Modelo para Predição da evasão escolar no curso de Ciência da Computação do IFSUL – Câmpus Passo Fundo utilizando Redes Neurais Artificiais

Gustavo Pedroso Gal[1]
Prof. Dr. João Mário Lopes Brezolin[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia
SUL-rio-grandense - Campus Passo Fundo

gustavogal.pf159@academico.ifsul.edu.br

Abstract. *Studies reveal that higher education courses in the area of Information Technology are among those most affected by dropout rates in Brazil. Bearing in mind that this affects both the students who drop out and the people responsible for their education, this work aims to create a machine learning algorithm that helps determine a profile of students prone to dropping out. Data from students at the Federal Institute of Education, Science and Technology SUL-rio-grandense - Campus Passo Fundo, currently enrolled in the Bachelor's degree in Computer Science, will be used.*

Resumo. *Estudos revelam que os cursos de ensino superior na área de Tecnologia da Informação estão entre os mais afetados pelas taxas de desistência no Brasil. Tendo em mente que isso afeta tanto os alunos que realizam a evasão quanto as instituições responsáveis pela sua educação, esse trabalho tem como objetivo criar um algorítmo de aprendizado de máquina que ajude a determinar um perfil de alunos propensos a evasão. Foram utilizados dados de alunos do Instituto Federal de Educação, Ciência e Tecnologia SUL-rio-grandense - Campus Passo Fundo, atualmente matriculados no curso de Bacharelado em Ciência da Computação.*

1. Introdução

A área da Computação apresenta-se como uma das mais afetadas pela evasão escolar em cursos superiores. Estudos mostram que a taxa de evasão calculada para cursos presenciais de Sistemas de Informação era de 37,6% em instituições da rede privada (SEMESP, 2021). O levantamento realizado pelo Ministério da Educação (MEC) em 2019 demonstrou que nas universidades federais, essa taxa esteve em 30,8% (PODER360, 2018). Esse cenário impacta o mercado de trabalho pois ocasiona a falta de profissionais da área, assim como afeta também o ambiente escolar, onde se observa uma escassez de alunos principalmente nas etapas mais avançadas dos cursos, além disso, o próprio evasor acaba se vendo frustrado pela sua situação de abandono do curso (HOED, 2016 apud CUNHA; NASCIMENTO; DURSO, 2016).

Com o avanço da tecnologia e da capacidade de processamento das máquinas, surgiram as Redes Neurais Artificiais (RNA), criadas inicialmente para a realização de tarefas relacionadas ao aprendizado de máquina, como reconhecimento de padrões e o

processamento de linguagem natural. Nesse sentido, o presente trabalho tem como objetivo avaliar o uso de Redes Neurais Artificiais no desenvolvimento de um modelo, que identifique quais são os fatores que mais influenciam o aluno na tomada de decisão acerca da desistência do curso.

O presente artigo está estruturado como segue: Na seção 2 é realizada a revisão bibliográfica onde foram apresentados conceitos dos temas presentes nesse trabalho, na seção 3 consta as tecnologias utilizadas para o desenvolvimento do modelo de inteligência artificial. Foram apresentados na seção 4 os trabalhos relacionados aos principais temas deste artigo, que serviram como fundação para o planejamento da metodologia a ser usada, que foi descrita na seção 5 e, por fim, os resultados são apresentados na seção 6 do presente artigo.

2. Revisão Bibliográfica

Nesta seção, está presente uma revisão da literatura, descrevendo os conceitos sobre tópicos e tecnologias relevantes ao artigo, e a problemática do fenômeno da evasão do ensino superior com foco na área da computação. Esta revisão orienta o trabalho, e fundamenta a abordagem metodológica utilizada.

2.1. Evasão Escolar no Ensino Superior

A evasão escolar de cursos superiores é um tema já conhecido em todo país, mostrando-se como um problema não só para os primariamente afetados, os estudantes, mas também para autoridades e gestores da área da educação. O número de trancamentos de matrículas em cursos à distância aumentou 35,6% de 2019 para 2020, outro número alarmante foi o crescimento desse dado em redes públicas, que saltou 94,5% no mesmo período (SEMESP, 2022). Podemos relacionar esse aumento com a pandemia da COVID-19 que teve seu início no ano de 2020, inviabilizando em um primeiro momento as aulas em todo o país. Os cursos da área da Ciência da Computação estão, como já dito anteriormente, tanto na rede privada quanto na pública, entre os que mais sofrem com o fenômeno da evasão, por outro lado, segundo dados da CNN Brasil (2021), a procura por profissionais dessa área cresceu 671% durante a pandemia, criando uma demanda que ficará cada vez mais difícil de ser preenchida devido essa contínua evasão dos estudantes da área.

Esse fenômeno não se limita somente ao Brasil, estudos semelhantes foram realizados em diversos países, trazendo à tona não somente a evasão na área da Ciência da Computação, como também na área denominada de STEM (Ciência, Tecnologia, Engenharia e Matemática). Lee e Ferrare (2019) observaram o tema em seu país, e notaram que estudantes matriculados em cursos dessas áreas possuem uma chance significativamente maior de obterem um diploma do que pessoas que iniciam e optam por trocar para um curso fora da área chamada STEM.

