Trabalho de Conclusão de Curso

Análise de Evasão Baseada em Modelos Preditivos Para o Curso de Engenharia de Computação da Universidade Federal de Alagoas

Igor da Cunha Araújo Theotônio
icat@ic.ufal.br

Orientador:

Prof. Dr. Thiago Damasceno Cordeiro

Maceió, maio de 2023

Igor da Cunha Araújo Theotônio

Análise de Evasão Baseada em Modelos Preditivos Para o Curso de Engenharia de Computação da Universidade Federal de Alagoas

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. Thiago Damasceno Cordeiro

Maceió, maio de 2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Helena Cristina Pimentel do Vale – CRB4 –661

T397a Theotônio, Igor da Cunha Araújo.
Análise de evasão baseada em modelos preditivos para o curso de Engenharia de Computação da Universidade Federal de Alagoas / Igor da Cunha Araújo Theotônio. – 2023.
29 f : il.

Orientador: Thiago Damasceno Cordeiro.
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2023.

Bibliografia: f. 28-29.

1. Engenharia de Computação. 2. Evasão universitária. 3. Modelo preditivo. 4. Árvore de decisão. 5. Floresta aleatória. (Random forest). I. Título.

CDU: 004.41:378

Agradecimentos

Primeiramente gostaria de agradecer aos meus pais, Rodrigo da Cunha Rocha Theotônio e Patrícia Almeida de Araújo Lima, irmã, Letícia Almeida de Araújo Theotônio e esposa, Mariana Calheiros Barroca Theotônio. Em especial minha mãe, que por muitas vezes foi a maior incentivadora para que eu finalizasse o curso independente dos rumos profissionais que tomei e tenho certeza que independente de onde esteja vai ser a pessoa mais feliz e realizada por esta entrega. Agradecê-los por todo o apoio, incentivo e parceria durante toda essa jornada.

Também, aos amigos que a faculdade me deu: Alfredo Lima, Lucas Peixoto e Vinícius Costa, que estiveram comigo do primeiro ao último dia de faculdade. Foram pessoas essenciais na minha formação acadêmica e desenvolvimento pessoal e com certeza sem eles o caminho teria sido muito mais difícil e menos engrandecedor.

Por fim, agradecer a todos os professores, alunos e profissionais que fazem o IC e o curso de Engenharia de Computação ser o que é, permitindo que eu e tantos outros alunos tenham uma formação rica e completa. Aqui, gostaria de nomear os professores que estiveram mais próximos durante minha jornada e com certeza sem eles minha formação não teria sido possível da forma que foi: Maria Andrade, Márcio Ribeiro, Aydano Machado, Erick Barboza, Ícaro Araújo, João Raphael, Heitor Savino e meu orientador Thiago Cordeiro.

Igor da Cunha Araújo Theotônio

"A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo"

– Albert Einstein

Resumo

Contexto: Evasão escolar é um problema crítico e crônico presente em diferentes níveis do ensino na educação brasileira. Este problema é extremamente presente nas faculdades públicas, em especial nos cursos de ciências exatas, afetando diretamente o curso de Engenharia de Computação do Instituto de Computação (IC) da Universidade Federal de Alagoas (Ufal).
Objetivo: Atualmente não há nenhum recurso sendo utilizado para auxiliar a coordenação do referido curso a mapear alunos com alto risco de evasão. Portanto, o objetivo deste trabalho consistem em avaliar a factibilidade do uso de ferramentas baseadas em modelos preditivos como meio de suporte para diminuição dos índices de evasão escolar.
Método: Esse trabalho apresenta a comparação entre dois modelos preditivos, sendo o primeiro baseado em árvores de decisão simples e o segundo baseado no algoritmo *Random Forest* utilizando a base de dados dos alunos do curso de Engenharia de Computação do IC da Ufal. Estes dados foram fornecidos pelo Núcleo de Tecnologia da Informação (NTI) da própria instituição e contém informações desde o início do curso até o primeiro semestre do ano 2020.
Resultados: Os resultados apontam que os modelos gerados possuem acurácia e *recall* acima de 85% em ambos os modelos, resultados considerados positivos dado o problema de pesquisa apresentado. Além disso, outras métricas e informações relevantes puderam ser extraídas dos dados utilizados.
Conclusão: Diante do trabalho desenvolvido e dos resultados obtidos, acreditamos que a utilização de modelos preditivos baseados em árvore de decisão como ferramenta de suporte à redução de evasão do no curso de Engenharia de Computação do IC/Ufal é factível.

Palavras-chave: Engenharia, Computação, Evasão, Modelo Preditivo, Árvore de Decisão, *Random Forest*

Abstract

Context: School dropout is a critical and chronic problem present at different levels of teaching in Brazilian education. This problem is extremely present in public colleges, especially in exact sciences courses, directly affecting the Computer Engineering course at the Institute of Computing (IC) of the Federal University of Alagoas (Ufal).
Objective: Cthere is no resource being used to help the coordination of the referred course to map students with high dropout risk. Therefore, the objective of this work is to evaluate the feasibility of using tools based on predictive models as a means of support for reducing school dropout rates.
Method: This work presents the comparison between two predictive models, the first based on simple decision trees and the second based on the *Random Forest* algorithm using the database of students of the Computer Engineering course at IC/Ufal. These data were provided by the Information Technology Center (NTI) of the institution and contain information from the beginning of the course until the first half of 2020.
Results: The results indicate that the generated models have accuracy and *recall* above 85% in both models, results considered positive given the research problem presented. In addition, other metrics and relevant information could be extracted from the data used.
Conclusion: In view of the work carried out and the results obtained, we believe that the use of predictive models based on decision trees as a tool to support the dropout reduction in the Computer Engineering course at IC/Ufal is feasible.

Keywords: Engineering, Computing, Evasion, Predictive Model, Decision Tree, Random Forest

Conteúdo

Lista de Figuras

Lista de Tabelas

1

Introdução

1.1 Motivação

No mundo de hoje, uma das discussões mais inócuas diz respeito à importância da educação como mola mestra do desenvolvimento econômico e social de uma nação. Há consenso em todas as partes do mundo de que a educação é a principal alavanca de melhoria nas condições de vida de qualquer povo.

Segundo Ioschpe [9], é possível calcular uma taxa de retorno à educação, onde observa-se uma relação consistentemente linear entre educação e salário. O coeficiente desta relação indica que um ano a mais de educação para cada indivíduo gera, em média, um aumento salarial de 10% para o mesmo. Também, do ponto de vista macroeconômico, pode-se observar que um ano a mais de educação gera um aumento de renda per capita entre 8% e 10%, além de um aumento do PIB e melhoria de indicadores de saúde.

Hoje no Brasil, segundo o censo da educação superior de 2021 divulgado pelo Ministério da Educação (MEC), há quase 9 milhões de pessoas matriculadas no ensino superior, público ou privado. Apesar disso, existem apenas cerca de 1,3 milhão de concluintes, o que representa menos de um terço dos novos ingressos que são quase 4 milhões de pessoas [8].

Partindo do pressuposto, devemos considerar que o desenvolvimento econômico de uma nação possui razões quantitativas e qualitativas no que diz respeito ao nível de instrução formal de uma população. Isto é, não adianta apenas relacionar o número de matrículas em instituições de ensino ao número de habitantes de uma região, pois essas matrículas podem, por meio de uma análise simplificada, não significar a permanência consistente e duradoura nos bancos escolares e, muito menos, atestar a qualidade da educação oferecida.

No Brasil, o tema evasão começou a ganhar força a partir da aprovação da Lei de Diretrizes e Bases da Educação (LDB), que motivou a criação da Comissão Especial de Estudos sobre evasão, que realizou estudos que apontavam altas taxas de evasão (50%) nas universidades públicas, além de um baixo número de concluintes [2].

