UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS JARDINS DE ANITA
CURSO DE TECNOLOGIA EM CIÊNCIA DE DADOS
MINERAÇÃO DE DADOS
DR. JÚLIO CÉSAR SANTOS DOS ANJOS

LARISSA VITÓRIA VASCONCELOS SOUSA - 519221

AVALIAÇÃO PARCIAL 1

ITAPAJÉ, CE
2024

I. INTRODUÇÃO

A evasão escolar, representada pela variável *"dropout"* em uma instituição de ensino superior, é uma preocupação significativa, impactando não apenas a trajetória acadêmica dos estudantes, mas também a eficácia do sistema educacional como um todo. Em resposta a esse desafio, este relatório apresenta um estudo detalhado baseado em *Data Mining* (Mineração de Dados e modelagem de Redes Neurais, com o objetivo de prever a ocorrência de evasão em uma disciplina específica.

Os dados fornecidos para análise incluem uma série de variáveis que abrangem desde informações demográficas até desempenho acadêmico. As variáveis, como idade do aluno, se ele trabalha ou ganha bolsa, reprovações, presença em aulas teóricas e práticas, entre outras, proporcionam uma visão abrangente do perfil dos alunos e de seus comportamentos acadêmicos.

A análise proposta segue as etapas da Mineração de Dados, começando pela exploração e preparação dos dados, passando pela identificação de features relevantes para o problema, até a criação e treinamento de uma Multi-Layer Perceptron (MLP) para a classificação binária do evento de evasão, e, em seguida um modelo de Random Forest também foi treinado para uma comparação dos resultados. O objetivo é descobrir padrões nos dados que possam contribuir para uma melhor compreensão e antecipação do fenômeno de *dropout*.

Ao longo deste relatório, serão apresentadas as principais etapas do processo, os critérios utilizados para a escolha das features mais relevantes, detalhes sobre a construção e treinamento da rede neural, bem como uma avaliação criteriosa do desempenho do modelo desenvolvido.

Espera-se que os resultados deste estudo forneçam insights valiosos para a instituição de ensino, permitindo a implementação de estratégias proativas para reduzir a evasão, proporcionando uma experiência acadêmica mais bem sucedida aos estudantes.

II. METODOLOGIA

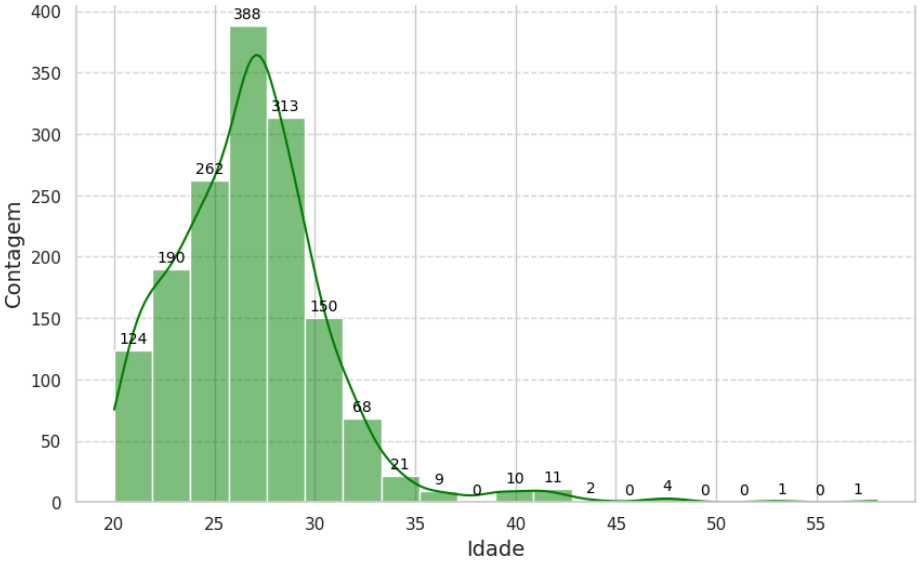
Para a condução deste estudo, foi utilizado o dataset "*DataSet_18_10_2023.csv"* no ambiente *Google Colab*, utilizando Python para a execução dos códigos. Para uma visualização abrangente do código, o mesmo encontra-se disponível no GitHub. A abordagem metodológica empregada neste estudo foi estruturada em cinco etapas, conforme ilustrado na Figura 1.