Hurka, Meelissen e Langen (2019, tradução nossa) concluíram que ”[...] parece que medidas com foco em conhecimento, habilidade, motivação e sentimentos de pertencimento podem aumentar o interesse e a persistência na educação em áreas da STEM”. Nota-se então que medidas podem ser tomadas para a mitigação do problema, porém devem ser tomadas com antecedência, já que na maioria das vezes só é descoberto o fenômeno de forma tardia, quando já ocorreu a evasão ou o aluno já está decidido a abandonar o curso.

2.2. Inteligência Artificial

O conceito de Inteligência Artificial (IA) pode ser definido como uma área de estudo dentro da tecnologia com o foco em agentes que são concebidos com um objetivo, alimentados com informações, e com elas entregam ações ou ideias (RUSSEL e NORVIG, 2013). Sistemas de IA são projetados para adquirir conhecimento, raciocinar, compreender, planejar, aprender e tomar decisões com base em dados e informações disponíveis. A área da IA busca emular a capacidade humana de processar informações complexas e resolver problemas de maneira eficiente, utilizando métodos computacionais avançados, como algoritmos de aprendizado de máquina, redes neurais artificiais e lógica simbólica. A aplicação da IA abrange diversas áreas, como medicina, finanças, robótica, educação, entre outras.

2.2.1. Aprendizado de Máquina

O aprendizado de máquina, por sua vez, como disse Arthur Samuel em 1959, é um subconjunto da IA que se concentra no desenvolvimento de algoritmos que permitem máquinas a aprender com os dados disponíveis, e façam previsões ou decisões sem serem explicitamente programadas para isso (SIMON, 2013). Existem diferentes tipos de abordagens de aprendizado de máquina, incluindo o aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço, são aplicados em contextos de reconhecimento de imagem e fala, processamento de linguagem natural e análise preditiva.

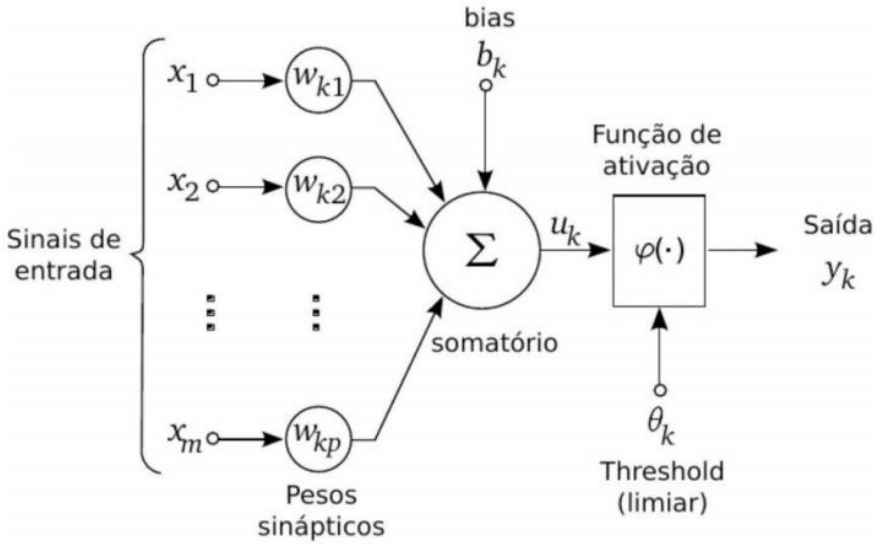
No que diz respeito aos diferentes métodos de aprendizado de máquina, eles se distinguem pela forma como processam conjuntos de dados. No aprendizado supervisionado, o algoritmo recebe um conjunto de dados juntamente com as respostas corretas que ele deve fornecer. Por outro lado, no aprendizado não supervisionado, o fornecedor dos dados não possui conhecimento prévio das respostas, e cabe ao algoritmo identificar relações entre as entradas e chegar em conclusões significativas. Para ambos os métodos, a eficácia do algoritmo aumenta proporcionalmente à quantidade de dados disponíveis. Existe também o método de aprendizado por reforço, frequentemente utilizado em ambientes controlados e simulações, nesse método, o objetivo é que o aprendizado ocorra por meio de tentativa e erro, incentivando o algoritmo a tomar decisões e penalizando aquelas que obtém resultados negativos, enquanto recompensa decisões positivas, dessa forma, o algoritmo aprende gradualmente o conjunto ideal de decisões que devem ser tomadas.

2.2.2. Redes Neurais Artificiais

Segundo BISHOP (1995, p.227, tradução nossa): ”Uma rede neural artificial (RNA) pode ser considerada como uma função matemática não linear que transforma um conjunto de variáveis de entrada em um conjunto de variáveis de saída”, com o progresso na área de estudos de IA suas aplicações expandiram-se e hoje são amplamente utilizadas para problemas dentro da área da pesquisa científica. As RNAs foram desenvolvidas com base na estrutura e no funcionamento do cérebro humano, em particular nos neurônios biológicos, buscam imitar os processos de aprendizado e processamento de informações do cérebro, onde os neurônios estão interconectados para transmitir e processar sinais.

Na Figura 1 pode-se observar um diagrama da arquitetura de uma rede neural artificial, onde os seus sinais de entrada são os dados ingeridos pelo modelo, e o seu limiar é o valor usado em classificações binárias para escolher entre duas classes, tipicamente 0 e 1.

