplomados no ano de 2016 foi de 36 alunos. Para k=4, a quantidade de ingressantes no segundo semestre de 2012 e primeiro semestre de 2013 foi respectivamente 44 e 73 totalizando 117. A TSG portanto é:

$$TSG(2016) = \frac{36}{117} = 0,3077 \tag{2.2}$$

[1] http://portal.mec.gov.br/setec/arquivos/pdf/indicadores.pdf

Para k=5, o número de diplomados não muda e a quantidade de ingressantes no segundo semestre de 2011 e primeiro semestre de 2012 foi respectivamente 47 e 59 totalizando 106. A TSG portanto é:

$$TSG(2016) = \frac{36}{106} = 0,3396 \tag{2.3}$$

Os dados disponibilizados pelo SIGA não são padronizados o que não é ideal para a realização de comparações da TSG entre vários anos e/ou cursos, um problema também descrito por Salles e Silva (2021). Dados melhor estruturados para esse tipo de análise podem ser encontrados no Censo da Educação Superior, realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). A razão para não terem sido utilizados os dados do Censo no calculo da TSG é que os ingressantes não estavam divididos por período, impossibilitando a análise para cursos semestrais.

A TSG, porém, é uma métrica bastante criticada e os principais questionamentos são que o valor dela diminui quando há expansão de vagas na IFES e que a TSG conta a mobilidade acadêmica numa mesma instituição como sendo evasão o que leva à diminuição da taxa. Uma explicação mais detalhada do problema dessa métrica se encontra em Lima et al. (2019).

Uma medida, detalhada em Lobo (2012), que pode ser feita é a evasão anual média que mensura o percentual dos alunos matriculados que não se inscreveram no ano seguinte na IES, excluindo-se aqueles que se formaram. Por exemplo, caso 80 alunos renovassem suas matrículas de um total de 100 que poderiam ter feito então a evasão anual média no curso seria de 20%. O cálculo do percentual da evasão média referente ao ano n é dado por:

$$E(n) = 1 - [M(n) - I(n)]/[M(n-1) - C(n-1)] \tag{2.4}$$

Onde:

- E(n) é evasão no ano n
- M(n) é o número de matrículas no ano n
- C(n) é o número de concluintes no ano n
- I(n) é o número de ingressantes ano n,

- n é o ano em estudo
- (n-1) é o ano anterior

Lobo (2012) também propõe um indicador denominado de evasão total, que avalia a quantidade de alunos que, após ingressarem em um curso, IES ou sistema de ensino, não obtiveram o diploma dentro de um determinado período de anos. Por exemplo, se 200 estudantes entraram em um curso em um determinado ano e 106 se formaram, o índice de titulação é de 54% e a evasão nesse curso é de 46%.

3 FUNDAMENTAÇÃO TEÓRICA: MINERAÇÃO DE DADOS

Segundo Han, Kamber e Pei (2012) a mineração de dados pode ser definida como o processo de descoberta de conhecimento e padrões de interesse a partir de grandes quantidades de dados. Esse processo de forma geral consiste nas seguintes etapas, em ordem:

- Limpeza dos dados, para remoção de dados inconsistentes e de ruído.
- Integração dos dados, onde múltiplas fontes de dados podem ser combinadas.
- Seleção dos dados, escolhendo apenas os dados relevantes na base de dados.
- Transformação de dados, onde os dados são transformados e consolidados para formas mais apropriadas para a mineração.
- Mineração, nessa etapa essencial, métodos inteligentes são aplicados aos dados transformados com o objetivo extrair padrões.
- Avaliação de padrões, para identificar aqueles que realmente são interessantes.
- Apresentação do conhecimento obtido à outras pessoas.

A etapa de mineração está conectada a área de aprendizado de máquina pois, conforme explicado por Kelleher, MacNamee e D'Arcy (2015), o aprendizado de máquina é um processo automatizado de extração de padrões dos dados.

Um tipo de aprendizado de máquina é o aprendizado supervisionado, nesse tipo o modelo aprende a relação entre um conjunto de atributos descritivos e um atributo ou classe alvo baseado em dados históricos. O modelo treinado então é capaz de prever o atributo alvo quando se deparar com novos atributos descritivos, não disponibilizados anteriormente no histórico.

3.1 ÁRVORE DE DECISÃO

Uma árvore de decisão, segundo Russell e Norvig (2010), é uma função que recebe um vetor de valores como entrada e retorna um único valor, em geral a predição da classe alvo desejada. A predição é encontrada atráves da realização de uma sequência de testes. Cada vértice interno representa um teste em um atributo de entrada, os ramos que saem de cada vértice são os possíveis valores daquele atributo. Um vértice sem nenhum ramo é chamado folha e representa o valor retornado na função.

Com o objetivo de possibilitar um melhor entendimento, criaremos uma árvore de decisão a partir dos dados no Quadro 1. O objetivo é criar uma árvore que utilize os

valores dos atributos `Gosta de Pipoca` e `Gosta de Refrigerante` para prever o valor do atributo `Gosta do filme Pantera Negra`. Cada linha representa os gostos de uma pessoa diferente, o atributo `id` serve como identificador dessa pessoa. Por exemplo, a pessoa com valor 1 no atributo id gosta de pipoca, refrigerante e não gosta do filme Pantera Negra.

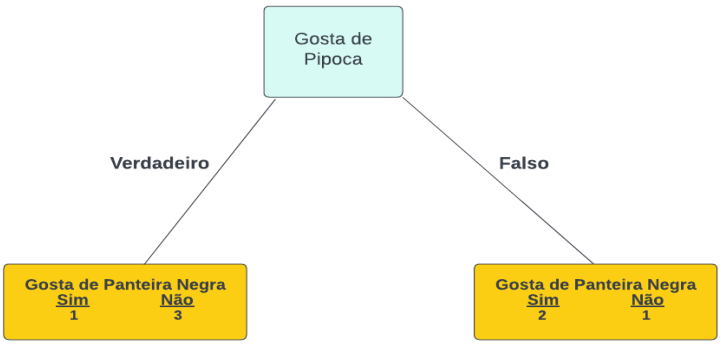
Quadro 1 – Dados a partir dos quais será criada árvore de decisão

id	Gosta de Pipoca	Gosta de Refrigerante	Gosta do filme Pantera Negra
1	Sim	Sim	Não
2	Sim	Não	Não
3	Não	Sim	Sim
4	Não	Sim	Sim
5	Sim	Sim	Sim
6	Sim	Não	Não
7	Não	Não	Não

Fonte: Própria

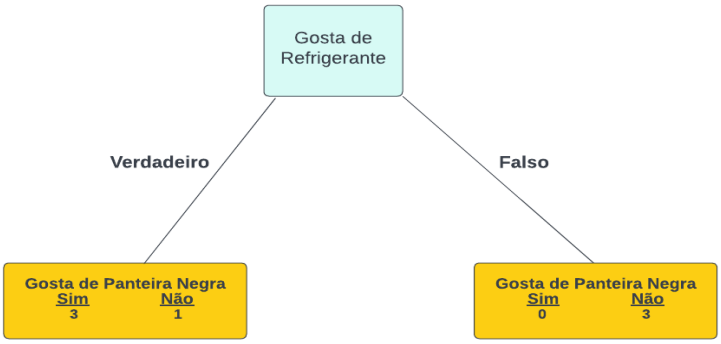
A primeira coisa a fazer é definir qual será a raiz da árvore, ou seja, qual será a primeira pergunta a ser feita dentre as possibilidades: perguntar se a pessoa gosta de pipoca ou perguntar se gosta de refrigerante. Para essa escolha devemos ver o quanto cada pergunta, ou atributo, prediz se uma pessoa gosta ou não do filme Pantera Negra. Começando pelo atributo `Gosta de Pipoca` , criamos uma árvore de decisão simples que utiliza apenas ele para predizer se a pessoa gosta ou não do filme. Essa árvore está exibida na Figura 1, a folha da esquerda contém todas as pessoas que gostam de pipoca, das quais uma gosta do filme e três não gostam. A folha da direita por sua vez contém todas as pessoas que não gostam de pipoca, duas dessas pessoas gostam do filme e uma delas não gosta. Realizando o mesmo procedimento para o atributo `Gosta de refrigerante` obtemos a árvore exibida na Figura 2.

Figura 1 – Árvore de Decisão para atributo pipoca



Fonte: Própria

Figura 2 – Árvore de Decisão para atributo refrigerante



Fonte: Própria

Considerando as duas árvores é possível notar que nenhuma das duas consegue predizer com perfeição quem gosta e não gosta do filme. Todas as folhas do atributo `Gosta de Pipoca` e a folha da esquerda do atributo `Gosta de Refrigerante` possuem misturas de pessoas que gostam e não gostam do filme e, por isso, essas folhas são chamadas de impuras. Como ambas as folhas do atributo `Gosta de Pipoca` são impuras enquanto apenas uma do atributo `Gosta de Refrigerante` o é, parece que o segundo faz um trabalho melhor ao predizer o valor do atributo `Gosta do filme Pantera Negra`. Conforme James et al. (2013) informa, é possível quantificar a impureza em cada folha e o mesmo autor indica que isso é tipicamente feito utilizando ou a entropia ou o coeficiente Gini, sendo este último bastante utilizado porque é menos custoso computacionalmente. Por causa disso e pelo fato de que a forma de cálculo dessas medidas é similar, optou-se por utilizar somente o critério de Gini. Mais detalhes sobre o cálculo da entropia podem ser vistos em James et al. (2013). Primeiramente devemos calcular a impureza de cada folha e depois a impureza do atributo. A fórmula para o cálculo da impureza Gini para uma folha é:

$$Impureza\ Gini\ para\ folha = 1 - (probabilidade\ de\ "Sim")^2 - (probabilidade\ de\ "Nao")^2$$

Começando pela impureza na folha à esquerda no atributo `Gosta de Pipoca`. A probabilidade de "Sim"é a probabilidade das pessoas gostarem do filme dado que elas gostam de pipoca. Isso é igual ao número de pessoas que gostam ao mesmo tempo de pipoca e do filme dividido pelo número total de pessoas que gostam de pipoca, realizando essa conta obtemos $1/4$. A probabilidade de "Não"é a probabilidade das pessoas não gostarem do filme dado que elas gostam de pipoca, o resultado desse cáclculo é $3/4$. Substituindo na equação obtemos:

$$Impureza\ Gini\ na\ folha\ esquerda = 1 - (1/4)^2 - (3/4)^2 = 0,375$$

De forma análoga, obtemos a impureza na folha direita.

$$Impureza\ Gini\ na\ folha\ direita = 1 - (2/3)^2 - (1/3)^2 = 0,444$$

Pode se perceber que a folha à esquerda representa um total de 4 pessoas, 3 pessoas que não gostam e 1 pessoa que gosta do filme. A folha à direita por sua vez representa um total de 3 pessoas. Portanto, ao calcular a impureza Gini total do atributo deve-se levar em conta o peso de cada folha. Logo a impureza Gini total é a média ponderada da inpureza das folhas. O peso da folha à esquerda é o número de pessoas nessa folha(4), dividido pelo número total de pessoas em ambas as folhas (4+3)=7, procedimento análogo é realizado na folha à direita, onde temos 3 pessoas e dividiremos por 7 . Com o valor dos pesos já encontrados podemos prosseguir para encontrar a impureza total do atributo.