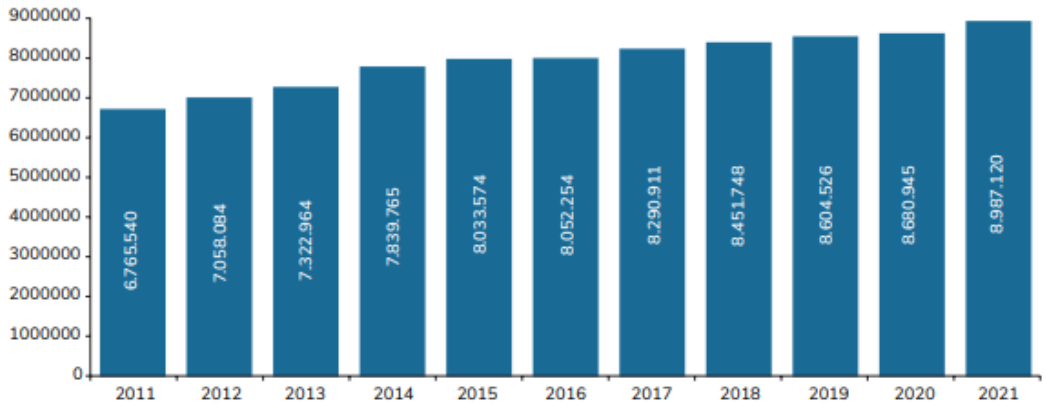

Figura 1.1: Número de matrículas em cursos de graduação [8].

Ainda hoje, a evasão é um tema complexo e que mobiliza diversos esforços nas diferentes esferas da sociedade. Além disso, é uma questão que preocupa e impacta diretamente todas as instituições de ensino, sejam elas públicas ou privadas, pois a saída de alunos causa graves prejuízos sociais, econômicos e acadêmicos [14, 6].

Inclusive fora do Brasil é possível observar o quão claro e gritante é o problema da evasão. "Na América Latina, as taxas de evasão do ensino superior somam 57%. Os níveis mais altos podem chegar a 82% em países como a Guatemala, e os mais baixos a 40% em países como Argentina. Em países desenvolvidos, como Espanha e Estados Unidos, a porcentagem de evasão não é tão diferente: cerca de 30% a 50%. Em países europeus, como Alemanha e Finlândia, essa mesma taxa varia de 10% a 25%"[15]. Isto é, até em países desenvolvidos e com ótimos indicadores socioeconomicos e educacionais, o problema é grave.

Levando isso em consideração, entendemos que o nível de evasão e taxa de concluintes no curso de Engenharia de Computação (EC) do Instituto de Computação (IC) da Universidade Federal de Alagoas (Ufal) merecem atenção, estudo e desenvolvimento de estratégias e ferramentas para diminuir a evasão e aumentar a taxa de concluintes.

Considerando que para aumento da taxa de concluintes é imprescindível que os alunos não evadam, faz sentido inicialmente focar no problema da evasão. Sendo assim, entendemos que um caminho possível para diminuir esta taxa é conseguir prever quem são os alunos mais propensos a evadir para que os órgãos competentes (coordenação do curso, por exemplo) possam atuar de forma mais próxima e reverter o cenário.

É pensando assim que surge o seguinte problema de pesquisa: a utilização de modelos preditivos baseados em árvores de decisão podem ser usados como uma ferramenta para identificar alunos propensos a evadir do curso de engenharia de computação do IC/Ufal?

1.2 Objetivos

A seguir apresentam-se os objetivos geral e específicos do trabalho.

1.2.1 Objetivo geral

Para responder ao problema de pesquisa apresentado, este trabalho tem como objetivo geral desenvolver modelos preditivos baseados em árvores de decisão como ferramentas de auxílio ao processo de identificação de alunos propensos a evadir no curso de Engenharia de Computação do Instituto de Computação da Universidade Federal de Alagoas.

1.2.2 Objetivos específicos

Para alcançar o objetivo geral deste trabalho, definiram-se os seguintes objetivos específicos:

a) Analisar as informações fornecidas pelo Núcleo de Tecnologia da Informação (NTI) da Ufal à respeito da situação acadêmica dos discentes do curso de EC do IC/Ufal;

b) Identificar correlações relevantes para o entendimento do perfil do aluno evasor;

c) Construir modelos de predição de evasão baseado em árvores de decisão que sejam capazes de auxiliar a coordenação do referido curso na tomada de decisões a fim de inibir a evasão escolar;

d) Testar os modelos construídos.

1.3 Justificativa

Este trabalho justifica-se na medida em que discute um assunto de grande relevância para o desenvolvimento econômico e social de qualquer país: a Educação. Mais especificamente, a evasão escolar dos alunos do curso de EC do IC/Ufal, curso este que possui baixas taxas de concluintes frente a outros cursos da mesma área do conhecimento, na mesma instituição de ensino.

Metodologicamente este trabalho pode ser classificado da seguinte forma: do ponto de vista da sua natureza, pesquisa aplicada; sob a ótica da sua problemática, o trabalho classifica-se como pesquisa quantitativa e qualitativa; sob o ponto de vista dos seus objetivos, o trabalho classifica-se como pesquisa exploratória, descritiva e explicativa; com relação aos instrumentos de coleta de dados: pesquisa documental, bibliográfica e estudo de caso.

Além disso, o trabalho será desenvolvido em cinco macro etapas:

- Estudo e análise da base de dados fornecida pelo NTI para entender quais dados tem real valor para o estudo e como serão utilizados;
- Desenvolvimento de *scripts* para extração e processamento dos dados de interesse;

- Desenvolvimento e avaliação de modelos preditivos baseados em árvores de decisão;
- Análises estatísticas dos resultados obtidos;
- Eventuais aprimoramentos nos modelos desenvolvidos anteriormente com base nas análises supracitadas.

Vale ressaltar que não há, neste trabalho, nenhuma intenção de identificar individualmente nenhum aluno e sim de desenvolver modelos preditivos que sejam capazes de representar o comportamento de alunos evasores de forma totalmente geral e não individual, preservando todo e qualquer dado pessoal e/ou sensível, nos termos da Lei Geral de Proteção de Dados (LGPD).

Também, levando-se em consideração as peculiaridades dos mais diversos cursos, a metodologia desenvolvida para este trabalho poderá ser replicada para qualquer curso da Ufal, respeitando-se as particularidades de cada um.

Estruturalmente o trabalho se desenvolve da seguinte forma:

- Capítulo 2: Fundamentação Teórica, onde se discutirá sobre informações e conceitos relevantes para o desenvolvimento e compreensão do trabalho;
- Capítulo 3: Metodologia, onde se discorrerá acerca da forma sobre a qual o trabalho foi desenvolvido;
- Capítulo 4: Resultados e Discussões, capítulo no qual serão apresentados os resultados obtidos e as discussões geradas com base neles;
- Capítulo 5: Conclusão, espaço final de sumarização do trabalho realizado e sugestão de trabalhos futuros.

2

Fundamentação Teórica

Neste capítulo será discutido o problema da evasão no ensino superior com ênfase no curso de Engenharia de Computação do IC/Ufal. Além disso, serão fornecidos os conceitos básicos sobre dois modelos de classificação (Árvore de Decisão e *Random Forest*) e outros aspectos técnicos, que serão utilizados para avaliar a evasão no curso citado.

2.1 Evasão no Ensino Superior

Neste trabalho, consideraremos evasão como "a saída definitiva do aluno de seu curso de origem, sem concluí-Io"[2]. Ou seja, consideraremos como um aluno evasor aquele que por quaisquer motivos, abandonou de forma definitiva o curso de Engenharia de Computação do IC/Ufal.

Em um contexto geral, é um problema global que afeta tanto países desenvolvidos quanto sub desenvolvidos, instituições tanto públicas quanto privadas e, direta e indiretamente, a sociedade como um todo.

Quando visto sob a ótica das instituições públicas brasileiras, temos um alto gasto sendo gerado nas universidades de um orçamento advindo de impostos e que ao evadir, a tendência é que o aluno não gere nenhum retorno desse investimento para a sociedade. No Brasil, apenas 19,7% dos jovens entre 18 e 24 anos estão no ensino superior e apenas 24% já concluiu ou frequenta o ensino superior [7].