Figura 1 - Fluxograma

Fonte: Autoria própria, 2024.

A primeira etapa iniciou-se com a limpeza e preparação dos dados, abrangendo a correção de valores ausentes, normalização, codificação de labels, entre outros. Posteriormente, realizou-se uma análise descritiva abrangente do conjunto de dados, visando compreender suas estatísticas fundamentais e identificar padrões. Como complemento, foram criados gráficos estratégicos para visualizar padrões e obter insights valiosos a partir do conjunto de dados. As Figuras 2, 3 e 4 apresentam algumas análises, como as idades dos alunos estão distribuídas, a quantidade de alunos que reprovaram por disciplina, bem como aqueles que possuem bolsa ou não.

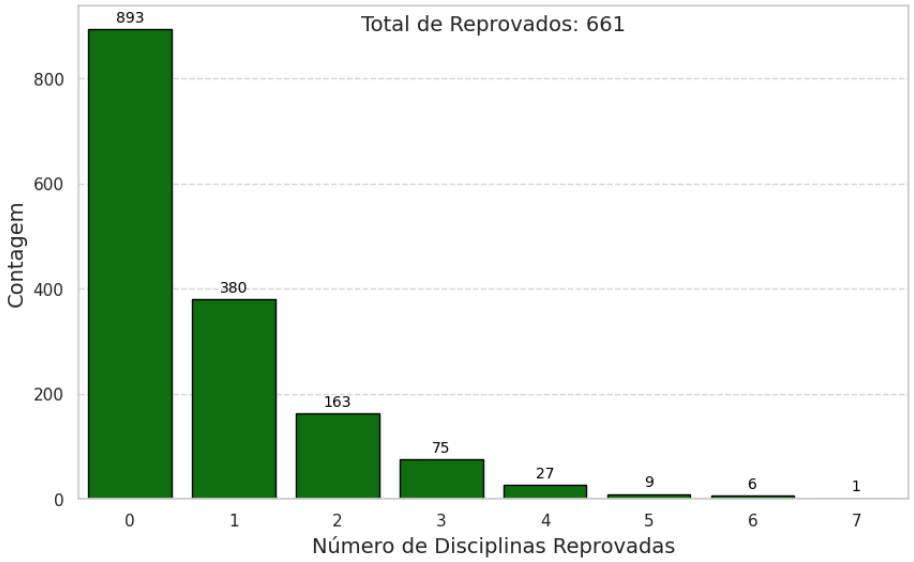
Figura 2 - Distribuição das idades



Fonte: Autoria própria, 2024.

Observa-se claramente que a faixa etária mais representativa dos alunos situa-se entre 25 e 30 anos. Além disso, destaca-se uma minoria de alunos com idades acima de 45 anos. Essa análise evidencia a predominância de estudantes na faixa etária mencionada, sugerindo uma distribuição etária concentrada, enquanto a presença de alunos mais velhos é notadamente menos expressiva.

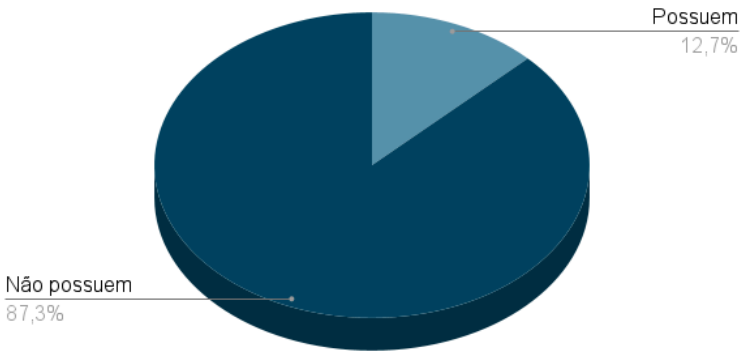
Figura 3 - Reprovações por disciplina



Fonte: Autoria própria, 2024.