$$Impureza\ Gini\ total\ do\ atributo = \frac{4}{4+3} * 0,375 - \frac{3}{4+3} * 0,444 = 0,405$$

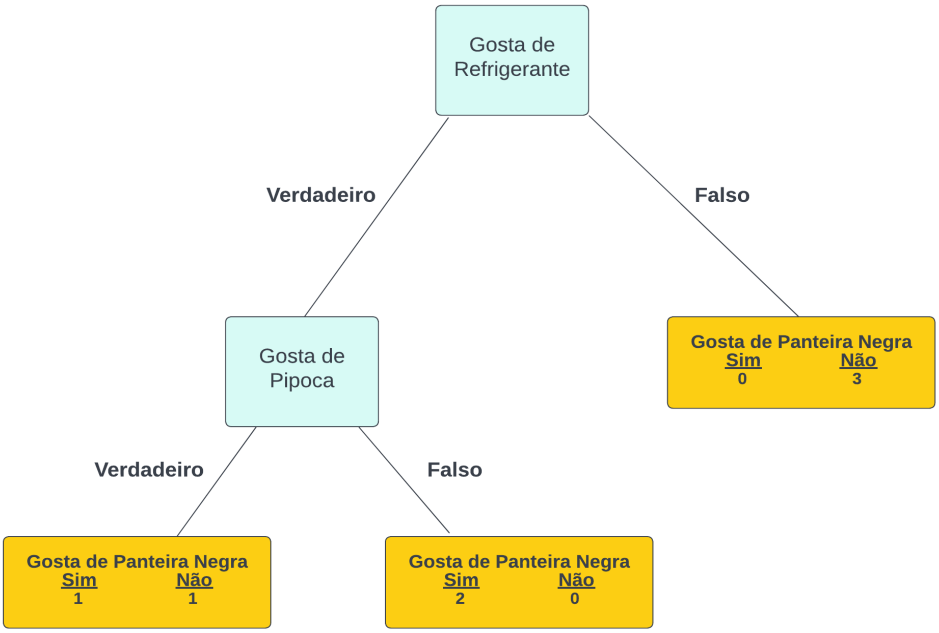
Portanto, o valor da impureza Gini para o atributo `Gosta de Pipoca` é 0,405. Fazendo o mesmo processo para o atributo `Gosta de Refrigerante` obtém-se impureza Gini de 0,214. Como a impureza total é menor neste último, é ele o escolhido para ser a primeira pergunta, a raiz da árvore. O próximo passo seria repetir esse processo todo para as folhas geradas a partir da raiz utilizando os atributos remanescentes, ou seja, iriamos ver qual o atributo restante com menor impureza em relação a folha esquerda e este atributo será o próximo teste a ser realizado nessa folha. Na folha direita não seria necessária mais nenhuma divisão ou teste pois já separa perfeitamente quem gosta e não gosta do filme. Ao final a árvore obtida é mostrada na Figura 3.

Dentre diversas possibilidades para algoritmos de aprendizado foi escolhido a árvore de decisão como classificador nesse trabalho, sendo o principal motivo a explicabilidade do modelo. Conforme informa Burkov (2019), a maioria dos algoritmos de aprendizagem de alta acurácia são "caixas pretas", ou seja, cometem poucos erros porém entender e explicar o por que de determinada predição é bem difícil. A árvore de decisão possui uma acurácia razoável ao mesmo tempo que é fácil entender quais atributos são responsáveis pelas predições, possibilitando a criação de um perfil do aluno com potencial de evasão.

3.2 TREINO, VALIDAÇÃO E TESTE

São utilizados três conjuntos de dados diferentes durante a construção e avaliação do modelo: de *treino*, de *teste* e de *validação*. O conjunto de treino é geralmente o maior deles e é utilizado para construir o modelo. Os conjuntos de validação e teste são mais ou menos do mesmo tamanho, e bem menores que o conjunto de treino. Não existe uma

Figura 3 – Árvore de Decisão obtida ao final



Fonte: Própria

proporção exata entre o tamanho desses três conjuntos porém Taboga (2021) menciona algumas escolhas populares:

- Treino: 60% , teste: 20% , validação: 20%.
- Treino: 70% , teste: 15% , validação: 15%.
- Treino: 80% , teste: 10% , validação: 10%.

Burkov (2019) explica o motivo de existirem três conjuntos de dados e não apenas um. Ocorre que um algoritmo que memoriza os dados de treino e depois usa essa memória para acertar todos os rótulos desse mesmo conjunto não tem utilidade. Esse caso aconteceria se utilizassemos um mesmo conjunto de dados para o treino e para a medição do desempenho. Queremos na verdade que o modelo possa predizer rótulos em exemplos ainda não vistos. Por isso usamos um conjunto de dados para treinar o algoritmo e mantemos separados os outros.

O conjunto de validação é utilizado para escolher o algoritmo de aprendizado e também para seleção dos hiperparâmetros que são parâmetros cujos valores são utilizados para controlar o processo de aprendizado. O hiperparâmetro neste trabalho foi a profundidade da árvore, definido pelo tamanho do maior caminho da raiz até uma folha ou então, de forma equivalente, o maior número de perguntas feitas até chegar em uma folha. O

conjunto de teste é utilizado para verificar o desempenho do modelo. Uma sequência de passos em como os dados são utilizados está disponibilizada abaixo:

1. Conjunto de treino é utilizado para treinar alguns modelos candidatos.
2. Conjunto de validação é utilizar para avaliar os modelos candidatos
3. Um dos candidatos é escolhido.
4. O modelo escolhido é treinado com um novo conjunto de dados de treino.
5. O modelo treinado é avaliado com o conjunto de dados de teste obtendo o desempenho final do modelo.

Ao invés de, nos passos 1 e 2, avaliar uma única vez cada modelo candidato, o melhor é avaliar cada um diversas vezes com diferentes conjuntos de dados e utilizar a média do desempenho para tomar a decisão no passo três. Isso é facilmente feito quando há uma grande quantidade de dados. Porém quando esse não é o caso podemos usar a validação cruzada. Confome Brownlee (2021) explica, essa técnica permite utilizar um mesmo conjunto de dados diversas vezes fingindo que eles são diferentes.

Na prática, pode ser utilizado o conjunto de treino para a validação cruzada. Nesse processo dividimos o conjunto de dados de treinamento em K subconjuntos de igual tamanho. Separamos o primeiro subconjunto para validação e treinamos o modelo nos outros *K-1* subconjuntos. Em seguida, avaliamos as predições do modelo no subconjunto separado. O próximo passo é separar o segundo subconjunto para validação e treinar o modelo nos *K-1* subconjuntos restantes e avaliamos as predições do modelo no segundo subconjunto , repetindo o processo até todos os subconjuntos terem sido utilizados para validação do modelo uma vez. O valor de K é em geral entre cinco ou dez. Ao final temos a média do desempenho de cada modelo.

A partir do resultado da validação cruzada é possível concluir qual modelo é o melhor. Após essa conclusão deve-se treinar o modelo de novo com o conjunto de dados de treino completo, sem validação cruzada. A motivação para tal é que a quantidade de dados não é grande e durante o processo de validação cruzada um subconjunto foi separado para validação e portanto não foi utilizado para treino. Treinando com todos os dados de treino teremos um modelo melhor.

3.2.1 Métricas de desempenho

Nesta seção busca-se mostrar como avaliar o desempenho de um modelo. Existem inúmeras métricas que podem ser utilizadas para tal e nosso foco será nas medidas acurácia, precisão, F1 score e recall. Suponha que um modelo de aprendizado de máquina é treinado para predizer tumores em pacientes e que o conjunto de dados contenha informações sobre 100 pessoas, a matriz de confusão mostrada no Quadro 2 apresenta um

exemplo de como está sendo feita a classificação. Uma matriz de confusão é um sumário dos resultados da predição em um problema de classificação, onde o número de predições corretas e incorretas é quebrado por classe, positiva ou negativa, permitindo entender os erros que o modelo está cometendo.

Quadro 2 – Exemplo de matriz de confusão

		Valor Real	
		Negativo	Positivo
Valor predito	Negativo	60	8
	Positivo	22	10

Fonte: Própria

- **Verdadeiro Positivo(VP):** o modelo prediz corretamente a classe positiva, ou seja, foi predito que a pessoa tinha um tumor e ela realmente tinha. No exemplo acima 10 pessoas que realmente possuem tumor foram preditas como tendo.
- **Verdadeiro Negativo(VN):** o modelo prediz corretamente a classe negativa, ou seja, foi predito que a pessoa não tinha um tumor e ela realmente não tinha. No exemplo acima 60 pessoas que realmente não possuem tumor foram preditas como não tendo.
- **Falso Positivo(FP):** o modelo prediz incorretamente a classe negativa, ou seja, foi predito que a pessoa tinha um tumor porém na realidade ela não tinha. No exemplo acima 22 tiveram predição de tumor porém na realidade não tinham.
- **Falso Negativo(FN):** o modelo prediz incorretamente a classe positiva, ou seja, foi predito que a pessoa não tinha um tumor porém na realidade ela tinha. No exemplo acima 8 pessoas com tumor foram preditas como não tendo.

A partir de VP, VN, FP e FN podemos calcular as seguintes métricas.

- **Precisão:** o percentual de casos verdadeiramente positivos dentre as predições positivas .

  $Precisao = \frac{VP}{VP+FP}$

- **Recall:** o percentual predito positivo dentre todos os casos verdadeiramente positivos.

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

- **Acurácia:** número de classificações realizadas corretamente em ambas as classes, positiva e negativa.

  $Acuracia = \frac{VP+VN}{VP+FN+VN+FP}$

- **F1 Score:** é a média harmônica entre a precisão e o recall, levando em conta tanto falsos negativos quanto falsos positivos.

  $f1\ score = \frac{2*(Precisao*Recall)}{Precisao+Recall}$

O recall e o F1 score foram as métricas selecionadas como principais neste trabalho. Quanto maior o recall maior o número de alunos evadidos identificados, o que é nosso objetivo. Já o F1 score é adequado para conjuntos de dados desbalanceados, nos quais uma classe possui um número muito maior de observações do que a outra. Esse é o caso de alguns dos conjuntos de dados utilizados, onde o número de alunos concluintes é superior ao de evadidos.

4 METODOLOGIA

Nesta seção é explicado como foram realizadas as etapas de coleta, tratamento e transformação dos dados e quais foram os experimentos realizados. O código criado para a realização dessas etapas foi disponibilizado pelo autor deste trabalho em https://github.com/victoraugustosouza/trabalho_conclusao_curso.

4.1 COLETA

Foram utilizados dados dos alunos da Universidade Federal do Rio de Janeiro (UFRJ) pertencentes ao curso de Ciências da Computação com data de ingresso entre 2000 e 2021 . Esses dados foram fornecidos pelo Sistema Integrado de Gestão Acadêmica (SIGA) da UFRJ formato `csv` e de forma anonimizada.