Considerando que um dos objetivos da meta 12 do Plano Nacional de Educação (PNE) é que 33% dos jovens entre 18 e 24 esteja no ensino superior [4], além de acelerar e proporcionar acesso à mais pessoas é imprescindível garantir que os altos indicadores de evasão diminuam, caso contrário é o mesmo que colocar uma torneira maior numa piscina com um buraco maior, por mais que entre mais água, vazará mais água e a piscina não será capaz de reter o volume de água esperado. No atual cenário, a estimativa é que essa meta seja apenas alcançada em 2040 [11].

No Brasil, em 2021, a taxa de evasão do ensino superior foi de 36,6% considerando as modalidades de ensino EAD e presencial [13], o que equivale a 3,42 milhões de alunos.

2.2 Modelos de Classificação

Algoritmos de aprendizagem de máquina são organizados em um tipo de taxonomia, de acordo com o resultado esperado, objetivo e método de construção. De acordo com [3] alguns dos tipos mais comuns são:

- Aprendizado supervisionado
- Aprendizado não supervisionado
- Aprendizado semi supervisionado
- Aprendizado por reforço

Neste trabalho o foco está em modelos de aprendizagem supervisionada, isto é, modelos que relacionam uma saída e uma entrada com base em dados rotulados. Nestes casos, a base utilizada para alimentar o modelo necessita ser composta de pares de entradas e saídas conhecidos e para cada saída um rótulo deve ser atribuído [16]. Aqui, os rótulos são evasor ou não evasor e após a determinação do modelo preditivo com base nos dados de treinamento, ele pode ser utilizado para predizer os rótulos de dados antes desconhecidos [1].

Algoritmos onde a saída pode assumir somente um conjunto de rótulos pré-definidos são chamados de Algoritmos de Classificação, e neste caso, como o problema apresenta apenas duas classes, se tratará de classificação binária [16].

É importante salientar que algoritmos treinados com base em aprendizado supervisionado tem seu resultado atrelado de forma totalmente dependente à qualidade dos dados usados para o treinamento. Ou seja, se o conjunto de dados utilizado não tiver qualidade e expressar de forma satisfatória o universo de dados que poderá ser usado como entrada para o algoritmo em momentos futuros a qualidade da predição do algoritmo será baixa [3].

Em [5] Santos aplicou técnicas de classificação binária com aprendizagem supervisionada baseadas em Árvores de Decisão, para tentar prever a evasão de cursos de engenharia e de computação da Universidade Federal Fluminense (UFF), com base no desempenho dos estudantes nas disciplinas cursadas. Neste trabalho, modelos específicos são gerados para prever se um determinado aluno, ao fim de um i-ésimo semestre, se ele ou ela irá evadir ou não. Experimentos mostraram que os modelos foram capazes de atingir acurácia entre 79,31% e 98,25%, utilizando Árvores de Decisão, e entre 81,18% e 97,06%, utilizando o algoritmo *Random Forest*.

2.2.1 Árvore de Decisão

Árvores de decisão são uma técnica popular de aprendizagem de máquina amplamente utilizada em diversas áreas de estudo e aplicação no mercado, é, inclusive, uma das técnicas mais bem sucedidas de aprendizagem de máquina [12].

Esta forma de modelagem consiste em construir um modelo de decisão em forma de árvore, em que cada nó interno representa uma decisão baseada em uma determinada variável, e cada folha representa o resultado final da decisão, na Figura 2.1 pode-se observar um exemplo usado para definir se uma família vai ou não esperar por uma mesa em um restaurante. Apesar de árvores de decisão serem utilizadas tanto para regressão quanto para classificação, neste trabalho o objetivo é apenas de classificação, e portanto, usaremos a técnica para este fim.

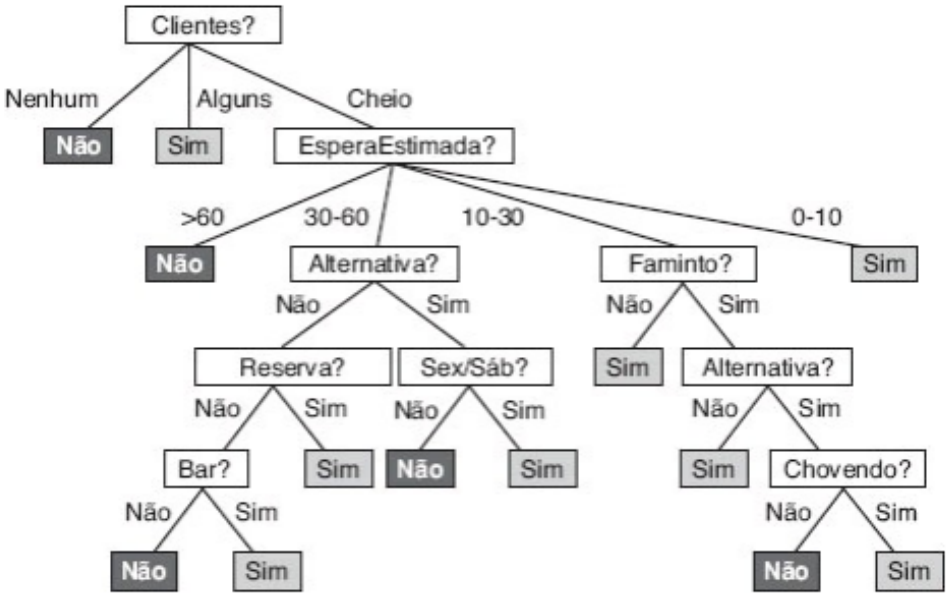

Figura 2.1: Exemplo de árvore de decisão [12].

2.2.2 Random Forest

O *Random Forest* ou floresta aleatória é um algoritmo que também possui função classificatória, introduzido por Breiman em 2001. Este algoritmo é uma combinação de diversas árvores de decisão construídas a partir de *subsets* do conjunto total de treinamento, objetivando a construção de árvores não relacionadas de forma aleatória.

Depois da construção do conjunto de árvores, cada árvore do conjunto terá uma resposta de saída para determinada entrada, classificando o exemplo dado. Ao final do algoritmo, através de um sistema de votos, a saída mais repetida entre as árvores que compõem a floresta, será a saída definida pelo algoritmo.

2.2.3 Importância dos Atributos

Na construção de uma árvore de decisão, é preciso escolher que atributo estará representado em determinado nó, isto é, com base em que atributos serão feitos os testes que definirão os caminhos a serem seguidos dentro da árvore até que um uma folha seja atingida e uma decisão seja indicada.

Para isso, utiliza-se a função de importância de um atributo, onde avalia-se o ganho de informação que determinado atributo representa dentro da árvore, definida em termos de entropia [12]. Este processo é capaz de indicar a relevância de cada variável no processo de tomada de decisão, onde um atributo perfeito divide os exemplos usados em conjuntos compostos apenas de variáveis positivas ou apenas de variáveis negativas, ou seja, folhas da árvore.

Em resumo, a importancia dos atributos é o que permite a construção de uma árvore mais inteligente e com menor profundidade, facilitando não só a tomada de decisão como a interpretação da árvore gerada pelo algoritmo utilizado.

2.3 Treinamento, teste e validação

"A primeira lição a ser aprendida é que, independentemente da medida de avaliação a ser usada para atestar a qualidade de um modelo, não é adequado avaliá-lo por seu desempenho em relação aos exemplares apresentados no processo de treinamento (indução). É sempre necessário saber como o modelo se comporta quando aplicado a exemplares que ainda não conhece, ou seja, não usados no processo de sintonização de seus parâmetros. O motivo para essa ressalva é que modelos preditivos, a depender de como são gerados, podem levar à manifestação de um fenômeno bastante conhecido, o sobreajuste (do inglês *overfitting*)."[1]

Isto é, é necessário encontrar formas de treinar o modelo corretamente sem que ele seja útil apenas em um conjunto específico dos dados e sim na maior porção possível dentro do universo de dados. Modelos que não são capazes de fornecer respostas adequadas a novos exemplares de dados apresentados, e apenas aos dados utilizados no treinamento, perdem desempenho e caem em erro no que diz respeito a generalização [1].

2.3.1 *Holdout*

Para solucionar o problema acima apresentado, neste trabalho a estratégia utilizada para treinamento, teste e validação foi a *Holdout*, que consiste em dividir, aleatoriamente, os dados entre treinamento e teste, tradicionalmente na proporção de 70% e 30%, ou seja, 70% dos dados possuídos serão utilizados para construção dos modelos e os 30% restantes serão utilizados para avaliação do desempenho dos mesmos.