É perceptível que a grande maioria dos alunos na base de dados não enfrentou reprovações em nenhuma disciplina. No entanto, é relevante ressaltar que cerca de 41% dos alunos que foram reprovados tiveram, pelo menos, uma reprovação ao longo de seu percurso acadêmico.

Figura 4 - Alunos que possuem ou não bolsa



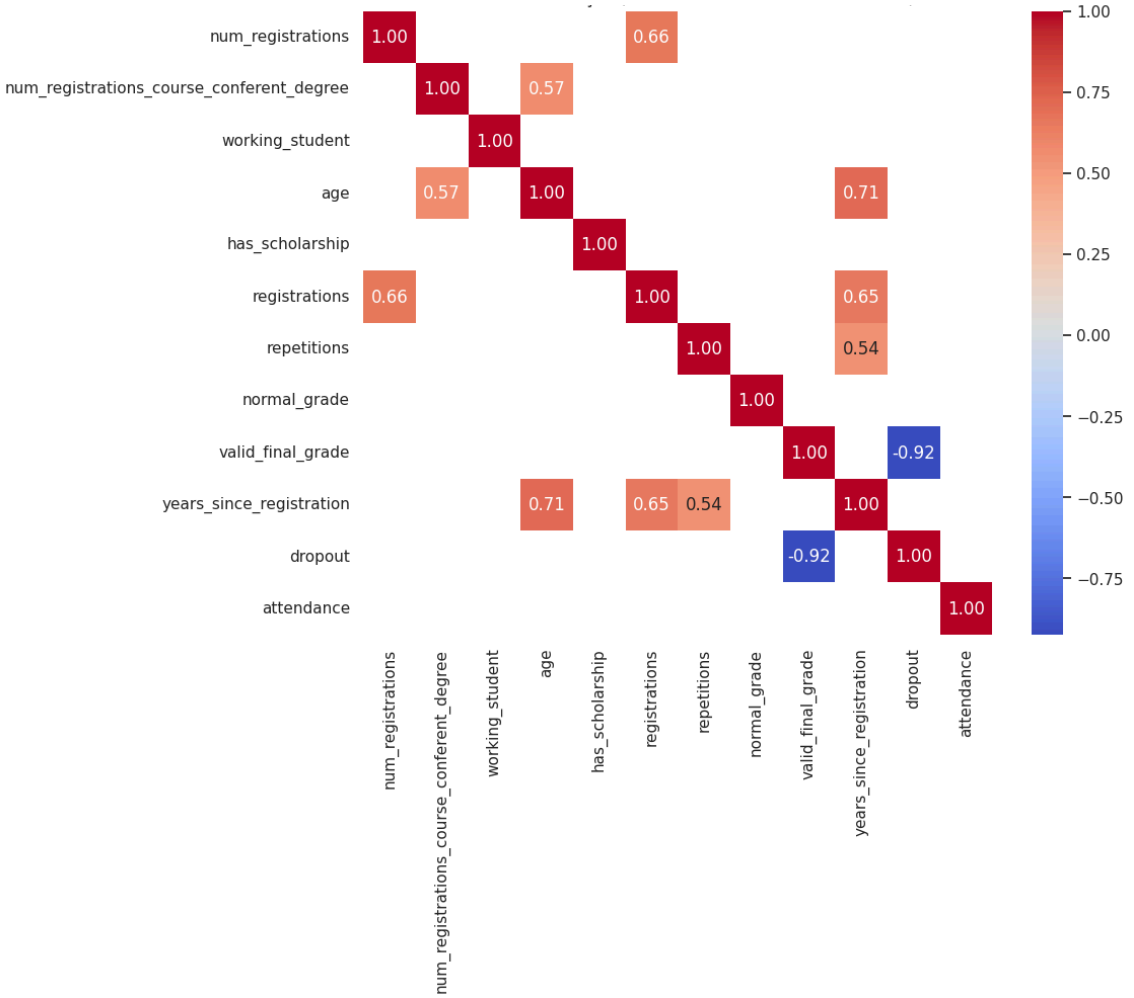
Fonte: Autoria própria, 2024.