A base de dados fornecida possuia 2578 instâncias e as informações disponibilizadas para cada aluno podem ser separadas em acadêmicas e socioeconômicas. As informações acadêmicas são aquelas relacionadas à universidade: (i) quais matérias o aluno reprovou e em qual período e nota obtida; (ii) quais períodos o coeficiente de rendimento (CR) ficou abaixo de três; (iii) os períodos cancelados; (iv) os períodos em que houve trancamento de matrícula; (v) o coeficiente de rendimento acumulado (CRA); (vi) a carga horária acumulada; (vii) o número de períodos integralizados; (viii) o valor do CR e do CRA em cada por período cursado; (ix) as disciplinas cursadas com nota final e o período em que foram realizadas; (x) período de ingresso na UFRJ; (xi) curso de ingresso e código do mesmo; (xii) curso atual e código do mesmo; (xiii) período de ingresso no curso atual; (xiv) nota no Exame Nacional do Ensino Médio (Enem); (xv) situação atual da matrícula (cancelada, ativa ou trancada).

Informações socioeconômicas são aquelas que não mudam durante a permanência do aluno na universidade e estão relacionadas a aspectos sociais e demográficos dos indivíduos, são elas: (i) a data de nascimento; (ii) modalidade da cota; (iii) sexo.

4.2 TRATAMENTO DE DADOS

A exploração dos dados fornecidos levou a identificação de inconsistências que se não tratadas poderiam atrapalhar a identificação de padrões e aprendizado dos algoritmos. Além disso, foi necessário filtrar os dados disponibilizados para se ter um melhor resultado na etapa de aprendizado. Esta seção busca esclarecer e explicar os tratamentos realizados.

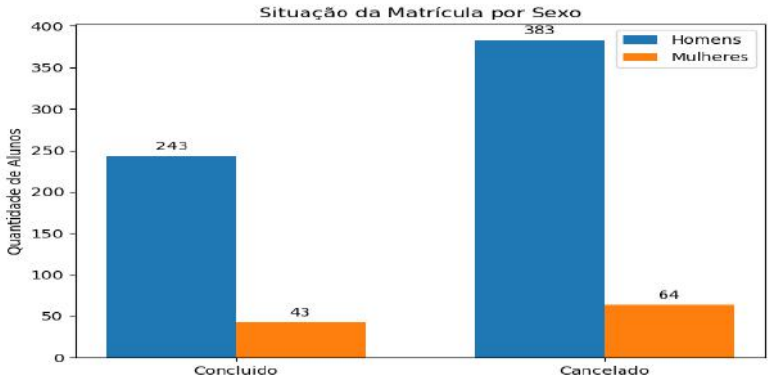
4.2.1 Limpeza e Seleção dos Dados

- **Cancelamento de matrícula antes de alguma disciplina ser feita:** Alguns alunos cancelaram a matrícula antes mesmo de finalizarem uma disciplina. Isso se refletiu nos campos CRA, CR e Disciplinas Cursadas tendo valor nulo. Um total de 128 estudantes se encontravam nessa situação, devido a falta dessas informações cruciais o tratamento utilizado foi remover os dados referentes a esses alunos.
- **Discrepância entre o CR por período e as Disciplinas Cursadas:** Em um único caso o campo CR por período indicava que o aluno havia feito sete períodos porém o campo disciplinas cursadas listava apenas três periodos com matérias feitas. Optou-se por remover esse aluno do conjunto de dados.
- **Alunos Ativos e Trancados:** Foram removidos os alunos com matrícula ainda ativa ou trancada pois não é possível saber se esses alunos vão evadir ou concluir o curso e portanto não podem ser utilizados nos conjuntos de teste e treino.
- **Alunos que entraram antes de 2007:** Houve uma mudança de currículo no ano de 2010 e para manter uma uniformidade nas disciplinas cursadas optou-se por remover os alunos cujo ano de ingresso é inferior a 2007.

4.2.2 Caracterização da Base

Após as etapas descritas anteriormente a base se encontrava com 733 registros de estudantes dos quais 286 eram de concluintes e 447 de alunos que evadiram, distribuídos conforme as Figuras 4 e 5.

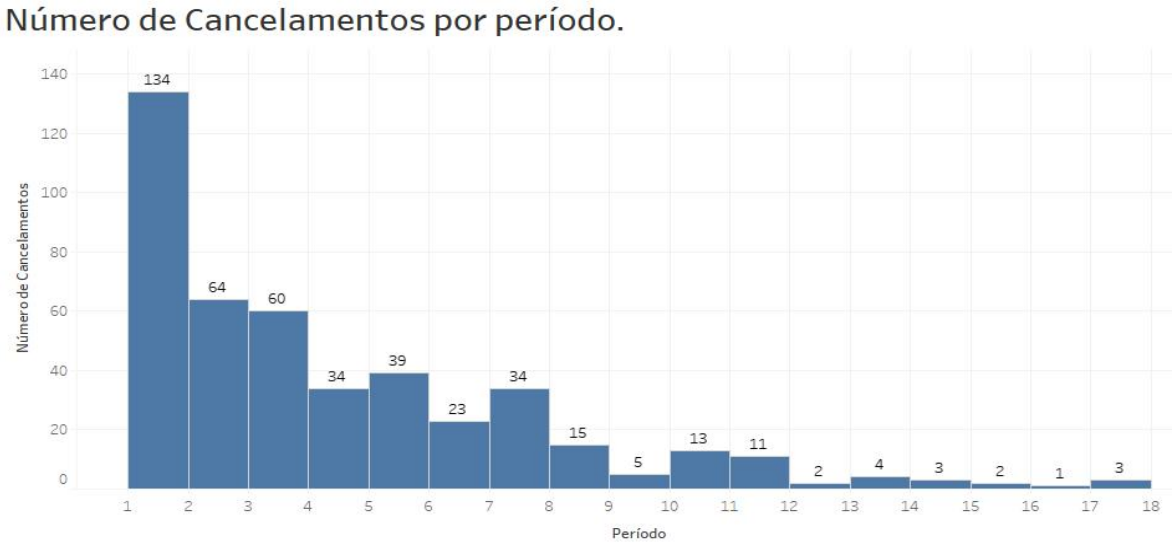
Figura 4 – Situação de Matrícula por Sexo



Fonte: Própria

A Figura 5 indica que dentre os 447 cancelamentos 134 deles ocorreram após o aluno finalizar apenas o primeiro período.

Figura 5 – Situação de Matrícula por Sexo



Fonte: Própria

4.2.3 Modelagem dos Dados

Nesta seção busca-se mostrar as transformações realizadas nos dados com objetivo de obter um melhor desempenho na etapa de aprendizado.

Foram gerados novos atributos a partir das informações originalmente disponibilizadas. São eles: *número de vezes que uma disciplina foi cursada* e *último período em que uma disciplina foi cursada*. Além disso, com o objetivo de identificar alunos propensos à evasão em diferentes momentos do curso foram criados quatro conjuntos de dados diferentes a partir dos dados fornecidos originalmente. O conjunto A é formado por alunos que efetivamente cursaram ao menos três períodos, e será utilizado para identificar alunos propensos a evasão a partir do terceiro período. Analogamente, os conjuntos B, C e D serão formados por alunos que efetivamente cursaram ao menos cinco, sete e um períodos,respectivamente, permitindo identificar alunos propensos a evasão após o quinto, sétimo e primeiro períodos. Embora fosse possível escolher outros períodos para análise, optou-se por manter um intervalo de um ano entre cada período selecionado.

Um aluno que cancela sua matrícula no terceiro período provalvemente faz isso por um motivo diferente daquele que cancela no sétimo, por isso a separação em diferentes conjuntos de dados permite uma melhor identificação das particularidades da evasão em cada período.

Também foi realizado tratamento dos dados para garantir que o conjunto A tivesse apenas os dados que estariam disponíveis até o segundo período. Por exemplo, quaisquer CRs e matérias cursadas do terceiro período em diante não foram incluídas. Isso permite que o modelo tenha resultados melhores já que estará comparando a mesma informações para todos os alunos ao invés de comparar um aluno do sétimo período, que já fez muito

mais matérias, com um aluno do segundo. De forma análoga, foi realizado tratamento dos dados para garantir que o conjunto B tivesse apenas dados que estariam disponíveis até o quarto período, conjunto C até o sexto período e conjunto D até o primeiro período.

Além disso, especialmente no conjunto D foram excluídos alunos que tiveram CR com valor zero no primeiro período. Isso foi feito porque a única forma de ter valor zero no CR é ter tirado zero em todas as matérias cursadas no período, sendo portanto um forte indicador que o aluno desistiu do curso antes do final do primeiro período porém não cancelou a matrícula. Dentre os 447 alunos que evadiram, vistos nas Figuras 4 e 5, 60 tiveram valor zero no CR do primeiro período e foram removidos do conjunto D.

A grade curricular do curso pode ser encontrada no apêndice A.

4.2.3.1 Caracterização da base

Após todas as transformações o número de alunos evadidos e o total em cada conjunto de dados é listado no Quadro 3. Em todos os conjuntos o número de concluintes é 286, que corresponde a diferença entre o número total de alunos e o número de alunos evadidos presentes no Quadro 3.

Quadro 3 – Número de alunos por conjunto de dados

Conjunto de Dados	Evadidos	Total de Alunos
A	249	535
B	155	441
C	93	379
D	387	673

Fonte:Própria

4.2.3.2 Codificação

Conforme Zheng e Casari (2018), uma variável categórica é usada para representar categorias ou rótulos. Um exemplo desse tipo de variável é o atributo *situação atual da matrícula* disponibilizado no conjunto de dados fornecido originalmente. Entretanto o modelo escolhido para mineração de dados foi a árvore de decisão da biblioteca *Scikit-learn* que aceita apenas variáveis númericas. É necessário, portanto, codificar essa informação não númerica em números.

Existem diversas formas de codificar e a utilizada nesse trabalho em todos os campos categóricos foi o *One Hot Encoding* que é mostrada na Figura 6. Para cada valor possível da variável categórica *situacaoMatriculaAtual* é gerado um novo atributo, por exemplo: para o valor *Cancelamento* é criado o atributo *situacaoMatriculaAtual_Cancelamento*. Uma instância cujo o atributo *situacaoMatriculaAtual* tem valor *Cancelamento* terá valor 1 no atributo *situacaoMatriculaAtual_Cancelamento* e 0 nos outros atributos criados.

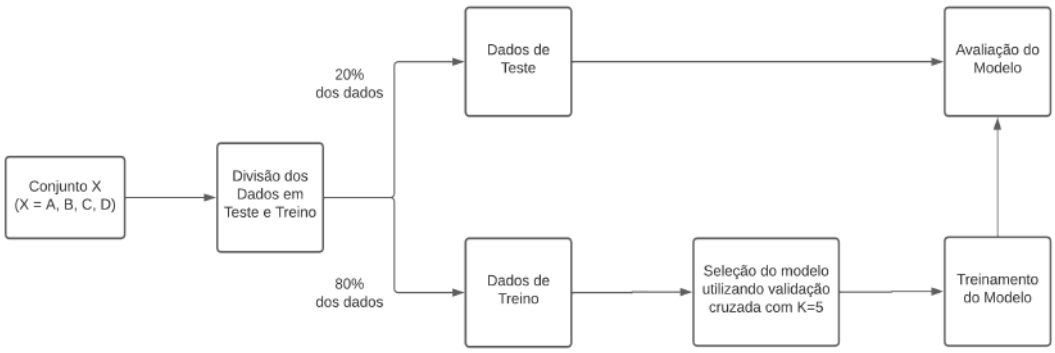
Figura 6 – Exemplo de codificação

Fonte: Própria

4.2.4 Escolha do Classificador e métrica de validação

Devido a facilidade de interpretar o resultado obtido foi escolhido como classificador a árvore de decisão. As métricas F1 score e recall foram selecionadas como principais para análise do desempenho. Partindo dos conjuntos A, B, C e D o procedimento para obtenção do desempenho do modelo foi a mesma e pode ser vista no diagrama abaixo.