2.4 Avaliação dos Modelos

Para entender e mensurar, não só a qualidade dos modelos, como também tornar possível a comparação entre eles, faz-se necessário a definição de métricas de avaliação, tornando-os assim, comparáveis. Neste trabalho observa-se um problema de classificação binário, isto é, buscamos classificar os alunos entre evasores e não evasores.

Neste tipo de problema, é extremamente importante identificar eventuais erros apresentados pelo modelo, bem como entender que tipo de erro é mais ou menos relevante para o problema estudado. Em [10] define-se: "Um erro do tipo I (α) corresponde ao erro de rejeitar a hipótese nula H_0 quando ela é, de fato, verdadeira (falso positivo). Um erro do tipo 2 (β) corresponde ao erro de não rejeitar H_0, quando ela é falsa (falso negativo).

Aqui, temos como H_0: "o aluno não é evasor"e, portanto, um falso positivo seria classificá-lo como evasor quando na verdade ele não é e um falso negativo seria classificá-lo como não evasor quando de fato ele é.

No caso específico deste trabalho, é mais danoso identificar um aluno evasor como não evasor (falso negativo) incorretamente do que o contrário, já que todo objetivo aqui é diminuir os índices de evasão e para isso, o modelo precisa ser capaz de identificar corretamente alunos em risco de evasão.

2.4.1 Matriz de Confusão

Através da matriz de confusão, é possível observar de forma mais clara os tipos de erros:

	Valor Predito: Evasão	**Valor Predito: Não Evasão**	**Total**
Valor Real: Evasão	Verdadeiro Positivo (TP) Decisão Correta	Falso Negativo (FN) Erro Tipo II	Positivo Real
Valor Real: Não Evasão	Falso Positivo (FP) Erro Tipo I	Verdadeiro Negativo (TN) Decisão Correta	Negativo Real
Total	Positivo Predito	Negativo Predito	total

Tabela 2.1: Matriz de Confusão

2.4.2 Métricas de Avaliação

Antes da apresentação dos resultados obtidos é necessário definir as métricas que serão utilizadas no decorrer do trabalho: acurácia, *precision* e *recall*.

2.4.2.1 Acurácia

Acurácia pode ser definida como a razão entre as predições feitas corretamente e o total de predições. Isto é, é uma métrica que representa o quanto o modelo faz as predições

de forma correta independente do tipo de predição.

$$\text{acurácia} = \frac{\text{predições corretas}}{\text{total de predições}}$$

2.4.2.2 *Precision*

Precision pode ser definida como a razão entre as predições classificadas corretamente como positivas e o total de predições classificadas como positivas.

Esta métrica é melhor aproveitada em casos onde os falsos positivos são mais danosos que os falsos negativos, visto que busca representar a precisão do modelo em classificar instâncias positivas. Um exemplo seria aplicações financeiras, onde é extremamente importante que o modelo não indique como positiva uma aplicação negativa, mas é pouco danoso indicar como negativa uma aplicação positiva.

$$\textit{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

2.4.2.3 *Recall*

Recall pode ser definida como a razão entre as predições classificadas corretamente como positivas e o total de instâncias positivas.

Esta métrica é melhor aproveitada em casos onde os falsos negativos são mais danosos que os falsos positivos, visto que busca representar o quão capaz é o modelo de identificar corretamente um caso positivo dentre todos aqueles que são efetivamente positivos. Neste caso, aplicações de saúde em geral buscam otimizar essa métrica, pois é extremamente perigoso indicar que um paciente não está doente quando de fato ele está, enquanto indicar que ele está quando não está apenas o levaria para uma nova bateria de exames.

$$\textit{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Para este trabalho, consideraremos esta como a mais importante para avaliação do modelo, considerando que é extremamente importante não classificar incorretamente um aluno propenso a evadir. Por outro lado, classificar incorretamente um aluno que não é propenso a evadir apenas geraria uma conversa adicional por parte da coordenação, por exemplo.

3

Metodologia

Esse capítulo tem como objetivo detalhar o desenvolvimento do trabalho e o caminho traçado na análise executada. A comparação foi feita entre dois algoritmos baseados em árvores: árvore de decisão e *Random Forest* com o objetivo de avaliar a factibilidade do uso de modelos preditivos como ferramentas de suporte ao controle e redução de taxas de evasão.

3.1 Tratamento inicial da base de Dados

A base de dados enviada pelo NTI, com informação sobre os alunos do curso de EC do IC/Ufal, foi disponibilizada em formato XML. A primeira etapa foi transformá-la numa base de dados estruturada, para facilitar a análise e a manipulação dos dados. Neste processo, construímos uma base com informações que entendemos serem relevantes e são citadas a seguir:

- número de matrícula do aluno, posteriormente desconsiderado para a análise, tendo sido utilizado apenas como chave primária da base;
- coeficiente de rendimento acumulado do aluno;
- carga horária obrigatória fixa (disciplinas obrigatórias do curso);
- semestres válidos para integralização;
- semestres trancados;
- evasão [sim/não].

Além dessas informações, também foram criadas 4 colunas por disciplina, sendo estas listadas a seguir:

- número de reprovações por falta;
- número de reprovações por média;
- número total de reprovações;
- status atual [aprovado/não aprovado].

Também para construção da base, desconsideramos todas as informações posteriores ao semestre 2020.1 por duas razões: não considerar o período letivo excepcional (PLE) e nenhum período ainda não finalizado, removendo também da análise alunos que ingressaram após o período supracitado. Cabe frisar que foram consideradas apenas as disciplinas obrigatórias do curso, não considerando as eletivas. Após as etapas mencionadas acima, a base de dados inicial continha 217 variáveis, com 567 alunos, sendo 335 evasores e 232 não evasores.

3.2 Divisão da base entre treinamento e teste

Em um esforço inicial, a divisão da base entre treinamento e teste foi feita de forma aleatória, onde 70% das linhas de dados foram consideradas treinamento e 30% foram consideradas para teste, conforme explicado em 2.3.

3.3 Escolha dos algoritmos

Após o trabalho realizado na base de dados, decidimos utilizar o algoritmo de árvore de decisão de classificação, por ser um modelo que, em geral, possui um bom poder preditivo e além disso seu resultado pode ser facilmente interpretado pelos usuários através de sua representação gráfica que nada mais que é um fluxograma. Somado a isso, escolhemos um outro algoritmo de classificação baseado em árvore, *Random Forest*, que como explicado em 2.2.2 nada mais é que um algoritmo que combina diversas árvores para tomar uma decisão, com o objetivo de comparar os resultados obtidos.

3.4 Métricas de comparação e avaliação dos modelos

Para comparar os modelos apresentados acima, vamos utilizar as seguintes métricas, apresentadas em 2.4.2:

- acurácia;
- matriz de confusão;
  - verdadeiro positivo (TP);

  - verdadeiro negativo (TN);
  - falso positivo (FP);
  - falso negativo (FN);

- *precision*
- *recall*

Além disso, vamos identificar as variáveis de maior importância em cada um dos modelos e fazer uma análise empírica da árvore gerada pelo algoritmo Árvore de Decisão, uma vez que não é possível analisar as N árvores uitilizadas no algoritmo *Random Forest*.