Na distribuição percentual entre alunos que possuem ou não bolsa, observa-se que apenas 12,7% dos estudantes recebem esse benefício. Essa constatação sugere que uma parcela reduzida dos alunos desfruta de bolsas, o que possivelmente indica que alguns

estudantes podem enfrentar desafios financeiros, considerando a possível necessidade de auxílio financeiro. Essa observação ressalta a importância de compreender o contexto socioeconômico dos alunos, uma vez que isso pode influenciar significativamente suas experiências acadêmicas.

Na segunda etapa, direcionou-se o foco para a identificação de correlações e a avaliação da importância das variáveis no contexto da predição de evasão. Para essa análise, empregou-se a matriz de correlação como ferramenta principal. A abordagem adotada considerou uma regra específica: quando duas variáveis apresentavam uma correlação positiva moderada, forte ou muito forte (igual ou superior a 0,50), essas variáveis foram selecionadas para análise mais aprofundada. Veja na Figura 5.

Figura 5 - Matriz de correlação



Fonte: Autoria própria, 2024.

Na terceira fase, foi realizada uma etapa crucial de preparação dos dados: a divisão do conjunto em conjuntos de treino e teste, que foi aplicada tanto para Random Forest quanto para Rede Neural Artificial (RNA). Decidiu-se reservar 20% dos dados para teste, enquanto os 80% restantes foram para o treinamento dos modelos. A divisão foi estratificada com base

na variável alvo *'dropout'*, essa estratégia assegura que a proporção de classes (abandono ou não) seja preservada nos conjuntos de treino e teste.

Na quarta etapa, foram implementados dois modelos, sendo a escolha inicial a técnica *Random Forest* para uma análise comparativa. Simultaneamente, construiu-se e treinou-se uma rede neural do tipo MLP, explorando diferentes abordagens para a predição da evasão.

A quinta e última etapa consistiu na avaliação dos modelos. Métricas como precisão, *recall*, acurácia, *F1-score* e matriz de confusão foram empregadas para analisar o desempenho. Uma análise crítica dos resultados conduziu a ajustes nos modelos, visando otimizar sua eficácia e promover a generalização para situações práticas.

III. RESULTADOS

Os resultados obtidos nos modelos *Random Forest* e Rede Neural Artificial (RNA) do tipo MLP são notáveis, evidenciando uma performance excepcional em ambas as técnicas, tal como mostra a Tabela 1.