Figura 7 – Fluxo para avaliação do modelo



Fonte: Própria

Uma descrição mais detalhada segue abaixo.

1. Foi dividido o conjunto de dados em teste e treino, o primeiro contento 20% dos dados e o segundo contendo 80%.

2. Para a escolha do melhor modelo, foram realizadas validações cruzadas utilizando K=5, um valor bastante utilizado na literatura, com os dados de treinamento. Esse

processo foi realizado com valores de 1 até 20 de profundidade máxima da árvore de decisão e o modelo escolhido foi aquele com maior valor na medida F1 score.

3. O modelo escolhido foi treinado utilizando todos os dados de treino, sem validação cruzada .

4. O modelo escolhido e já treinado é utilizado nos dados de teste, que não haviam sido utilizados anteriormente, resultando nas métricas de desempenho do modelo, das quais a mais importante é o recall.

Além disso, foram analisadas as árvores de decisão produzidas no passo 3 quando o desempenho do modelo foi bom, permitindo a identificação de indicadores do potencial de evasão dos alunos.

5 EXPERIMENTOS E RESULTADOS

Foram desenhados três experimentos com o objetivo de entender se os atributos acadêmicos ou socioeconômicos trariam o melhor desempenho em identificar alunos propensos a evasão. Para cada conjunto de dados A, B, C e D foram realizados os seguintes experimentos:

- **Experimento 1:** alunos propensos à evasão utilizando apenas os dados socioeconômicos.
- **Experimento 2:** alunos propensos à evasão utilizando apenas dados acadêmicos.
- **Experimento 3:** alunos propensos à evasão utilizando tanto dados acadêmicos quanto socioeconômicos.

Os resultados estão consolidados na Figura 8 e as árvores de decisão geradas após o treinamento do modelo estão disponibilizadas no Apêndice 2.

Figura 8 – Consolidação dos Resultados

	Experimento 1						
		Validação			Teste		
	Prof. Máxima	F1	Acurácia	Recall	F1	Acurácia	Recall
A	3	25,8%	58,4%	15,7%	38,7%	64,5%	25,5%
B	3	8,4%	63,6%	15,8%	16,7%	66,3%	10,3%
C	1	2,2%	75,6%	1,2%	0,0%	73,7%	0,0%
D	1	74,9%	59,8%	100,0%	65,0%	48,1%	100,0%
	Experimento 2						
		Validação			Teste		
	Prof. Máxima	F1	Acurácia	Recall	F1	Acurácia	Recall
A	3	78,8%	80,6%	77,2%	81,3%	84,1%	78,7%
B	2	79,7%	86,1%	77,0%	71,7%	83,1%	65,5%
C	3	70,4%	86,1%	69,7%	71,8%	85,5%	70,0%
D	2	77,7%	74,5%	74,5%	79,4%	79,2%	83,1%
	Experimento 3						
		Validação			Teste		
	Prof. Máxima	F1	Acurácia	Recall	F1	Acurácia	Recall
A	3	79,8%	81,3%	79,1%	81,3%	84,1%	78,7%
B	2	79,7%	86,1%	77,0%	71,7%	83,1%	65,5%
C	3	70,3%	86,1%	69,7%	75,5%	85,4%	69,0%
D	4	78,6%	75,1%	77,0%	75,4%	74,8%	80,0%

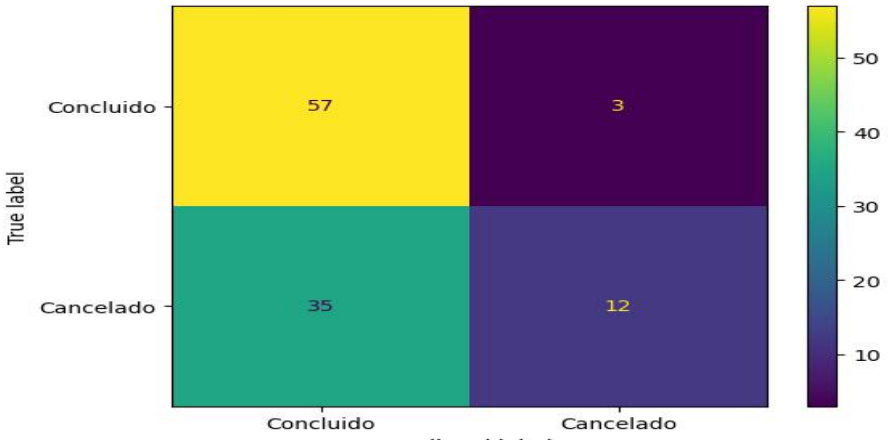
Fonte: Própria

Ao analisar os resultados dos experimentos no conjunto A observamos que o experimento 1 teve o pior resultado em todas as métricas utilizadas. O experimento 1 resultou em um modelo com uma alta chance de predizer que um aluno concluiu o curso mesmo quando o aluno evadiu, isso fica claro na matriz de confusão disponibilizada na Figura 9 onde apenas 12 dos 47 alunos que evadiram foi corretamente classificado enquanto

que o modelo acertou 57 dos 60 alunos que concluiram. Como o número de alunos que concluiram(60) era maior do que o de evadidos(47) no conjunto de teste a acurácia é relativamente alta apesar dos baixos valores de recall e F1 score . O desempenho ruim desse experimento demonstra que não há conexão direta entre dados socioeconômicos de um aluno e a evasão do mesmo. Isso pode ter ocorrido uma vez que temos poucos dados socioeconômicos.

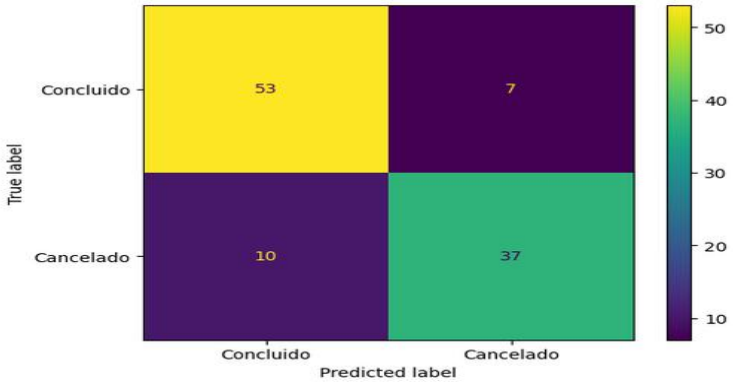
Os experimentos 2 e 3 por sua vez tiveram resultados idênticos e bem melhores que o experimento 1. As matrizes de confusão (Figuras 10 e 11) e as árvores de decisão geradas também foram iguais e contém apenas atributos acadêmicos, inclusive a do experimento 3, que utilizou tanto dados acadêmicos quanto socioeconômicos como entrada para o treinamento. Novamente, a falta de informações socioeconômicas mais robustas não trouxe ganho no desempenho do modelo 3 quando comparado ao modelo 2.

Figura 9 – Matriz de confusão utilizando apenas dados socioeconômicos com conjunto A



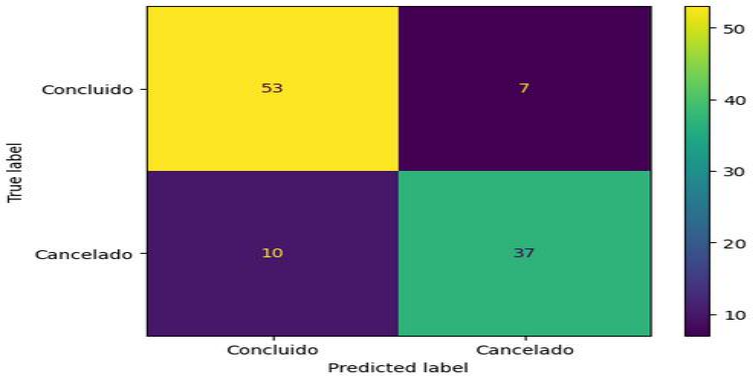
Fonte: Própria

Figura 10 – Matriz de confusão utilizando apenas dados acadêmicos com conjunto A



Fonte: Própria

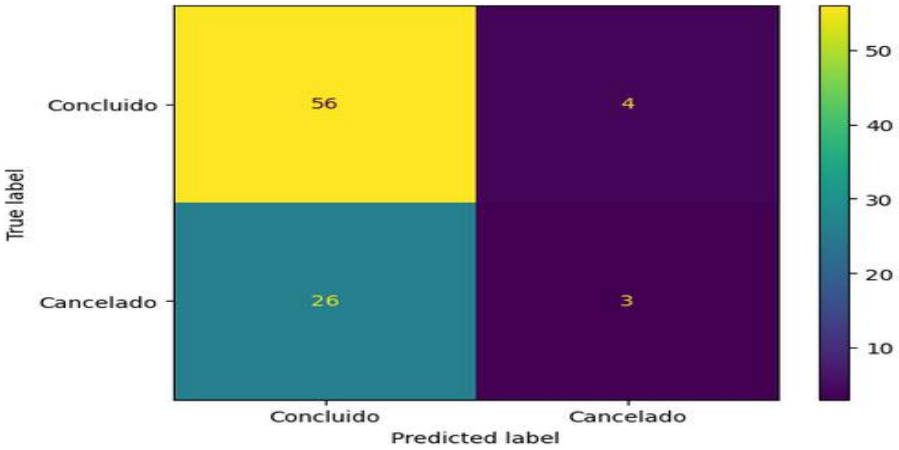
Figura 11 – Matriz de confusão utilizando dados acadêmicos e socioeconômicos com conjunto A



Fonte: Própria

Os experimentos no conjunto B tiveram um padrão parecido com aqueles no conjunto A. Experimentos 2 e 3 com resultados idênticos e bem superiores ao resultado do experimento 1, o modelo gerado neste último classificou erroneamente 26 dos 29 alunos evadidos como tendo concluído o curso. Mais uma vez o desempenho do experimento 2, que contém apenas dados acadêmicos, foi superior.