Matrícula	Coeficiente	Carga Horaria Obrigatória Obtida	Semestres Válidos para Integralização	Semestres Trancados	Evasão	Cálculo 1 rep_falta	Cálculo 1 rep_media	Cálculo 1 rep_total	Cálculo 1 status	...	Mecânica dos Sólidos rep_falta	Mecânica dos Sólidos rep_media	Mecânica dos Sólidos rep_total	Mecânica dos Sólidos status
11xxxxxx	5.6	3060	17	1	False	0	4	4	1	...	0	4	4	0
14xxxxxx	0.0	0	10	0	True	1	0	1	0	...	0	0	0	0
16xxxxxx	4.76	1170	7	2	False	1	3	4	1	...	0	0	0	0
11xxxxxx	0.99	60	6	0	True	0	3	3	0	...	0	0	0	0
11xxxxxx	3.5	510	8	0	True	0	7	7	0	...	0	0	0	0
17xxxxxx	3.57	840	7	0	False	3	1	4	0	...	0	0	0	0
12xxxxxx	0.02	0	5	0	True	2	1	3	0	...	0	0	0	0
15xxxxxx	6.79	1110	7	3	True	0	0	0	1	...	0	0	0	0
14xxxxxx	1.82	120	11	0	True	0	2	2	0	...	0	0	0	0
14xxxxxx	6.61	2640	12	0	False	1	4	5	1	...	0	1	1	0
17xxxxxx	8.32	1980	6	0	False	0	1	1	1	...	0	0	0	0
11xxxxxx	7.73	3060	13	0	False	0	0	0	1	...	0	0	0	1
17xxxxxx	3.45	660	7	0	False	4	0	4	0	...	0	0	0	0
12xxxxxx	6.61	1890	7	0	True	0	2	2	1	...	0	0	0	0
15xxxxxx	5.4	2880	9	2	False	0	0	0	1	...	0	1	1	0
18xxxxxx	0.0	0	4	0	True	0	1	1	0	...	0	0	0	0
14xxxxxx	8.45	3060	11	0	False	0	0	0	1	...	0	1	1	1
15xxxxxx	8.57	3060	10	0	False	0	0	0	1	...	0	0	0	0
17xxxxxx	5.33	1020	7	0	False	0	1	1	1	...	0	0	0	0
17xxxxxx	9.24	2640	7	0	False	0	0	0	1	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12xxxxxx	4.0	360	3	0	True	1	2	3	0	...	0	0	0	0
14xxxxxx	7.85	3060	11	0	False	0	0	0	1	...	0	0	0	1
18xxxxxx	0.0	0	4	0	True	0	1	1	0	...	0	0	0	0
11xxxxxx	5.26	1140	11	4	True	1	3	4	1	...	0	0	0	0
16xxxxxx	7.3	2280	8	0	False	0	2	2	1	...	0	0	0	0
17xxxxxx	4.22	900	6	0	True	0	2	2	1	...	0	0	0	0
11xxxxxx	6.81	3000	16	0	True	0	0	0	1	...	1	1	2	0
12xxxxxx	0.82	60	4	0	True	1	1	2	0	...	0	0	0	0
11xxxxxx	0.05	0	7	0	True	3	2	5	0	...	0	0	0	0

Tabela 3.1: Recorte da base de dados inicial.

4

Resultados e Discussões

Desde a etapa de pré-processamento, todo o trabalho foi desenvolvido na linguagem Python, devido a facilidade de implementação, manutenção, aplicabilidade do mercado e meio acadêmico e simplicidade para aprofundamento do estudo em trabalhos futuros. Nos apoiamos nas bibliotecas Panda e Sklearn, amplamente utilizadas no âmbito de análise e mineração de dados e aprendizagem de máquina.

4.1 Resultados Experimentais Iniciais

Na primeira análise realizada, obtivemos um resultado de 0,77 de acurácia no modelo gerado através da Árvore de Decisão, enquanto 0,84 no modelo gerado através do algoritmo *Random Forest*. Isto é, 84% dos dados incluídos na base de teste foram classificados corretamente entre evasores e não evasores.

Analisando mais atentamente o resultado, com a expectativa de encontrar dados mais palpáveis para apoiar a coordenação do curso, buscamos compreender as variáveis de maior relevância em ambos os modelos, ou seja, as variáveis consideradas mais importantes para definição de um aluno como evasor ou não evasor. Os resultados obtidos, apresentados na tabela 4.1, nos fizeram questionar se não estavamos diante de um *overfitting* no caso do modelo da árvore de decisão, visto que as variáveis apresentadas como mais importantes são dependentes de outras variávies e em outra parte são pouco relevantes na experiência empírica do curso, além do grande volume de variáveis incluídas no estudo.

Também, pelo volume de variáveis utilizadas foi impossível extrair alguma informação analisando o fluxograma, ou seja, a árvore em si, gerada pelo algoritmo de Árvore de Decisão.

Já o resultado obtido com o algoritmo *Random Forest* parece mais promissor. Além de um resultado com maior acurácia, este algoritmo tem como variáveis mais relevantes as

Árvore de Decisão	***Random Forest***
Coeficiente	Coeficiente
Carga Horária Obtida	Carga Horária Obtida
Qtd de semestres válidos para integralização	Qtd de semestres válidos para integralização
Qtd de semestres trancados	Cálculo 1 - Qtd total de reprovações
Inglês - Status	Cálculo 1 - Status
Física 1 - Qtd reprovações por média	Programação 1 - Qtd total de reprovações
	Cálculo 2 - Status
	Sistemas Digitais - Status
	Química Tecnológica - Status

Tabela 4.1: Resultados iniciais: variáveis de maior importância

disciplinas que compõem o ciclo básico do curso e que, coincidentemente, possuem alto grau de reprovação.

4.2 Resultados Experimentais Aprimorados

Com os resultados acima citados, partimos para melhorar o que já havíamos conseguido executando os seguintes passos:

- Criação de novas variáveis;
  - número de semestres com matrícula e aprovação em pelo menos uma disciplina, para melhor identificar aqueles que fazem a matrícula e abandonam o semestre;
  - número de semestres com matrícula sem aprovação em nenhuma disciplina, para melhor identificar aqueles que fazem a matrícula e abandonam o semestre;
  - categorização do coeficiente do aluno por quartis, com o objetivo de categorizar os alunos baseado no desempenho acadêmico geral;
  - número total de reprovações, para entender o impacto do volume de reprovações;
  - % de reprovações por matrículas, com o objetivo de avaliar se a proporção entre número de reprovações e matrículas tem influência direta na evasão.
- Desconsiderar variáveis que entendemos gerar pouco conhecimento e informação;
  - carga horária obtida;
  - semestres válidos para integralização;
  - número de reprovações por média em cada disciplina;
  - número de reprovações por falta em cada disciplina, ou seja, considerar apenas o total de reprovações.

- Analisamos a matriz de confusão gerada para ambos os modelos.

Nesta segunda etapa da análise, os resultados obtidos indicam uma acurácia de 0,824 no modelo criado com a árvore de decisão e 0,906 no modelo criado através da *Random Forest*. Ou seja, após as mudanças, o segundo modelo foi de 84% para 90% de acurácia na classificação correta dos dados incluídos na base teste.

Além disso, analisamos a matriz de confusão para cada um dos modelos, na tabela 4.2 observamos que no caso do modelo gerado com a Árvore de Decisão obtivemos um resultado de 15 falsos negativos, isto é, evasores classificados incorretamente como não evasores, dentro de um universo de 98. Isso significa que classificamos incorretamente 15,3% dos dados relativos aos evasores na base de teste. No caso dos não evasores, temos 15 falsos positivos em um universo de 73, ou seja, 20,5%. Também, em 4.3 podemos observar outras métricas relacionadas ao modelo, onde entendemos que a mais importante aqui é *Recall*, já que ter falsos negativos - pessoas que evadem sendo classificadas como não evasores - é o mais danoso ao objetivo do estudo.

	Valor Predito: Evasão	**Valor Predito: Não Evasão**	**Total**
Valor Real: Evasão	TP = 83	FN = 15	98
Valor Real: Não Evasão	FP = 15	TN = 58	73
Total	98	73	171

Tabela 4.2: Matriz de Confusão: Árvore de Decisão

Métrica	**Resultado**
Precision	0,846
Recall	0,846

Tabela 4.3: Indicadores: Árvore de Decisão - Modelo Aprimorado

Já na tabela 4.4 observamos que no caso da *Random Forest* chegamos a um resultado de 10 falsos negativos, ou seja, 10,2% dos alunos evasores vieram a ser classificados incorretamente. No caso dos não evasores esta relação foi de 8,2%. Em 4.5 observamos mais alguns indicadores relacionados ao modelo e mais uma vez o foco aqui fica em *Recall*.