Figura 1. Diagrama da arquitetura de uma RNA



Fonte: Haykin, 2001

Cada neurônio artificial recebe uma entrada, realiza uma computação e produz uma saída, sendo capaz de capturar relações não lineares e lidar com dados de alta dimensionalidade. Essa interação entre os neurônios é estabelecida através das conexões existentes, onde pesos são atribuídos a cada conexão, esses pesos multiplicam o valor do neurônio e o encaminham para o próximo neurônio na rede, resultando em uma modificação dos valores iniciais à medida que avançamos nas camadas da RNA. Além dos pesos, as funções de soma também desempenham um papel importante na matemática das RNAs, essas funções são aplicadas entre as camadas da rede e são responsáveis por combinar os valores de saída dos neurônios de uma camada e fornecer as entradas para os neurônios da camada seguinte, essa etapa de soma e transferência dos valores é fundamental para o processamento e propagação das informações pela rede neural.

A combinação das operações de multiplicação dos pesos e das funções de soma possibilita que as RNAs realizem cálculos complexos e aprendam a partir dos dados. Ao ajustar os pesos das conexões durante o processo de treinamento, a RNA pode adaptar-se e encontrar padrões relevantes nos dados de entrada, permitindo a resolução de diversos problemas, como reconhecimento de padrões, classificação, regressão, entre outros.

3. Tecnologias utilizadas no desenvolvimento do modelo

Nesta seção estão detalhadas as tecnologias utilizadas na criação do modelo de aprendizado de máquina, além de uma breve explicação sobre seus conceitos.

3.1. Python

Python é uma linguagem de programação que oferece não apenas simplicidade e legibilidade, mas também se destaca no campo de aprendizado de máquina, contando com um ecossistema robusto de bibliotecas especializadas, como TensorFlow e PyTorch. Essas ferramentas simplificam o desenvolvimento de modelos complexos, e também permitem que os desenvolvedores se concentrem no design e na personalização, inserindo parâmetros de forma eficiente para construir redes neurais e soluções de aprendizado de máquina.

Além disso, como deixa claro seu criador, a versatilidade do Python estende-se abrangendo uma gama ampla de aplicações, desde análise de dados e desenvolvimento web até automação. Por esses fatores, Python se apresenta como uma linguagem de programação sólida para a construção de soluções robustas e inovadoras em diferentes campos (VENNERS, 2007).

3.2. Tensorflow

Atualmente ferramentas como o Tensorflow, projeto da Google Brain iniciado em 2011 com objetivo de oferecer uma variedade de algoritmos de treinamento e inferência de redes neurais, nas mais diversas áreas de aplicação como robótica, visão computacional, processamento de linguagem natural, entre outras (ABADI et al., 2015), combinadas com linguagens de programação como Python, oferecem a possibilidade da construção manual de redes neurais, pois se encarregam da criação da parte de menor customização do modelo e o encapsulando em funções que permitem ao desenvolvedor um foco maior no design e na escolha das variáveis do seu modelo por meio da inserção de parâmetros nessas funções.

3.3. Scikit Learn

Scikit-learn, comumente referido como Scikit, é uma popular biblioteca de aprendizado de máquina de código aberto para Python. Ela fornece uma ampla gama de ferramentas e algoritmos para várias tarefas em aprendizado de máquina, incluindo classificação, regressão, agrupamento, redução de dimensionalidade e seleção de modelo (PEDREGOSA et al., 2011).

Um dos principais pontos fortes da Scikit é sua coleção abrangente de algoritmos de aprendizado de máquina. Ela oferece funções para aprendizado supervisionado e não supervisionado, como máquinas de vetores de suporte (SVMs), florestas aleatórias, agrupamento k-means e análise de componentes principais (PCA).

3.4. Grid Search

O *Grid Search*, ou busca em grade, é uma técnica fundamental na otimização de hiperparâmetros em modelos de aprendizado de máquina. Quando se constrói um modelo, é comum ter parâmetros que não são aprendidos durante o treinamento, mas que precisam ser configurados de forma manual, esses parâmetros são chamados de hiperparâmetros e influenciam diretamente o desempenho e a capacidade de generalização do modelo.

A técnica de *Grid Search* consiste em definir um conjunto de valores possíveis para cada hiperparâmetro e avaliar o desempenho do modelo para todas as combinações possíveis desses valores. Isso é feito por meio de validação cruzada, onde o conjunto de

dados é dividido em subconjuntos de treinamento e teste em várias iterações. Para cada combinação de hiperparâmetros, o modelo é treinado nos dados de treinamento e avaliado nos dados de teste, e o desempenho é registrado.

É uma abordagem sistemática e exaustiva, garantindo que todas as combinações possíveis sejam consideradas, isso ajuda a encontrar a combinação de hiperparâmetros que resulta no melhor desempenho do modelo.

3.5. Google Colaboratory

O Google Colaboratory, ou simplesmente Google Colab, é uma plataforma de código aberto baseada na nuvem desenvolvida pelo Google. Apresentando-se como um ambiente colaborativo em Python, o Colab oferece uma maneira fácil e eficiente de escrever, executar e compartilhar código.

Uma característica distintiva do Colab é a sua integração direta com o Google Drive, permitindo o armazenamento e compartilhamento simples de projetos. Além disso, oferece acesso a recursos computacionais poderosos, como GPUs e TPUs, sem a necessidade de configurações complexas.

No contexto do trabalho, o Google Colab será a plataforma escolhida para o desenvolvimento e execução de modelos de aprendizado de máquina, aproveitando sua facilidade de uso, colaboração e recursos computacionais disponíveis na nuvem.