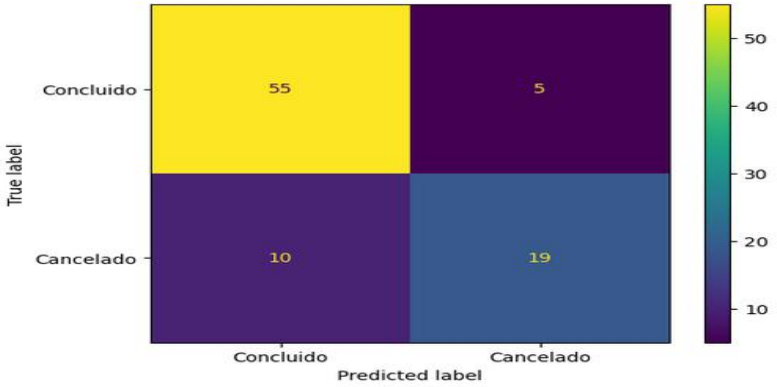
Figura 12 – Matriz de confusão utilizando apenas dados socioeconômicos com conjunto B



Fonte: Própria

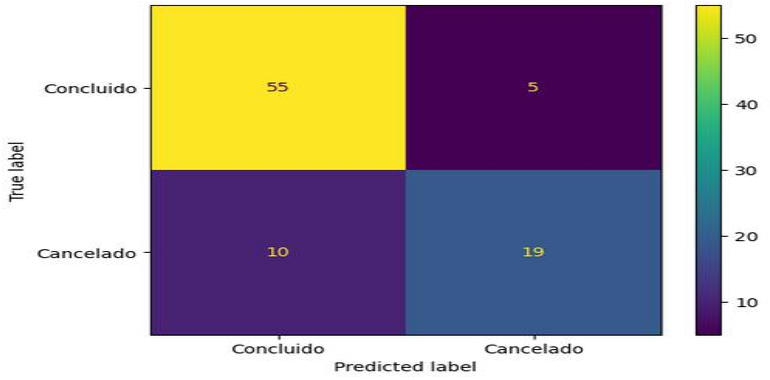
No conjunto C o experimento 1 teve resultado ainda pior que nos outros conjuntos, sendo incapaz de identificar alunos que evadiram. Como pode ser visto na Figura 15 o modelo gerado no experimento 1 classificou absolutamente todos os alunos como tendo concluído e portanto errou todos aqueles que na verdade evadiram, como consequência obteve 0% de recall e F1 score. A acurácia de 73,7% é alta porém se deve ao fato que no conjunto de teste a maioria dos alunos realmente tinha concluído o curso.

Figura 13 – Matriz de confusão utilizando apenas dados acadêmicos com conjunto B



Fonte: Própria

Figura 14 – Matriz de confusão utilizando dados acadêmicos e socioeconômicos com conjunto B

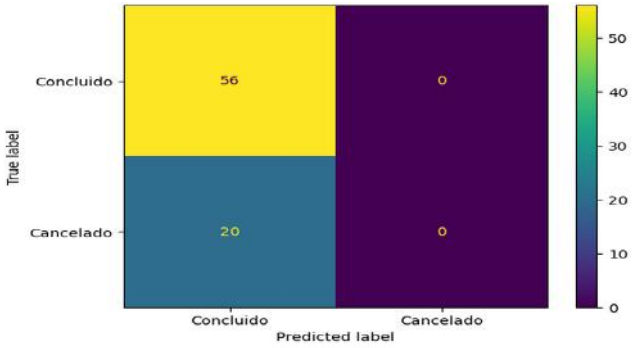


Fonte: Própria

Os experimentos 2 e 3 no conjunto C tiveram resultados diferentes porém muito próximos. O recall é a métrica mais relevante para a identificação dos evadidos e a diferença do experimento 2 para o 3 é de apenas 1%. O experimento 2 ainda é o melhor porque utiliza menos atributos como entrada e gera um resultado praticamente igual ao experimento 3.

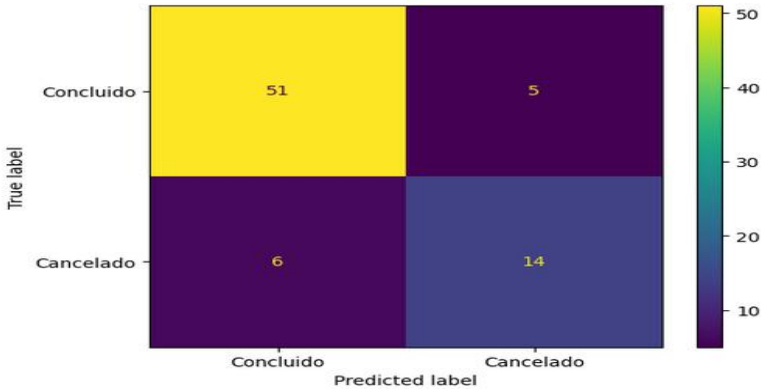
O experimento 1 no conjunto D conseguiu classificar todos os 65 evadidos corretamente obtendo 100% de recall. Examinando atentamente a matriz de confusão da Figura 18 é possível ver que o modelo gerado no experimento 1 classificou absolutamente todos os alunos como tendo evadido e portanto errou as predições de alunos que concluiram o curso. Como a acurácia leva em conta as predições realizadas corretamente em ambas as classes, concluídos e evadidos, o valor dela é mais baixo que o do recall. Esse experimento não foi bom porque não conseguiu separar quem é propenso a evasão e quem não é. Os experimentos 2 e 3 no conjunto D tiveram resultados bem próximos e muito bons, ambos com pelo menos 80% de recall. O experimento 2 teve resultado melhor que o experimento 3 pelos mesmos motivos citados anteriomente.

Figura 15 – Matriz de confusão utilizando apenas dados socioeconômicos com conjunto C



Fonte: Própria

Figura 16 – Matriz de confusão utilizando apenas dados acadêmicos com conjunto C



Fonte: Própria

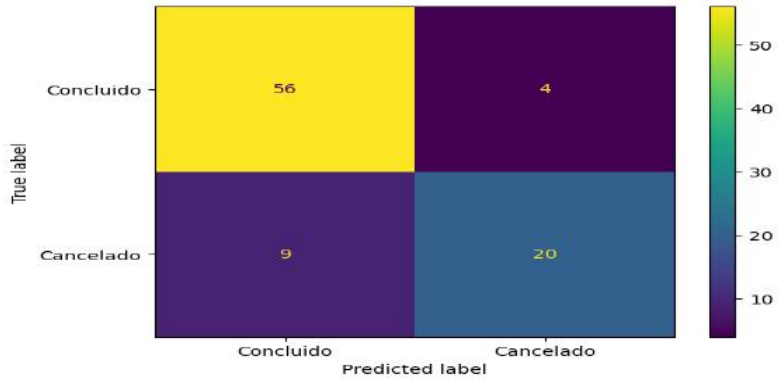
Os bons resultados do experimento 2 demonstram a viabilidade de treinar modelos com dados acadêmicos para identificar alunos com potencial de evasão. Além disso, o fato de que o desempenho foi superior em relação aos experimentos 1 e 3 indica que os dados acadêmicos são mais relevantes do que os dados socioeconômicos que tinhamos disponíveis para identificar esses alunos.

5.1 PERFIL DOS ALUNOS EVADIDOS

Nesta seção é montado um perfil do aluno evadido a partir das árvores de decisão do modelo 2, o melhor modelo testado. O período ao qual cada disciplina pertence pode ser encontrado no Apêndice 1.

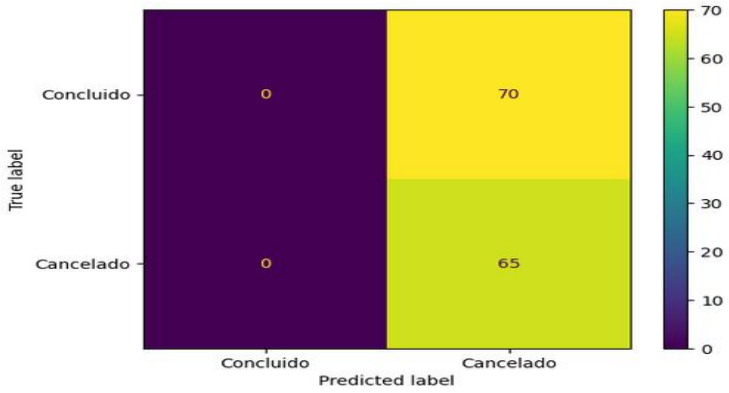
O conjunto A é utilizado para prever a evasão em alunos a partir do terceiro na faculdade e analisando a árvore gerada treinando o modelo 2 com esse conjunto (Figura 32) podemos ver que: (i) a maioria dos evadidos no conjunto de treino (128 alunos do total de 202 que evadiram - 63,4%) tiveram CR no primeiro período menor que 4,9 e

Figura 17 – Matriz de confusão utilizando dados acadêmicos e socioeconômicos com conjunto C



Fonte: Própria

Figura 18 – Matriz de confusão utilizando apenas dados socioeconômicos com conjunto D

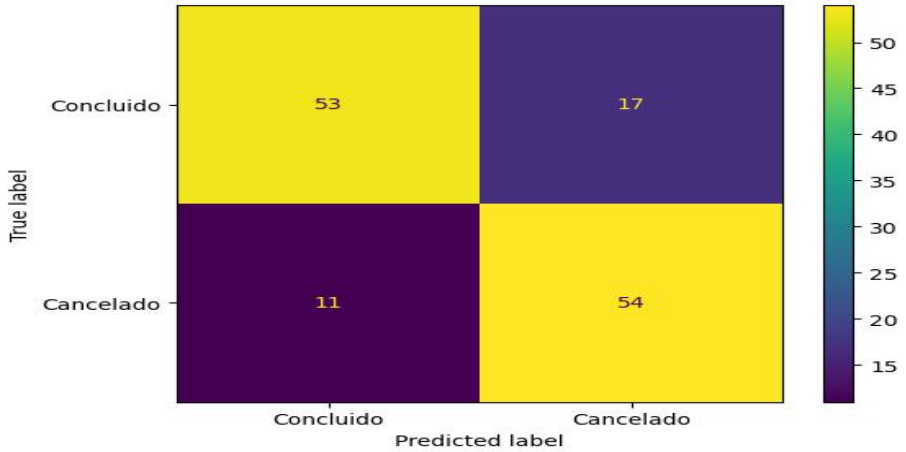


Fonte: Própria

destes 128, boa parte (78) tiveram uma nota na disciplina Fundamentos da Computação Digital inferior a 5,5. Dos alunos aprovados em Fundamentos com nota superior a 5,5, um fator relevante é que eles não cursaram a disciplina Matemática Combinatória (43 de 50). Como esta disciplina possui como pré-requisito a disciplina de Números Inteiros e Criptografia, o resultado pode indicar que o aluno ainda não tenha conseguido aprovação nesta disciplina. Já no caso de alunos com CR do primeiro período maior que 4.9, o CR no segundo período abaixo de 5,4 e uma nota inferior a 5,7 na disciplina Computação I sugerem uma maior possibilidade de evasão.

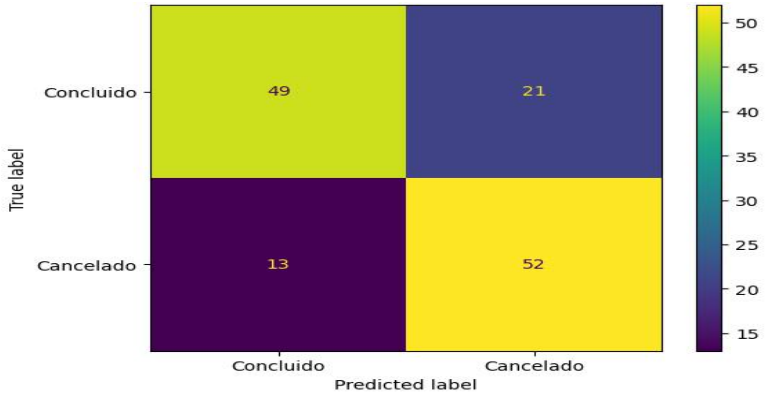
O conjunto B é utilizado para prever a evasão a partir do quinto período na faculdade. Analisando a árvore gerada (Figura 22) pode-se ver que: (i) a maioria dos evadidos (84 de 126 alunos - 66,7%) possui CR no primeiro período menor que 4,9, como no caso do modelo A. Deste, o CR no quarto período inferior a 5,5 se apresentou como um indicador relevante para a evasão (78 dos 84 alunos - 92,9%). Para os alunos com CR do primeiro

Figura 19 – Matriz de confusão utilizando apenas dados acadêmicos com conjunto D



Fonte: Própria

Figura 20 – Matriz de confusão utilizando dados acadêmicos e socioeconômicos com conjunto D



Fonte: Própria

período maior que 4,9 (42 alunos de 126 - 33,3%), ter CR no segundo período inferior a 5,4, mostrou-se um fator importante para identificar a evasão (25 de 42 aluno - 59,5%).

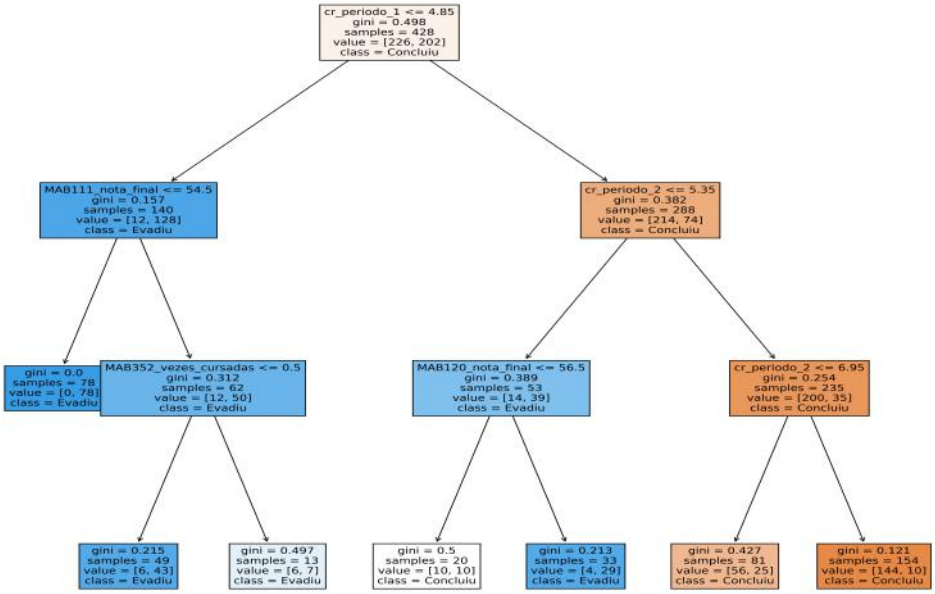
O conjunto C é utilizado para prever a evasão a partir do sétimo período. Analisando a árvore gerada treinando o modelo 2 nesse conjunto (Figura 23) vê-se que: (i) a maioria dos evadidos no conjunto de treino (52 alunos de 73 - 71,2%) tiveram nota final abaixo de 3,8 em Linguagens Formais(disciplina do terceiro período). Destes 52 alunos, 36 (69,2%) possuem CR do primeiro período inferior a 4,8, além de quase todos (35) terem nota na disciplina Arquitetura de Computadores (disciplina de quinto período) inferior a 4,6. Quase todos os alunos que possuem CR do primeiro período superior a 4,8 e que evadiram (14 dos 16 alunos), haviam cursado a disciplina de Circuitos Lógicos até o terceiro período. Como esta disciplina é do segundo período, isso indica que o aluno pode ter sido reprovado em Fundamentos da Computação Digital, atrasando em 1 período a conclusão de Circuitos

Lógicos. Isso corrobora o que foi apontado pelo modelo usado no conjunto A, que apontou a nota na disciplina de Fundamentos da Computação Digital como um fator importante para a evasão. Dos alunos que cancelaram e tiveram uma nota em Linguagens Formais superior a 3,8 (21 alunos de 73, 28,8%), 10 alunos tiveram nota final em Lógica (quinto período) inferior a 3,8 e nota final de Estrutura de Dados inferior a 5,4. Importante notar que as disciplinas que aparecem como fator importante para evasão possuem uma relação de entre si: Fundamentos é pré-requisito de Circuitos Lógicos, que é requisito de Arquiterura de Computadores, e Linguagens Formais é pré-requisito de Lógica, o que sugere que os alunos com maior possibilidade de evasão possuem dificuldades em partes específicas do curso.

O conjunto D é utilizado para prever a evasão a partir do primeiro período. Analisando a árvore gerada treinando o modelo 2 nesse conjunto (Figura 24) vê-se que: (i) a maioria dos evadidos no conjunto de treino (200 alunos em 322 - 62,1%) tiveram CR no primeiro período menor 4,9, sendo que 161 deles tiveram nota final em Fundamentos da Computação Digital inferior a 5,3. Do restantes (122 alunos de 322, 37,9%), pouco menos da metade (54) obteve nota final na disciplina Números Inteiros e Criptografia inferior a 5,5. A utilização dessa árvore identificou corretamente 83,1% dos evadidos no conjunto de teste, e isso é relevante porque a maioria dos alunos que cancela a matrícula completou o primeiro período mas não o segundo, conforme a Figura 5 indica. Observa-se que (i) é bem parecido com o conjunto A, utilizando o CR e a disciplina de Fundamentos como indicadores e isso faz sentido já que o conjunto A busca identificar alunos propensos a evasão a partir do segundo período e o conjunto D a partir do primeiro, ou seja, diferença de apenas um período.

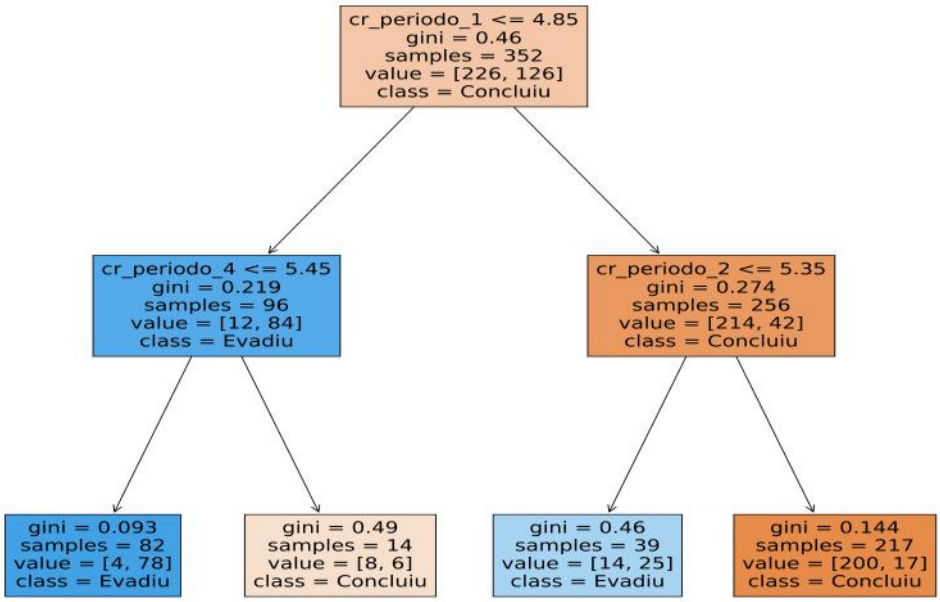
A análise dos conjuntos de dados A, B, C e D permitiu identificar fatores que indicam a propensão de alunos à evasão em diferentes momentos da graduação em Ciência da Computação. É possível observar que os indicadores mais relevantes incluem o desempenho em disciplinas específicas, como Fundamentos da Computação Digital, Circuitos Lógicos e Linguagens Formais, além do CR dos primeiros períodos. A relação entre essas disciplinas sugere que os alunos que possuem dificuldades em partes específicas do curso têm uma maior propensão à evasão, os resultados apresentados possibilitam a criação de estratégias para auxiliá-los e, assim, aumentar as chances de conclusão do curso.

Figura 21 – Árvore de decisão do modelo 2 com conjunto A



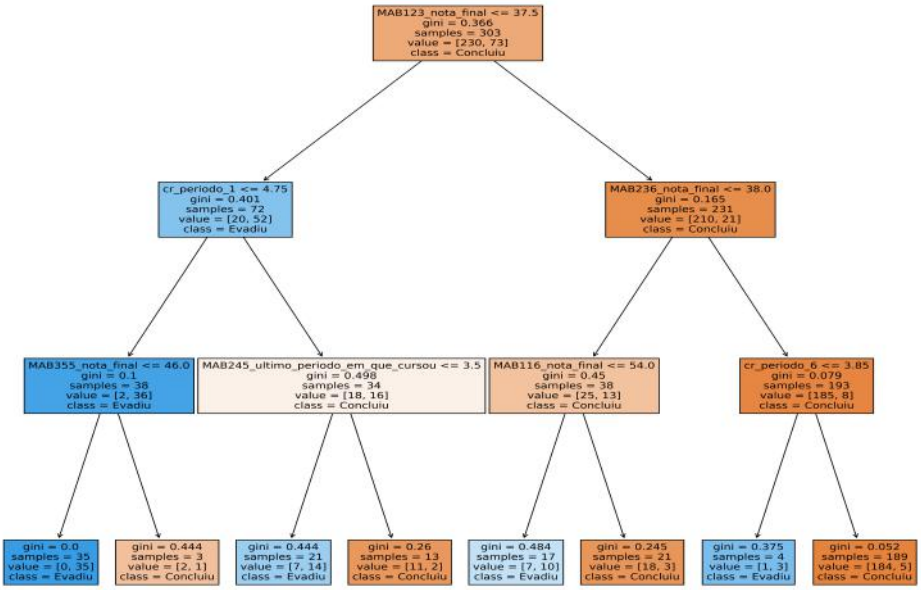
Fonte: Própria

Figura 22 – Árvore de decisão do modelo 2 com conjunto B



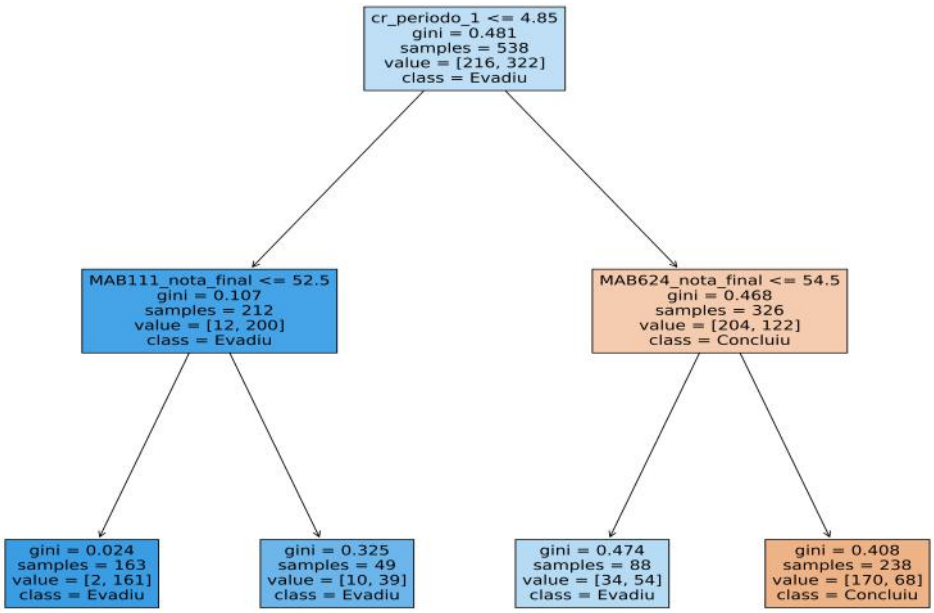
Fonte: Própria

Figura 23 – Árvore de decisão do modelo 2 com conjunto C



Fonte: Própria

Figura 24 – Árvore de decisão do modelo 2 com conjunto D



Fonte: Própria

6 CONCLUSÃO

Neste trabalho, utilizamos o algoritmo de árvore de decisão para identificar alunos propensos a evasão. Foi possível criar um modelo de classificação que apresentou índices de F1 score, acurácia e recall entre 65% e 85%, índices que são considerados bons para esse tipo de análise.

A partir das árvores de decisão geradas foi possível criar um perfil dos alunos que evadem. Observou-se que os indicadores mais importantes para evasão consistem no desempenho acadêmico em disciplinas específicas como Fundamentos da Computação Digital, Circuitos Lógicos e Linguagens Formais, além do Coeficiente de Rendimento (CR) dos primeiros períodos. Compreender as características comuns desses alunos pode ajudar a desenvolver estratégias para evitar a evasão, como a oferta de orientação acadêmica e apoio emocional. O perfil pode ser utilizado por professores e orientadores para identificar e prevenir a evasão.

A utilização de dados acadêmicos nesse modelo de classificação é um fator chave para o sucesso da abordagem. Isso ocorre porque esses dados são altamente relevantes para a compreensão do desempenho dos alunos e podem indicar alguns dos problemas que levam à evasão, como notas baixas ou atrasos nas disciplinas. Além disso, a utilização de técnicas de aprendizado de máquina permite que o modelo analise grandes quantidades de dados de forma rápida e eficiente, o que é fundamental em um ambiente com muitos alunos e informações a serem processadas. A expectativa é que os resultados desse estudo possam contribuir para a implementação de estratégias que permitam reduzir a evasão no curso analisado.

Os processos de seleção, tratamento e modelagem de dados podem ser aplicados a outros cursos da instituição e serem replicados em outras universidades, o que pode contribuir para a redução da evasão no ensino superior como um todo.

Outro aspecto relevante desse trabalho foi confirmar que técnicas de aprendizado de máquina podem ter um papel relevante em projetos que visam reduzir a taxa de evasão nas universidades, abrindo novas possibilidades para o desenvolvimento de novas abordagens.

Com base nos resultados obtidos neste trabalho, os próximos passos incluem a comparação do modelo de árvore de decisão com outras técnicas de aprendizado de máquina, como redes neurais e algoritmos de clustering. Essa comparação pode permitir a identificação da técnica mais adequada para a análise de dados acadêmicos e a criação de modelos mais precisos e eficientes para a previsão da evasão.

REFERÊNCIAS

BROWNLEE, J. **Training-validation-test split and cross-validation done right**. 2021. Disponível em: https://machinelearningmastery.com/training-validation-test-split-and-cross-validation-done-right/. Acesso em: 10/01/2023.

BURKOV, A. **The hundred-page machine learning book**. [S.l.]: Andriy Burkov Quebec City, QC, Canada, 2019. v. 1.

DUTRA, R. **O Uso de Inteligência Artificial para Predição de Evasão na Rede Doctum de Ensino**. Monografia (TCC) — FACULDADES INTEGRADAS DE CARATINGA – FIC, 2015.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, p. 641–659, 2007.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques, third edition**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. Disponível em: http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1.

JAMES, G. et al. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013. Disponível em: https://faculty.marshall.usc.edu/gareth-james/ISL/.

JUNIOR, I. P. d. B. **Uso de mineração de dados educacionais para a classificação e identificação de perfis de evasão de graduandos em Sistemas de Informação**. Dissertação (B.S. thesis) — Universidade Federal do Rio Grande do Norte, 2018.

KELLEHER, J. D.; MACNAMEE, B.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies**. Cambridge, MA: MIT Press, 2015. ISBN 978-0-262-02944-5.

LIMA, P. et al. Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior. **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 27, p. 157–178, 2019.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos**, v. 25, p. 14, 2012.

LOBO, R. A evasão no ensino superior brasileiro–novos dados. 2017. Disponível em: https://www.institutolobo.org.br/core/uploads/artigos/art_088.pdf. Acesso em: 25/04/2023.

PEREIRA, P. **O fantasma da evasão**. 2014. Disponível em: https://revistaeducacao.com.br/2014/02/18/o-fantasma-da-evasao/. Acesso em: 08/03/2023.

PINHEIRO, R. **Evasões na Universidade de Brasília causam prejuízo de R$ 95 mi**. 2015. Disponível em: https://www.correiobraziliense.com.br/app/noticia/cidades/2015/10/10/interna_cidadesdf,501999/evasoes-na-universidade-de-brasilia-causam-prejuizo-de-r-95-mi.shtml. Acesso em: 08/03/2023.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. [S.l.]: Prentice Hall, 2010.

SALLES, G.; SILVA, J. **Evasão, retenção e orientação acadêmica: reflexões para o ensino de graduação**. 2021. Disponível em: https://www.youtube.com/watch?v=kV0Y9X_Eg0A. Acesso em: 10/03/2023.

SANTOS, J. C. B. dos. **Usando mineração de dados para predição da evasão escolar**. Monografia (TCC) — Instituto Federal de Santa Catarina, 2021.

TABOGA, M. **Training, validation and test samples**. 2021. Disponível em: https://www.statlect.com/machine-learning/training-validation-and-test-samples. Acesso em: 10/03/2023.

ZHENG, A.; CASARI, A. **Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists**. O'Reilly, 2018. ISBN 9781491953242. Disponível em: https://books.google.com.br/books?id=Ho0UvgAACAAJ.

GLOSSÁRIO

abnTeX2 suíte para LaTeX que atende os requisitos das normas da ABNT para elaboração de documentos técnicos e científicos brasileiros. *veja* LaTeX

componente descrição da entrada componente.

equilíbrio da configuração consistência entre os componentes. *veja também* componente

APÊNDICES

APÊNDICE A – GRADE CURRÍCULAR DO BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO NA UFRJ.

Segue abaixo uma lista completa das matérias obrigatórias para a conclusão do curso de Bacharelado em Ciências da Computação na UFRJ, separadas pelo período em que devem ser realizadas. Nessa lista não é considerado o trabalho de conclusão de curso ou matérias optativas.

No primeiro período as matérias são: Fundamentos da Computação Digital, Sistemas de Informação, Computação I, Números Inteiros Criptografia e Cálculo Infinitesimal I.

As matérias do segundo período são: Organização da Informação, Computação II, Circuitos Lógicos, Matemática Combinatória e Cálculo Integral e Diferencial II.

No terceiro período: Mecânica, Oscilação e Ondas, Álgebra Linear Algorítmica, Estrutura dos Dados, Computadores e Programação, Linguagens Formais e Cálculo Integral e Diferencial III.

No quarto período: Eletromagnetismo e Ótica, Computação Concorrente, Cálculo Numérico, Algoritmos e Grafos e Cálculo Integral e Diferencial IV.

No quinto período: Lógica, Computadores e Sociedade, , Arquitetura de Computadores I, Compiladores I, Banco de Dados I e Fundamentos da Engenharia de Software.

No sexto período: Computação Gráfica I, Programação Linear I, Inteligência Artificial e Estatística e Probabilidade.

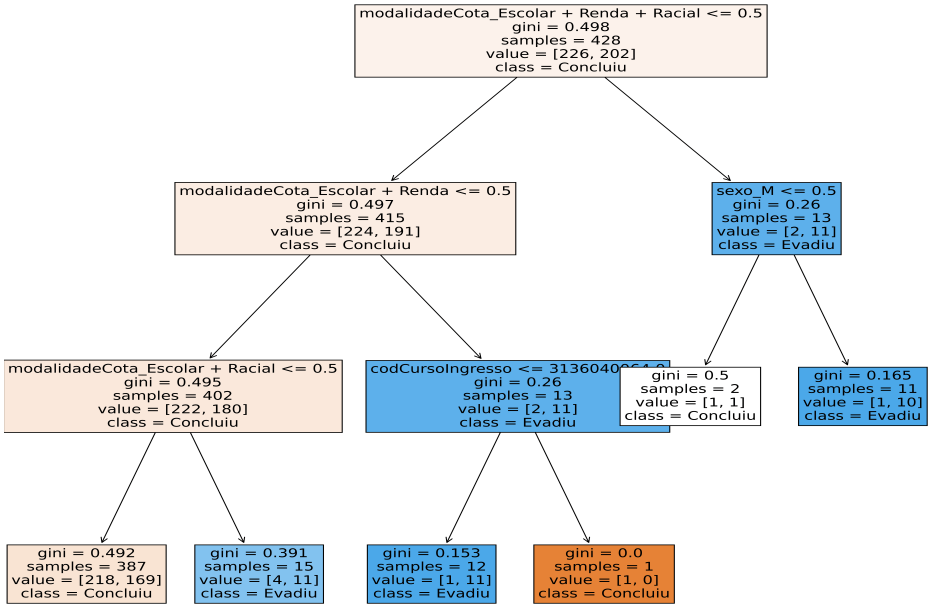
No sétimo período: Sistemas Operacionais I e Avaliação e Desempenho.

No oitavo período: Teleprocessamento e Redes.

APÊNDICE B – ÁRVORES DE DECISÃO GERADAS NOS EXPERIMENTOS UTILIZANDO DADOS DE TREINO.

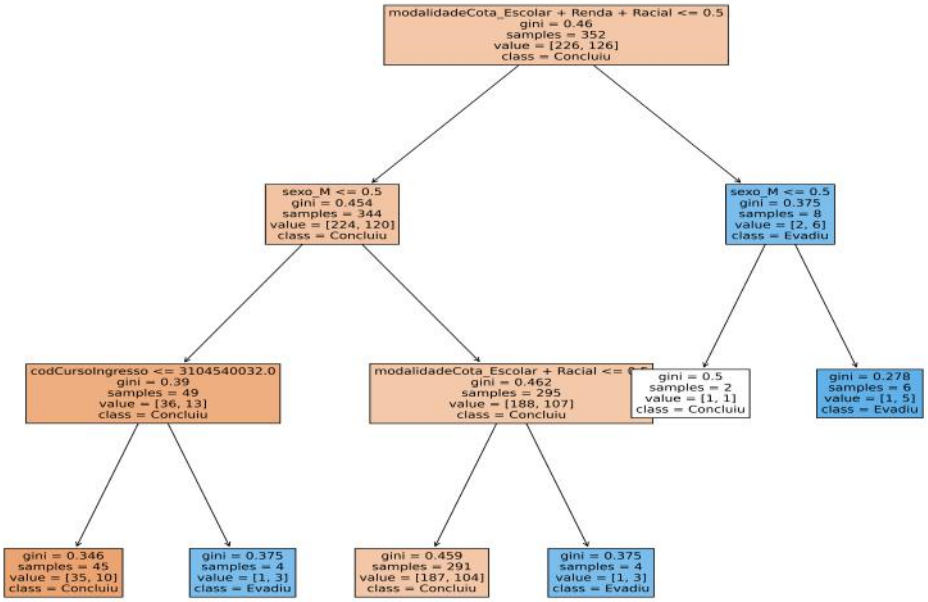
Neste apêndice são disponibilizadas as árvores de decisão geradas nos experimentos utilizando dados de treino.