	Valor Predito: Evasão	**Valor Predito: Não Evasão**	**Total**
Valor Real: Evasão	TP = 88	FN = 10	98
Valor Real: Não Evasão	FP = 6	TN = 67	73
Total	94	77	171

Tabela 4.4: Matriz de Confusão: *Random Forest*

Além dos indicadores acima apresentados, também verificamos uma mudança nas variáveis de maior importância em cada um dos modelos após os ajustes, chegando nos resultados expressados na tabela 4.6, onde resumimos as variáveis com importância maior ou

Métrica	Resultado
Precision	0,936
Recall	0,897

Tabela 4.5: Indicadores: *Random Forest* - Modelo Aprimorado

igual a 0,02. Aqui, vale notar que em ambos os modelos a nova variável incluída, razão entre quantidade de reprovações e quantidade de matrículas, passou a ser a mais importante, inclusive com uma diferença grande das outras variáveis, principalmente no modelo da árvore de decisão. Vale notar também que ter um coeficiente acima do terceiro quartil parece ser determinante na permanência ou não de um aluno no curso, o que faz sentido se partimos da hipótese que poucos são os alunos com excelentes notas que, ainda assim, optarão por evadir.

Árvore de Decisão	***Random Forest***
Razão entre reprovações e matrículas	Razão entre reprovações e matrículas
Qtd de semestres sem aprovação	Qtd total de reprovações
Semestres trancados	Coeficiente - Acima do 3º quartil
LFAC - Qtd de reprovações	Cálculo 1 - Status
Cálculo 1 - Qtd de reprovações	Química - Status
Sistemas Digitais - Qtd de reprovações	Cálculo 1 - Qtd de reprovações
Qtd total de reprovações	Programação 1 - Qtd de reprovações
Matemática Discreta - Qtd de reprovações	Qtd semestres trancados
	Geometria Analítica - Qtd de reprovações
	Qtd de semestres com aprovação

Tabela 4.6: Resultados aprimorados: variáveis de maior importância

No caso do modelo gerado com o *Random Forest*, aparecem ainda mais disciplinas relacionadas ao ciclo básico do curso e no modelo gerado através da Árvore de Decisão, as variáveis de maior relevância ainda não nos entregam muita informação e/ou conhecimento palpável para suporte na tomada de decisão.

4.3 Resultados Experimentais Finais

Motivados pela intenção de buscar modelos com maior *recall* e que o modelo baseado na Árvore de Decisão seja mais palpável e gere uma árvore de mais fácil interpretação, além da suspeita que parte da complexidade está sendo gerada pelo alto volume de variáveis, executamos mais um ajuste na base, removendo todas as *features* com importância menor ou igual a 0,01 baseado no modelo *Random Forest*. Além disso, acrescentamos mais algumas métricas na análise que serão apresentadas no decorrer da seção.

Com os ajustes, o número de variáveis na base de dados foi reduzido a 28:

- Número de matrícula (posteriormente desconsiderado para a análise, tendo sido utilizado apenas como chave primária da base)
- Quantidade semestres trancados
- Evasão (sim/não)
- Quantidade total de reprovações
- Quantidade de semestres com aprovação
- Quantidade de semestres sem aprovação
- Razão entre quantidade de reprovações e quantidade de matrículas
- Cálculo 1 - Quantidade de reprovações
- Cálculo 1 - Status
- Geometria Analítica - Quantidade de reprovações
- Geometria Analítica - Status
- Programação 1 - Quantidade de reprovações
- Programação 1 - Status
- Introdução a Engenharia de Computação - Quantidade de reprovações
- Matemática Discreta - Quantidade de reprovações
- Matemática Discreta - Status
- Estrutura de Dados - Quantidade de reprovações
- Física 1 - Quantidade de reprovações
- Física 1 - Status
- Cálculo 2 - Quantidade de reprovações
- Álgebra Linear - Quantidade de reprovações
- Álgebra Linear - Status
- Circuitos Digitais - Quantidade de reprovações
- LFAC - Quantidade de reprovações
- Metodologia Científica - Status

- Química Tecnológica - Status
- Coeficiente - Acima do 3º quartil (sim/não)

4.3.1 Comparação entre os resultados apresentados

Na terceira e última etapa da análise, os resultados obtidos indicam uma acurácia de 0,836 no modelo criado com a Árvore de Decisão, superior aos resultados da obtidos na seção 4.2. Já no caso do modelo de *Random Forest* não houve mudança no indicador supracitado.

No que diz respeito aos resultados obtidos com os novos modelos, partindo a análise da árvore de decisão, observa-se que sua matriz de confusão apresenta uma leve mudança, como visto na tabela 4.7 que, em comparação com a tabela 4.2, apresenta tanto menos falsos negativos como menos falsos positivos, corroborando com o aumento de acurácia citado acima. Além disso, ao analisar a tabela 4.8 e comparar com a tabela 4.3 fica claro o crescimento de ambas as métricas de *precision* e *recall*.

	Valor Predito: Evasão	**Valor Predito: Não Evasão**	**Total**
Valor Real: Evasão	TP = 84	FN = 14	98
Valor Real: Não Evasão	FP = 14	TN = 59	73
Total	98	73	171

Tabela 4.7: Matriz de confusão: Árvore de Decisão - Modelo Final

Métrica	**Resultado**
Precision	0,857
Recall	0,857

Tabela 4.8: Indicadores: Árvore de Decisão - Modelo Final

No caso do modelo *Random Forest* observa-se que apesar de não ter havido mudança na acurácia, houve mudança na matriz de confusão e suas métricas, conforme observado nas tabelas 4.9 e 4.10. Comparando as tabelas 4.4 e 4.9 fica claro que apesar da melhora na classificação de alunos evasores houve piora na classificação de alunos não evasores, isto é, o modelo agora apresenta maior *recall* e menor *precision*, na comparação entre as tabelas 4.10 e 4.5.

Entende-se que o modelo melhorou neste caso, visto que para esta aplicação ter um melhor *recall* é o que se busca, já que o mais importante é garantir que o mínimo de alunos que vão evadir passem despercebidos, ou seja, sejam classificados incorretamente como não evasores.

Além do acima exposto, houveram mudanças nas *features* mais importantes, expostas na tabela 4.11 aquelas com importância maior ou igual a 0,02. Analisando a tabela, observa-se uma concentração de disciplinas do ciclo básico em ambos os modelos, o que

	Valor Predito: Evasão	**Valor Predito: Não Evasão**	**Total**
Valor Real: Evasão	TP = 90	FN = 8	98
Valor Real: Não Evasão	FP = 8	TN = 65	73
Total	98	73	171

Tabela 4.9: Matriz de confusão: *Random Forest* - Modelo Final

Métrica	**Resultado**
Precision	0,918
Recall	0,918

Tabela 4.10: Indicadores: *Random Forest* - Modelo Final

reforça a hipótese que é sumamente importante que o aluno consiga se encaixar desde o começo do curso para que permaneça no mesmo.

Árvore de Decisão	***Random Forest***
Razão entre reprovações e matrículas	Razão entre reprovações e matrículas
Qtd de semestres trancados	Qtd total de reprovações
Qtd total de reprovações	Coeficiente - Acima do 3º quartil
Qtd de semestres sem aprovação	Qtd de semestres sem aprovação
LFAC - Qtd de reprovações	Cálculo 1 - Status
Cálculo 1 - Qtd de reprovações	Geometria Analítica - Status
Geometria Analítica	Química - Status
Qtd de semestres com aprovação	Qtd de semestres trancados
Física 1 - Qtd de reprovações	Cálculo 2 - Status
Estrutura de dados - Qtd de reprovações	Qtd de semestres com aprovação
	Cálculo 1 - Qtd de reprovações
	Geometria Analítica - Qtd de reprovações
	Física 1 - Qtd de reprovações
	Matemática Discreta - Qtd de reprovações
	Cálculo 2 - Qtd de reprovações
	Matemática Discreta - Status

Tabela 4.11: Resultados finais: variáveis de maior importância

4.3.2 Análises complementares realizadas

Com o intuito de enriquecer ainda mais o estudo, foram agregadas algumas novas análises nesta terceira etapa:

- Quantidade de evasores e não evasores por intervalo de quantidade de reprovações;
- Percentual de evasores e não evasores com pelo menos uma reprovação e matrícula em Cálculo 1, Cálculo 2, Geometria Analítica, Álgebra Linear, Física 1, Física 2, Programação 1 e Estrutura de Dados;

- Percentual de evasores e não evasores por quartil de coeficiente do curso

Ao analisar a tabela 4.12, fica claro que o aumento do número de reprovações acompanha de forma consistente o crescimento da quantidade de evasores, ou seja, os intervalos com mais reprovações tem mais alunos evasores. Vale notar, também, que há uma queda da quantidade de evasores no intervalo de 7 a 10 reprovações, porém a relação entre quantidade de evasores e total de alunos neste intervalo segue em alta.