4. Trabalhos Relacionados

Nesta seção, constam estudos que contribuíram diretamente para o entendimento do fenômeno da evasão na área da tecnologia, analisando metodologias utilizadas e resultados obtidos com o uso das mesmas.

O estudo de Hoed (2016) oferece uma análise quantitativa da evasão em cursos superiores da área da Computação utilizando dados obtidos via Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) no escopo de instituições públicas e privadas, tendo como resultado o encontro de uma taxa anual de evasão de estudantes. Além disso também é realizado um estudo qualitativo usando questionários entregues para estudantes evadidos de cursos superiores da área da Computação. Relacionando esses dois métodos obtiveram resultados que ilustram a ligação entre a evasão e o requerimento de conhecimentos matemáticos e algorítmicos do curso, além do apontamento de outros fatores diretamente ligados a esse fenômeno evasivo, como o sexo, forma de ingresso e a participação do indivíduo em bolsas destinadas para cotistas.

Outro estudo a ser notado engloba áreas da educação dentro do sistema conhecido como *STEM* (ciência, tecnologia, engenharia e matemática), que inlcui por consequência também a área da Computação. Aponta para dificuldades dependendo do fato do aluno ser bolsista, do sexo e da idade do estudante, com mulheres de maior idade apresentando níveis superiores de dificuldades (CASANOVA et al., 2023).

Na UFERSA (Universidade Federal Rural do Semiárido) foi realizado um estudo de caso sobre a evasão dos alunos matriculados no curso de Ciência da Computação, buscando identificar os motivos da alta taxa de evasão acadêmica, extraindo informações de uma base de dados fornecida pela instituição e apresentando uma ideia da existência de fatores internos às instituições que influenciam na evasão acadêmica, como currículos

ultrapassados, a falta de clareza sobre o projeto pedagógico, estruturas ultrapassadas e métodos inadequados de avaliação do desempenho do aluno (MARQUES et al., 2020).

Em outro trabalho, foram obtidos dados demográficos e acadêmicos de alunos por Melo, Souza e Santos (2022) através do sistema utilizado no IFMG fornecido pela empresa TOTVS, além de dados presentes em planilhas eletrônicas. A preparação dos dados foi necessária para a anonimização de seus donos, a exclusão de dados sobre uma pessoa caso tenha um valor em branco no lugar de uma informação essencial, além da exclusão de valores duplicados. Ainda sobre o tratamento dos dados, alguns atributos tiveram seu nome modificado para melhor representar o seu significado, e outros atributos foram criados pelos autores e tiveram seus valores devidamente calculados. O estudo fez uma análise da importância das variáveis referentes aos alunos, o percentual de reprovações no primeiro semestre letivo foi a principal causa direta na evasão, com cerca de 27% de importância. O segundo atributo indicado também apresenta alta importância (25%) e refere-se ao coeficiente de rendimento do estudante durante o primeiro semestre letivo, enquanto o terceiro atributo indica a carga horária total do curso.

Alban e Mauricio (2019) apresentaram também uma proposta de utilização de Redes Neurais para a predição do nível de evasão em um curso de nível superior, coletando dados de 2670 alunos matriculados em diversos cursos na Universidade do Equador. Os dados foram coletados por meio de um questionário online disponibilizado para os alunos matriculados no período entre o primeiro e o quarto ano acadêmico, informações coletadas incluem dados demográficos, padrões de comportamento, informações sobre a universidade e dados socioeconômicas do entrevistado. Os dados receberam um tratamento para a limpeza de informações redundantes e de campos vazios. A rede neural desenvolvida pelos autores obteve uma acurácia de 96,8% no conjunto de testes, se apresentando como uma melhor opção comparada ao outro algoritmo explorado.

Em Baranyi, Nagy e Molontay (2020) é encontrado uma amostragem maior de dados de alunos matriculados na Budapest University of Technology and Economics, após realizadas as transformações e a limpeza dos dados os autores obtiveram informações referentes a 8.319 estudantes presentes entre 2013 e 2019 e que tiverem seus estudos interrompidos por motivos de evasão ou de graduação na universidade. Esse trabalho usou uma extensa lista contendo 27 atributos sobre os estudantes observados, atributos que foram divididos em classes, referentes a informações do estudante, performance acadêmica anterior a faculdade, detalhes pessoas, dentre outros.

Na proposta de Viana et al (2022) observou-se a utilização de modelos treinados em cada janela semestral para a predição da evasão em ensino superior, com base nos dados de dois cursos da Universidade Federal do Piauí (UFPI), Computação e Sistemas de Informação. Foram utilizados algoritmos de aprendizado de máquina por meio da biblioteca Scikit-Learn, na linguagem de programação Python. Foram realizados os treinamentos e testes utilizando a técnica de validação cruzada, alternando os dados utilizados no treinamento e no teste em cada iteração do treinamento do algoritmo. Sobre os resultados, o algoritmo que obteve a maior acurácia na predição foi o *Random Forest* quando usado para avaliar alunos do 5º período, porém ele não somente foi o que obteve a maior acurácia em um período único mas o que apresentou a maior média de acurácia entre os períodos dentre todos os algoritmos (VIANA; SANTANA; RABÊLO, 2022).

Foram encontrados trabalhos e estudos realizados na área da Ciência da Computação, que investigam o fenômeno da evasão escolar em instituições públicas e privadas, tendo como objetivo encontrar os motivos que influenciam os estudantes a chegarem ao ponto de evadir instituições de ensino superior, e revelam que existem dificuldades encontradas na ligação do curso com a matemática, sendo necessário um nível maior de empenho e conhecimento requerido ao ingressar na área, e esse desafio está presente nos semestres iniciais dos cursos da área, explicando também a influência direta da nota do estudante referente ao primeiro semestre com uma possível desistência do curso.

Sobre a extração e manipulação dos dados referentes a alunos e ex-alunos de universidades, se diferenciou um trabalho para o outro principalmente na quantidade e na divisão proposta dos atributos escolhidos, usados em algoritmos e análises na tentativa de traçar um perfil do aluno evadido.

Acerca das tecnologias e métodos utilizados para o processamento desses dados, foram encontrados diversos trabalhos com propostas e resultados diferentes, muito por conta da diferença entre amostragem de dados disponíveis em cada um dos estudos, os resultados também podem variar de acordo com o tamanho da porção de dados destinados a avaliação do algoritmo, entretanto no geral constatou-se um desempenho satisfatório da maioria dos trabalhos encontrados, reforçando a ideia da boa performance de algoritmos de inteligência artificial em cenários de predição de valores.

5. Desenvolvimento do Sistema Proposto

Utilizando como referência o estudo de Colpo, Primo e Aguiar (2021), foi desenvolvido um documento contendo uma lista de dados específicos de alunos, com intuito de serem utilizados no desenvolvimento desse trabalho. Foram então exportados os dados de 100 estudantes por meio da plataforma SUAP (Sistema Unificado de Administração Pública) em uma planilha, com as seguintes colunas:

A coleta de dados descrita na tabela 1 foi realizada utilizando dados não-identificados de alunos do curso de Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia SUL-rio-grandense (IFSUL).

Os atributos pessoais têm como objetivo de investigar, se características não relacionadas ao desempenho acadêmico estão ligadas ao fenômeno da evasão. O termo ”Estudante com necessidades específicas”, refere-se a pessoas que são o público-alvo da educação especial, ou seja, estudantes com deficiências, transtornos, superdotações. Já a variável ”Situação no Curso”, representa se o aluno ainda continua matriculado no curso, se graduou ou realizou a evasão.

O conjunto de variáveis seguinte medem o desempenho do aluno no seu primeiro semestre matriculado na instituição, esses dados são relevantes, pois é nesse período que muitos estudantes têm seu primeiro contato com o funcionamento do curso. As variáveis apresentam desde a sua frequência nas aulas, a quantidade de disciplinas em que o aluno se matriculou e em quantas ele foi aprovado ou reprovado. São requisitados dados também de qual foi o seu desempenho nas cinco disciplinas iniciais do curso, para assim saber se o desempenho dos estudantes em uma disciplina específica afeta a sua jornada acadêmica. E por final a variável ”C.R.”, representa a média de notas do aluno durante sua carreira acadêmica.

Tabela 1. Atributos escolhidos para seleção dos dados.

Grupo	Variáveis
Atributos pessoais	Idade Gênero Etnia Estado de Nascimento Renda Familiar Estudante com necessidades específicas Situação no Curso
Primeiro Semestre	% Frequência Disciplinas Matriculadas Disciplinas Aprovadas Disciplinas Reprovadas Nota Circuitos Digitais Nota Algoritmos Nota Matemática Discreta Nota IHCC Nota CPW C.R.
Último Semestre Completo	% Frequência Disciplinas Matriculadas Disciplinas Aprovadas Disciplinas Reprovadas C.R.

Fonte: Elaborada pelo autor, 2023

O último semestre completo visa mostrar o desempenho mais recente do estudante, uma vez que alguns alunos podem começar com um desempenho considerado ideal, mas ao longo do curso acabam se tornando evasores. Tendo mapeado o primeiro e último semestre, é possível ter uma ideia se o aluno regrediu ou evoluiu o seu desempenho acadêmico, outros semestres poderiam estar presentes nas pesquisas, porém para que a maior quantidade de alunos possível seja comparada com os mesmos parâmetros, foi escolhido o último semestre completo tendo em vista que não é possível definir um semestre que todos os alunos tenham cursado além do primeiro, pelo mesmo motivo não foram avaliadas as notas de disciplinas específicas no último semestre completo.

5.1. Pré-processamento dos dados

Os dados disponibilizados foram carregados no ambiente do Google Colaboratory e, em seguida, após serem analisados, foi observado a necessidade de uma etapa de pré-processamento para que os valores fossem introduzidos em um algoritmo de aprendizado de máquina.

Posteriormente as variáveis foram transformadas em tipo numérico, etapa essencial em toda manipulação de dados visando a inserção em algorítmos de aprendizado de máquina, porém nas ocorrências de ”Renda per Capita”, onde já se encontravam nessa formatação, apenas os valores nulos foram convertidos para 0. As porcentagens

de frequência foram convertidas em valores entre 0,0 e 1,0, removendo o caractere ”%”, e substituindo a vírgula por ponto.

Foi realizado o processamento da variável contendo a situação do curso, pois existiam diversos valores diferentes informando a situação da matrícula do aluno, como ”Evasão”, ”Cancelamento Compulsório”, ”Trancado Voluntariamente”, ”Transferido”, entre outros. Todas essas categorias foram transformadas em ”Descontinuação”, tendo em mente que o aluno não estava matriculado no semestre atual, para os casos em que o aluno não realizou a evasão, como em ”Matriculado”e ”Graduado”, foram transformados em ”Continuação”.