Figura 25 – Árvore de decisão do modelo 1 com conjunto A



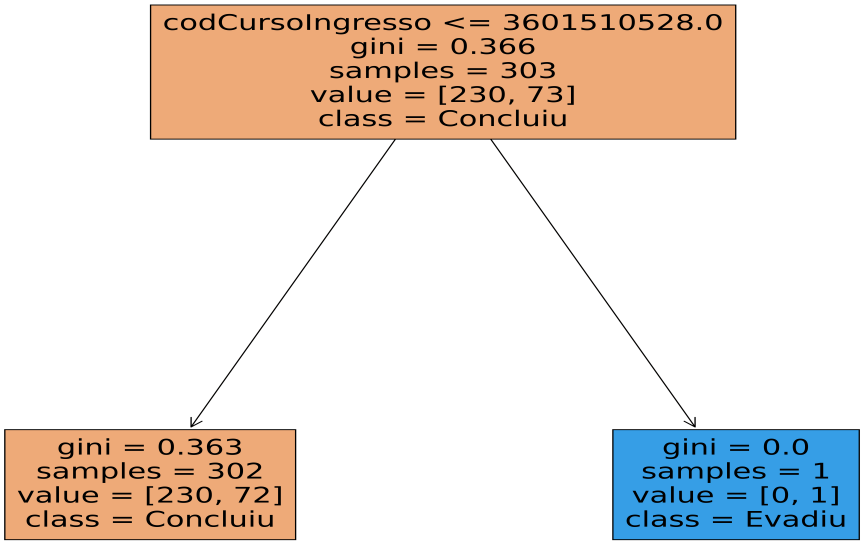
Fonte: Própria

Figura 26 – Árvore de decisão do modelo 1 com conjunto B



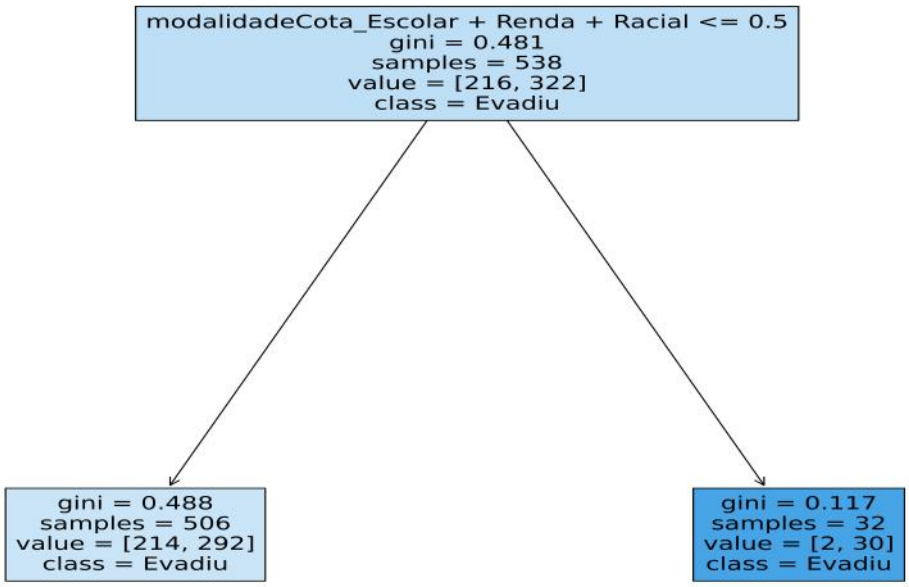
Fonte: Própria

Figura 27 – Árvore de decisão do modelo 1 com conjunto C



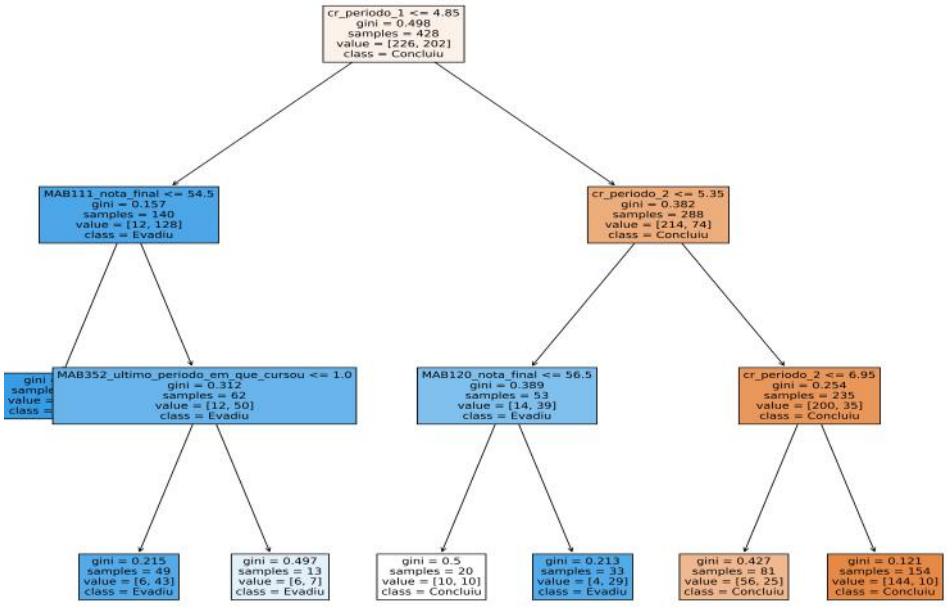
Fonte: Própria

Figura 28 – Árvore de decisão do modelo 1 com conjunto D



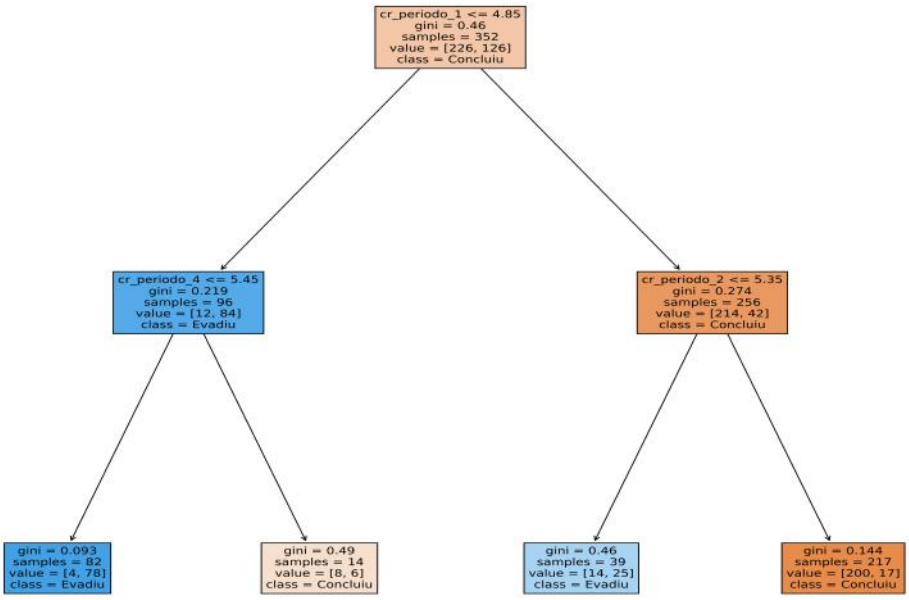
Fonte: Própria

Figura 29 – Árvore de decisão do modelo 3 com conjunto A



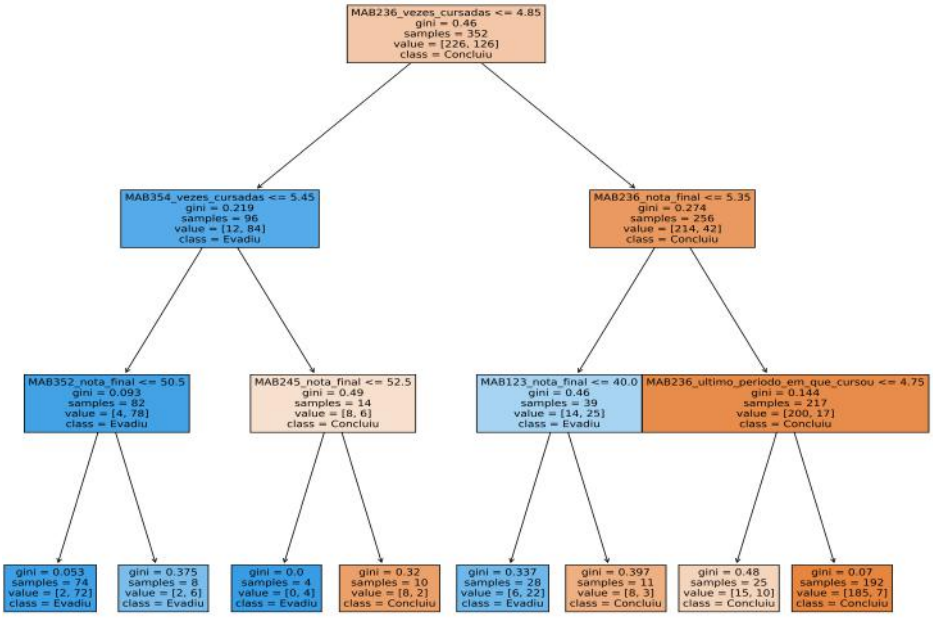
Fonte: Própria

Figura 30 – Árvore de decisão do modelo 3 com conjunto B



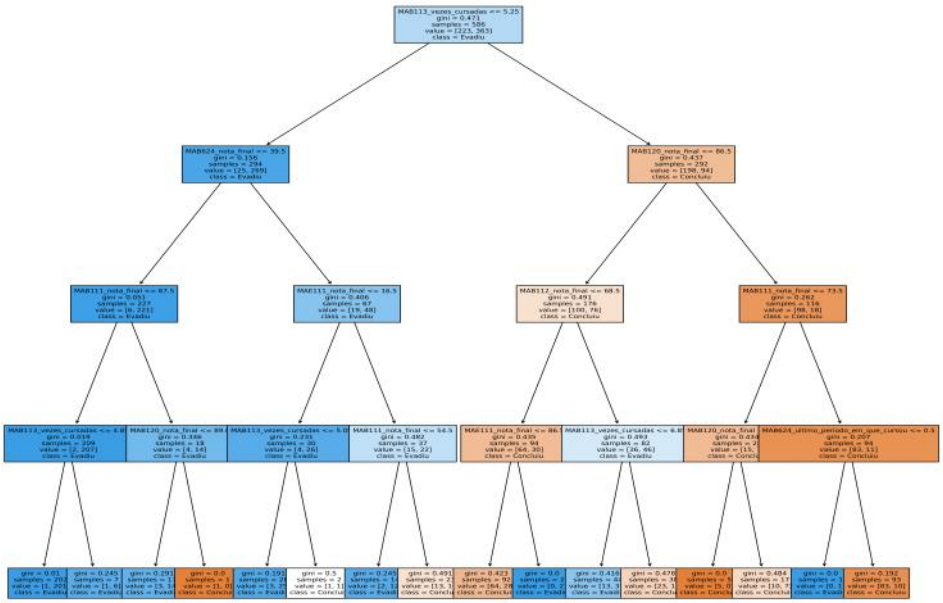
Fonte: Própria

Figura 31 – Árvore de decisão do modelo 3 com conjunto C



Fonte: Própria

Figura 32 – Árvore de decisão do modelo 3 com conjunto D



Fonte: Própria