Quantidade de Reprovações	**0**	**1 a 3**	**4 a 6**	**7 a 10**	**mais que 10**
Evasores	2	10	98	55	170
Não evasores	33	68	47	15	69
Total	35	78	145	70	239

Tabela 4.12: Análises complementares: quantidade de alunos evasores e não evasores por quantidade de reprovações

Já em relação as tabelas 4.13 e 4.14, observa-se que há uma grande concentração de alunos evasores com pelo menos uma reprovação nas disciplinas analisadas referentes ao primeiro período do curso: Cálculo 1, Geometria Analítica e Programação 1. Também, é importante notar que a disciplina Cálculo 1 também apresenta um grande percentual de não evasores com pelo menos uma reprovação, chegando a quase 50% dos alunos contidos neste universo, o que traz o seguinte questionamento: o problema está no ensino superior ou na preparação do ensino médio?

Disciplina	**Cálculo 1**	**Cálculo 2**	**Geometria Analítica**	**Álgebra Linear**
Evasores	89,55%	17,61%	76,11%	26,86%
Não evasores	48,27%	30,17%	30,60%	21,98%

Tabela 4.13: Análises complementares: percentual de alunos evasores e não evasores com pelo menos uma reprovação na referida disciplina - Parte 1

Disciplina	**Física 1**	**Física 2**	**Programação 1**	**Programação 2**
Evasores	30,14%	8,05%	71,64%	25,37%
Não evasores	20,25%	16,37%	20,25%	28,87%

Tabela 4.14: Análises complementares: percentual de alunos evasores e não evasores com pelo menos uma reprovação na referida disciplina - Parte 2

No que diz respeito a distribuição dos alunos e seus coeficientes por quartis, podemos analisar na figura 4.1 e na tabela 4.15. Na tabela 4.16 fica claro observar que o coeficiente, que nada mais é que a média das notas adquiridas ao longo do curso, tem relação inversa com os evesaores, isto é, quanto maior o coeficiente, menos alunos evasores.

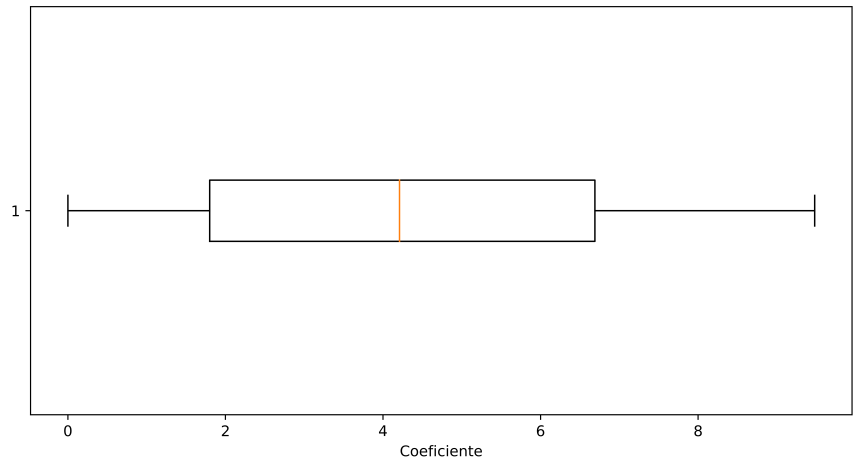

Figura 4.1: Análises complementares: distribuição dos coeficientes dos alunos

	Mínimo	1º Quartil (Q1)	2º Quartil (Q2)	3º Quartil (Q3)	Máximo
Coeficiente	0	1,8	4,21	6,69	9,48

Tabela 4.15: Análises complementares: distribuição dos coeficientes por quartil

4.4 Discussão

Nesta seção será discutido a respeito dos resultados obtidos nas seções 4.1, 4.2 e 4.3 e entender como os modelos apresentados podem vir a ser relevantes como suporte na tomada de decisão e redução dos índices de evasão.

4.4.1 Modelos

Ao comparar os modelos gerados pela Árvore de Decisão e pelo *Random Forest*, fica claro que o segundo apresenta-se como superior neste estudo, obtendo uma maior acurácia e também maior *recall*, que como explicado anteriormente, é a métrica mais relevante para esta aplicação.

Apesar disso, a Árvore de Decisão é mais facilmente interpretável, já que pode ser representada por um fluxograma e portanto possui uma representação gráfica. No caso apresentado a árvore gerada ainda possui muitos nós, conforme pode ser observado na figura 4.2[1], o que dificulta a interpretação e portanto necessita de maior aprofundamento e ajustes finos se a intenção for obter uma mais simplificada.

[1]Figura em alta qualidade: https://drive.google.com/file/d/12r-o3not94XzM9dOxSgWREMvLlBBvX9b/view?usp=share_link

	Abaixo de Q1	Entre Q1 e Q2	Entre Q2 e Q3	Acima de Q3
Evasores	137	118	68	12
Não evasores	5	24	73	130
Total	142	142	141	142

Tabela 4.16: Análises complementares: distribuição de evasores e não evasores por quartil

4.4.2 Variáveis de maior importância

Depois das distintas análises executadas e aprimoramentos feitos no modelo, observa-se que:

- O ciclo básico possui grande relevância na determinação da continuidade ou não do aluno no curso, com foco nas disciplinas de fora do IC;
- Permitir o alto volume de matrículas em disciplinas de alunos com histórico de muitas reprovações pode intensificar a evasão do curso, já que aumenta a razão entre quantidade de reprovações e quantidade de matrículas;
- Manter o aluno engajado desde o primeiro período e ao longo de toda a sua formação parece, também, ser determinante. Vale notar que uma das variáveis que aparece com maior relevância é a quantidade de semestres trancados, indicando que ao trancar o curso há grande chance de que o aluno não retorne;
- Mesmo no caso dos alunos com coeficiente acima do terceiro quartil observam-se casos de evasão, o que indica que a performance escolar não é o único fator de gerador de evasão.