Sobre a quantidade de dados dos alunos que continuaram ou descontinuaram o curso, observou-se uma diferença de 36,3% a mais de alunos matriculados contra evadidos, considerada aceitável para a rede neural, eliminando a necessidade de técnicas de tratamento de desbalanceamento de rótulos. Os dados foram padronizados usando a classe *StandardScaler* do Sklearn, garantindo média zero e desvio padrão próximo de 1 para a correta avaliação do modelo, e finalmente, divididos em conjuntos de treinamento e validação.

5.2. Desenvolvimento e comparativo entre modelos de classificação

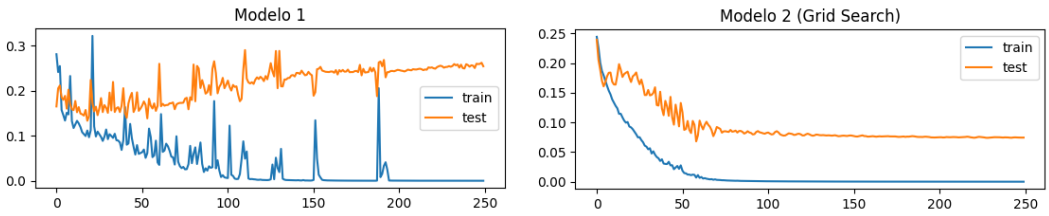
Para fins de comparação, optou-se pela utilização de dois modelos diferentes de redes neurais, fazendo com que a escolha do modelo final se baseie na melhor performance entre os dois quanto a sua acurácia na predição da situação dos estudantes. A primeira rede neural foi criada manualmente, utilizando uma arquitetura previamente testada em situações de classificações, caracteriza-se pelo uso do otimizador RMSprop e a utilização do cálculo de MSE (erro quadrático médio) como função de perda, essa métrica calcula a média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais, ao elevar as diferenças ao quadrado, o MSE destaca discrepâncias mais significativas, sendo sensível a erros maiores.

Na criação da segunda rede neural, foi utilizada a biblioteca Kerastuner, que realiza o processo de *Grid Search*, testando várias arquiteturas diferentes para descobrir qual delas tem a melhor performance diante dos dados disponíveis. Nesse processo, foram exploradas redes com diferentes quantidades de camadas, quantias de neurônios por camada, otimizadores e funções de perda. Para contornar o problema em que o modelo selecionado como vencedor era o que performava melhor apenas na última época de validação, foi colocada uma função de *callback*, onde caso o modelo não melhore sua perda em no máximo 10 épocas durante seu treinamento, o aprendizado será encerrado. O treinamento dos modelos foi realizado após a divisão dos dados, destinando 70% da população para treinamento e os outros 30% para a validação da acurácia após o treinamento.

6. Resultados e discussões

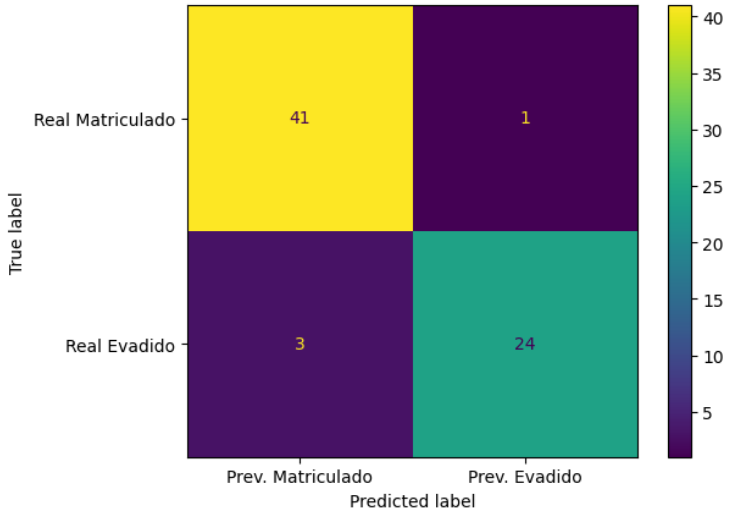
A matriz de confusão é uma ferramenta importante para avaliar as previsões de um modelo em relação à quantidade de dados disponíveis para cada classe. Organizada em linhas e colunas, as respostas verdadeiras são armazenadas nas linhas, enquanto as predições do modelo ficam nas colunas. Para facilitar a compreensão, os acertos do modelo podem ser identificados ao longo da diagonal principal da matriz.

Figura 2. Valor da perda durante treinamento e validação



Fonte: Elaborado pelo autor, 2023

Figura 3. Matriz de confusão do conjunto de dados completo



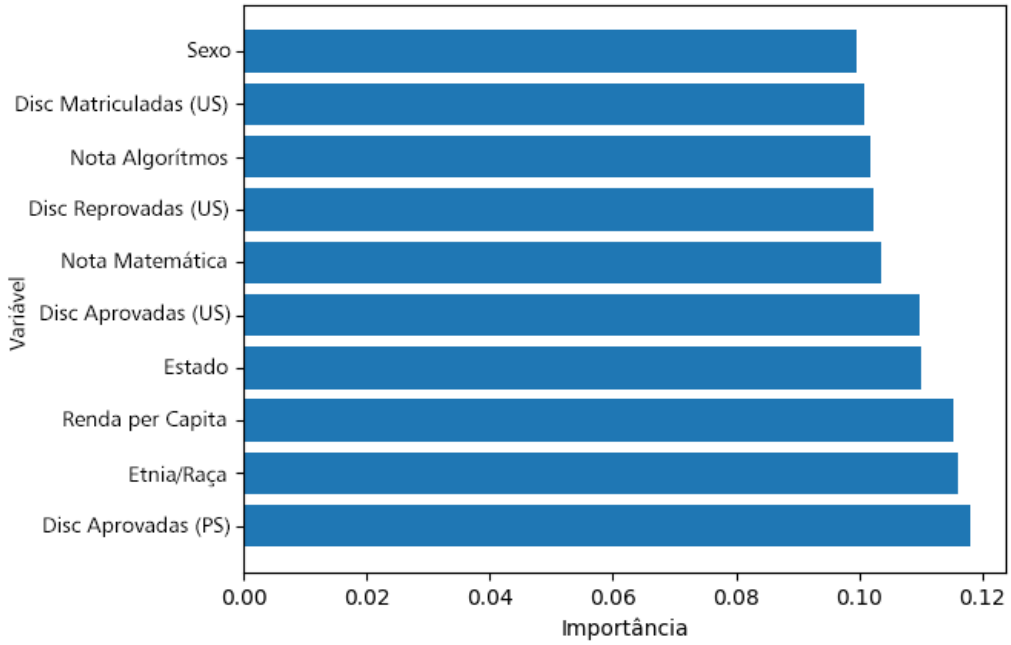
Fonte: Elaborado pelo autor, 2023

A Figura 3, apresenta a matriz de confusão com as predições abrangendo todo o conjunto de dados. Na primeira linha da matriz, encontram-se as informações relacionadas à classe ”matriculado”, a primeira coluna representa quando o modelo corretamente fez a predição como matriculado, já na segunda coluna estão os casos de alunos que continuaram matriculados mas o modelo fez a predição da evasão. Já na segunda linha, dados referentes à classe ”evadido”, as colunas seguem a mesma lógica, representando as predições do modelo para cada classe.

Observa-se que o modelo demonstra uma notável capacidade de acertar casos de alunos que continuam matriculados, com um desempenho ligeiramente inferior nos casos de alunos evadidos, com 24 acertos e 3 erros.

Após a identificação do modelo com a melhor acurácia, foi realizada a observação da importância das 10 variáveis mais importantes no cálculo da chance de evasão de um estudante, cálculo esse que se baseia nos pesos presentes no modelo de rede neural, em cada um dos neurônios ligados à variável em observação. Para a leitura do gráfico, deve-se levar em consideração que as variáveis com a nomenclatura ’US’ referem-se ao último semestre cursado, e no caso de ’PS’ presente na sua nomenclatura, consequentemente representam as informações do primeiro semestre.

Figura 4. Importância das variáveis na evasão



Fonte: Elaborado pelo autor, 2023

Com a leitura da Figura 4, é possível analisar que um fator crucial para a permanência do estudante foi a quantidade de disciplinas em que conseguiu ser aprovado no primeiro semestre, após essa variável, observa-se a presença de dados representando características pessoais, como etnia, renda per capita da família e estado de nascimento (Figura 4). Além disso, é relevante observar que, após essas características pessoais, retornamos à análise de variáveis relacionadas ao desempenho escolar do aluno, desta vez no último semestre completo, seguidas por variáveis que representam suas notas nas primeiras disciplinas do curso. Sobre a diferença dos valores de importância, como mostrado no trabalho de Melo, Souza e Santos (2022) e também Alban e Mauricio (2019), a diferença entre os valores acaba sendo mínima aos olhos humanos, porém, os mesmos pesos associados às variáveis representam para o modelo a força e a direção da influência de cada variável na saída do modelo, fazendo com que uma pequena diferença no seu valor resulte em uma grande diferença no cálculo final devido a quantidade de camadas em uma Rede Neural.

Conclui-se, portanto, com base no gráfico apresentado, que o primeiro semestre exerce uma influência significativa nas decisões dos alunos sobre a continuidade no curso, pois em casos em que o aluno apresenta desempenho negativo, a propensão ao abandono do curso é maior. Observa-se também que três características pessoais dos alunos surgem como fatores importantes para o desempenho acadêmico, embora a abordagem do trabalho não se aprofunde em cada uma dessas características, a identificação da relevância desses elementos destaca a complexidade e a interconexão de fatores pessoais que podem influenciar a trajetória acadêmica dos estudantes.

7. Considerações Finais

Em relação ao desenvolvimento dos modelos, o segundo, otimizado pelo *Grid Search*, se destacou pela performance mais estável, com menos picos e quedas de desempenho durante sua fase de testes, além de melhor acurácia no conjunto de validação. A matriz de confusão mostrou que ele teve êxito em identificar casos de continuidade, enquanto teve uma quantidade baixa de erros quanto ao caso da predição de alunos evadidos.

Ao analisar as variáveis, ficou claro que o desempenho no primeiro semestre, especialmente a quantidade de disciplinas aprovadas, é decisivo para a continuidade no curso. Observa-se também a necessidade de um aprofundamento na avaliação dos impactos das características pessoais nas decisões tomadas pelos alunos. Nesse sentido, observa-se que a implementação de estratégias específicas, adaptadas às características individuais dos estudantes, podem ser implementadas no primeiro semestre, visando fornecer o suporte necessário para mitigar os desafios iniciais. A atenção nesses aspectos pessoais pode servir de base para políticas institucionais, com o objetivo de melhorar a retenção e o sucesso acadêmico dos estudantes.

Ao projetar o futuro desta pesquisa, há oportunidades significativas para aprimorar ainda mais nossa compreensão da evasão acadêmica. Um ponto para exploração são variáveis adicionais, como participação em atividades extracurriculares ou histórico acadêmico prévio ao ingresso na instituição, podendo enriquecer nossa compreensão do fenômeno, oferecendo perspectivas mais diversificadas.

Aprofundar a análise dos aspectos pessoais identificados também é crucial. Compreender como esses fatores se interconectam e influenciam o desempenho acadêmico pode oferecer clareza sobre os determinantes da evasão.

Referências

ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (Preliminary White Paper)*. 2015. ⟨https://www.tensorflow.org/whitepapers/tensorflow-whitepaper.pdf⟩. Accessed: April 5, 2023.

ALBAN, M.; MAURICIO, D. Neural networks to predict dropout at the universities. *International Journal of Machine Learning and Computing*, International Association of Computer Science and Information Technology Press (IACSIT Press), v. 9, n. 2, p. 185–189, April 2019.

BARANYI, M.; NAGY, M.; MOLONTAY, R. Interpretable deep learning for university dropout prediction. In: *Proceedings of the 21st annual conference on information technology education*. [S.l.: s.n.], 2020. p. 13–19.

BISHOP, C. M. Neural networks and their applications. *Proceedings of the IEEE*, v. 83, n. 2, p. 227–239, 1995.

BRASIL, C. *Procura por profissionais de tecnologia cresce 671% durante a pandemia*. 2021. Disponível em: ⟨https://www.cnnbrasil.com.br/economia/procura-por-profissionais-de-tecnologia-cresce-671-durante-a-pandemia/⟩. Acesso em: 12 Mar. 2023.

CASANOVA, J. R. et al. The dropout of first-year stem students: Is it worth looking beyond academic achievement? *Sustainability*, v. 15, n. 2, 2023. ISSN 2071-1050. Disponível em: ⟨https://www.mdpi.com/2071-1050/15/2/1253⟩.

COLPO, M.; PRIMO, T.; AGUIAR, M. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In: *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2021. p. 873–884. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/sbie/article/view/18114⟩.

CUNHA, J. V. A. d.; NASCIMENTO, E. M.; DURSO, S. d. O. Razões e influências para a evasão universitária: Um estudo com estudantes ingressantes nos cursos de ciências contábeis de instituições públicas federais da região sudeste. v. 9, p. 141–161, Aug. 2016. Citado em: Raphael Magalhães Hoed, Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação, 2016. Disponível em: ⟨https://asaa.anpcont.org.br/index.php/asaa/article/view/260⟩.

HOED, R. M. *Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação*. xvi, 164, [8] p. Mestrado Profissional em Computação Aplicada — Universidade de Brasília, Brasília, 2016. Dissertação.

HURK, A. van den; MEELISSEN, M.; LANGEN, A. van. Interventions in education to prevent stem pipeline leakage. *International Journal of Science Education*, Routledge, v. 41, n. 2, p. 150–164, 2019. Disponível em: ⟨https://doi.org/10.1080/09500693.2018.1540897⟩.

LEE, Y.-M.; FERRARE, J. J. Finding one's place or losing the race? the consequences of stem departure for college dropout and degree completion. *The Review of Higher Education*, Johns Hopkins University Press, v. 43, n. 1, p. 221–261, 2019.

MARQUES, L. et al. Evasão acadêmica e suas causas em cursos de bacharelado em cincia da computação: Um estudo de caso na ufersa. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2020. p. 1042–1051. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/sbie/article/view/12860⟩.

MELO, E. C.; SOUZA, F. S. H. de; SANTOS, E. B. dos. Predição da evasão escolar nos cursos superiores do ifmg–campus bambuí com o apoio de técnicas de aprendizado de máquina. *Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica*, v. 12, n. 1, 2022.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PODER360. *Universidades federais têm evasão de 15% em 2018*. 2018. Disponível em: ⟨https://www.poder360.com.br/governo/universidades-federais-tem-evasao-de-15-em-2018/⟩. Acesso em: 29 Mar. 2023.

RUSSELL, S.; NORVIG, P. *Inteligência Artificial*. 3ª edição. ed. Porto Alegre, Brazil: Bookman, 2013. ISBN 9788577808345.

SEMESP, I. *Evasão – Dados Brasil – 11º Mapa do Ensino Superior*. 2021. Disponível em: ⟨https://www.semesp.org.br/mapa/edicao-11/brasil/evasao/⟩. Acesso em: 12 Mar. 2023.

SEMESP, I. *Evasão – Dados Brasil – 12º Mapa do Ensino Superior*. 2022. Disponível em: ⟨https://www.semesp.org.br/mapa/edicao-12/brasil/evasao/⟩. Acesso em: 29 Mar. 2023.

SIMON, P. *Too Big to Ignore: The Business Case for Big Data*. S.l.: Wiley, 2013. 89 p. ISBN 978-1-118-63817-0.

VENNERS, B. *The Making of Python*. 2007. Artima Developer. Consultado em 22 de março de 2007. Disponível em: ⟨https://www.artima.com/articles/the-making-of-python⟩.

VIANA, F.; SANTANA, A.; RABÊLO, R. Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In: *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2022. p. 908–919. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/sbie/article/view/22469⟩.