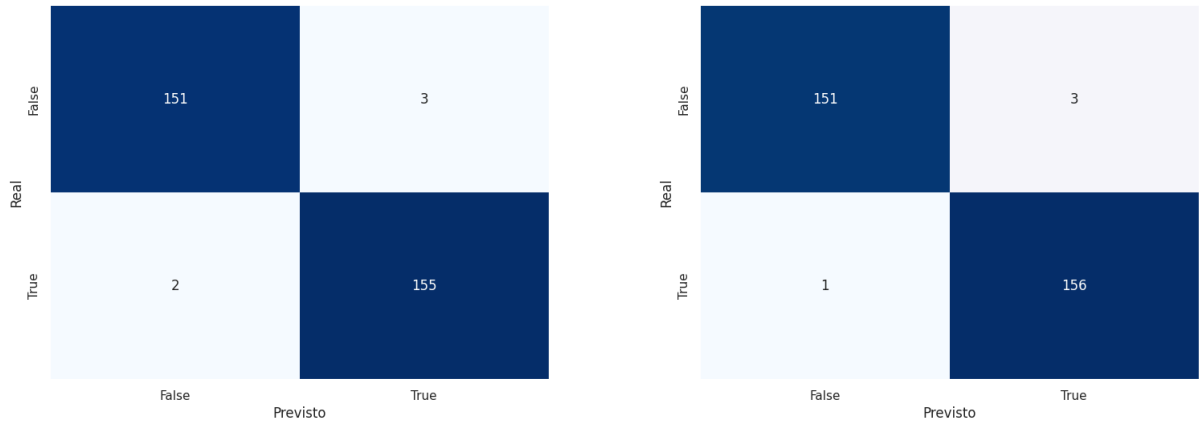
Tabela 1 - Avaliação dos modelos

Métricas	Rede Neural MLP	Random Forest
Precision	98,11%	98,10%
Recall	99,36%	98,73%
F1-Score	98,73%	98,41%
Acurácia	98,71%	98,39%

Fonte: Autoria própria, 2024.

Ambos os modelos alcançaram resultados notáveis, com altas acurácias e métricas de precisão, recall e F1-score muito próximas de 1, porém a RNA do tipo MLP se destacou. Esses resultados indicam que os modelos foram eficientes em prever o abandono de disciplinas, demonstrando um desempenho consistente e confiável. A análise das matrizes de confusão também destaca a capacidade dos modelos em minimizar falsos positivos e falsos negativos, como destaca a Figura 6 a e b.

Figura 6 - Matrizes de confusão



(a) Random Forest

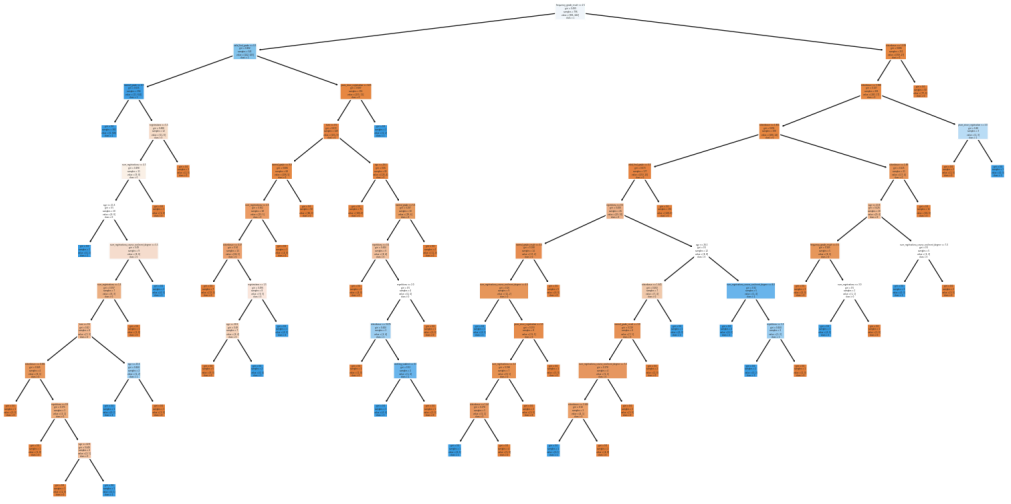
(b) RNA do tipo MLP

Fonte: Autoria própria, 2024.

Na matriz de confusão da Random Forest, a observação revela que a grande maioria dos casos foi corretamente classificada, registrando apenas alguns falsos positivos e falsos negativos, totalizando 5 classificações incorretas. Por outro lado, a matriz de confusão da RNA MLP apresenta um desempenho ligeiramente superior, com apenas 4 classificações erradas. Embora a diferença seja mínima, é notável, destacando a eficácia de ambos os modelos na tarefa de prever o abandono de disciplinas.

Além disso, a figura da árvore da Random Forest (Figura 7) proporciona uma visualização adicional das decisões do modelo, contribuindo para a interpretação do processo de classificação. Essa abordagem híbrida, combinando técnicas de aprendizado de máquina clássico e redes neurais, resultou em modelos robustos e eficazes para lidar com o desafio da predição de evasão.

Figura 7 - Árvore da Random Forest



Fonte: Autoria própria, 2024.