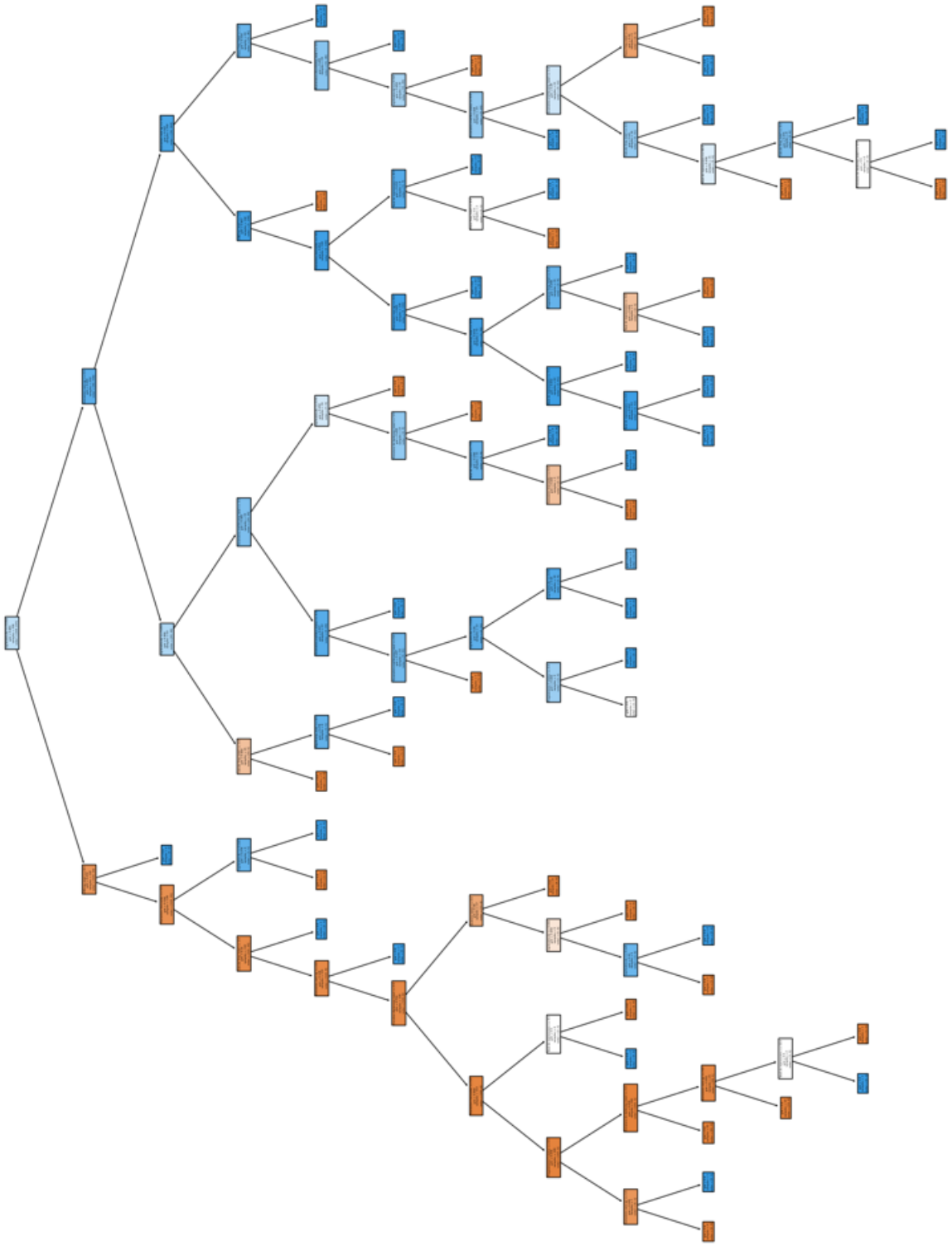

Figura 4.2: Análise Final: diagrama da árvore de decisão

5

Conclusão

Mostramos neste trabalho que é possível utilizar modelos preditivos baseados tanto em árvores de decisão como no algoritmo *Random Forest* como ferramentas de suporte à redução de evasão no curso de Engenharia de Computação do IC/Ufal. O modelo gerado através da Árvore de Decisão apresentou uma acurácia de 83,6% com *recall* de 85,7% enquanto o modelo gerado através do algoritmo *Random Forest* atingiu 90,6% e 91,8%. respectivamente. Desta forma, fica claro observar, que neste caso, o segundo modelo apresenta melhor desempenho no problema de pesquisa apresentado. Os algoritmos foram escolhidos objetivando principalmente uma maior facilidade de compreensão dos modelos.

Apesar do esforço, a árvore gerada pelo primeiro modelo ainda é bastante extensa, dificultando a extração simplificada de conhecimento apenas analisando o diagrama.

Vale salientar que analisando os atributos apontados como de maior importância observa-se que o ciclo básico do curso (disciplinas iniciais de cunho generalista e não específico) possui maior relevância, a razão entre número de reprovações e matrículas pode indicar que uma estratégia relevante para redução de evasão seja diminuir o volume de matrículas e garantir maior taxa de aprovação e que é importante manter o engajamento do aluno desde o seu primeiro dia no curso.

Além disso, como trabalhos futuros, indicamos:

- Utilização de outros algoritmos para geração e avaliação de novos modelos preditivos;
- Avaliação de outras estratégias para geração dos modelos, por exemplo: um modelo para cada semestre do curso, avaliar os alunos que vão evadir ou se formar e não somente evadir ou não evadir
- Inclusão de dados demográficos na análise, já que neste trabalho foram usados dados puramente acadêmicos;

- Realização de entrevistas com alunos que já evadiram do curso em diferentes etapas e por diferentes razões, objetivando compreender aspectos qualitativos da evasão para somar à análise quantitativa;
- Analisar a possibilidade de *overfitting* nos modelos apresentados e sua capacidade de generalização, bem como a utilização de técnicas de aumento de dados, redução de dados e validação cruzada.

Referências bibliográficas

[1] Leandro Algusto da Silva, Sarajane Marques Peres, and Clodis Boscarioli. 2016. *Introdução à Mineração de Dados com Aplicações em R.* Elsevier Editora Ltda.

[2] ANDIFES, ABRUEM, and SESu/MEC. 1996. 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* 1, 2 (ago. 1996). http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/739

[3] Taiwo Ayodele. 2010. *Types of Machine Learning Algorithms.* https://doi.org/10.5772/9385

[4] Danilo Giroldo. 2022. *Censo da educação superior: desafios e possibilidades.* Andifes. Retrieved January 01, 2023 from https://www.andifes.org.br/?p=94986#:~:text=Em%202021%2C%20havia%20apenas%2039,8%25%20entre%202012%20e%202016.

[5] CARLOS HENRIQUE DOMINGOS CORREIA SANTOS. 2021. Modelos Preditivos Baseados em Arvores de Decisão para um Sistema de Previsão de Evasão no Ensino Superior. (2021). Retrieved January 31, 2023 from https://app.uff.br/riuff/bitstream/handle/1/22786/TCC%20CARLOS%20HENRIQUE%20DOMINGOS%20CORREIA%20SANTOS.pdf;jsessionid=B69899914C673E0661234145DA991299?sequence=1

[6] Frederick Hess. 2018. *The College Dropout Problem.* Forbes. Retrieved October 03, 2021 from https://www.forbes.com/sites/frederickhess/2018/06/06/the-college-dropout-problem/?sh=52eb2b605fd2

[7] BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). 2022. Censo da Educação Superior 2021: divulgação de resultados. (2022). Retrieved January 01, 2023 from https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2021/apresentacao_censo_da_educacao_superior_2021.pdf

[8] BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). 2022. Censo da Educação Superior 2021: notas estatísticas. (2022). Retrieved January 01, 2023 from https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_da_educacao_superior_2021.pdf

[9] Gustavo Ioschpe. 2016. *A ignorância custa um mundo: o valor da educação no desenvolvimento do Brasil.* Objetiva.

[10] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge: Cambridge University Press.

[11] LUIZ EDUARDO KOCHHANN. 2022. *Brasil deve atingir meta de matrículas no ensino superior apenas em 2040.* Desafios da Educação. Retrieved January 01, 2023 from https://desafiosdaeducacao.com.br/brasil-meta-matriculas-ensino-superior/

[12] P. Norvig and S. Russell. 2013. *Inteligência Artificial.* ELSEVIER EDITORA. https://books.google.com.br/books?id=KhUQvgAACAAJ

[13] REDAÇÃO. 2022. *Evasão bate recordes no ensino superior.* Desafios da Educação. Retrieved January 01, 2023 from https://desafiosdaeducacao.com.br/evasao-bate-recordes-no-ensino-superior/

[14] Cristiane Aparecida dos Santos Baggi and Doraci Alves Lopes. 2011. *Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica.* Revista da Avaliação da Educação Superior (Campinas). Retrieved October 03, 2021 from https://doi.org/10.1590/S1414-40772011000200007

[15] Isabel Cristina Torres-Patiño, Cristhiam Mauricio Rojas-Hernandez, and Herney Andres García-Perdomo. 2021. Barreiras de acesso e permanência na universidade: um olhar. (2021). Retrieved January 01, 2023 from https://journal.einstein.br/pt-br/article/barreiras-de-acesso-e-permanencia-na-universidade-um-olhar/#:~:text=Em%20pa%C3%ADses%20desenvolvidos%2C%20como%20Espanha,de%2010%25%20a%2025%25.

[16] Prof. Éliton Fontana. 2020. *Introdução aos Algoritmos de Aprendizagem Supervisionada.* UFSC. Retrieved January 30, 2023 from https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf