Inteligência Artificial

Trabalho II - Mineração de dados

Meios para remediar a evasão escolar no Brasil

Autores:	*N° USP :*
Vítor Augusto Paiva de Brito	13732303
Marcus Vinicius da Silva	13833150
Rafael Cunha Bejes Learth	13676367
Gabriela Passos de Andrade	12625142
Matheus Pereira Dias	11207752
João Pedro Gomes	13839069
José Matheu Alves da Silva	15505755

Docente:

Solange Oliveira Rezende

2024

Conteúdo

1 Introdução

A evasão escolar é um problema significativo que afeta a educação e o desenvolvimento socioeconômico do Brasil. Portanto, compreender os fatores que influenciam a evasão escolar é crucial para desenvolver políticas eficazes que possam reduzir suas taxas e melhorar a qualidade da educação. Desse modo, a partir das bases de dados do censo escolar de 2021 disponibilizadas pelo **INEP** (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) e dos dados do PIB dos municípios disponibilizados pelo **IBGE** (Instituto Brasileiro de Geografia e Estatística), tem-se como fito abordar a problemática da evasão escolar por intermédio de técnicas de inteligência artificial.

O objetivo deste trabalho é analisar os diversos fatores que têm influência sobre o índice de evasão escolar em todas as escolas do Brasil, utilizando métodos de aprendizado de máquina e mineração de dados. A análise busca identificar fatores que contribuem positivamente, negativamente ou que não possuem influência significativa sobre a evasão escolar. Para alcançar esse objetivo, foram utilizados dois algoritmos de machine learning: Random Forest e Redes Neurais.

Ao utilizar essas técnicas, este estudo pretende fornecer uma visão abrangente sobre os fatores que influenciam a evasão escolar, oferecendo subsídios para a formulação de políticas públicas e intervenções educacionais que possam efetivamente combater esse problema. A análise e os resultados obtidos podem servir como uma base sólida para futuras pesquisas e iniciativas voltadas à melhoria do sistema educacional brasileiro.

2 Desenvolvimento

2.1 Formulação da base de dados

Na abordagem executada para o problema, foi feita uma combinação de bases de dados de origem governamental para fornecer substrato para resolução do problema.

2.1.1 Bases utilizadas

Primeiramente, a fim de se obter uma base detentora de conhecimento acerca dos dados do sistema educacional brasileiro, foi feita a integração de bases do censo escolar de 2021 sobre indicadores educacionais de dois níveis de ensino no país, ensino médio e ensino fundamental,

como as taxas de rendimento escolares e o grau de esforço docente, realizado pelo **INEP**, e da base de dados municipais fornecida pelo **IBGE** do mesmo ano, integrando informações como o PIB e o PIB per capta municipais à base principal. Cada exemplo na base representa uma escola no território nacional, com atributos sobre município, dependência e categoria da entidade.

2.1.2 Imputação de dados

Com o intuito de tratar os dados faltantes das bases integradas foi feita a imputação de dados por meio do algoritmo do *KNN*, com parâmetro K igual a cinco, considerando os objetos que pertenciam à mesma unidade federativa nacional a fim de se obter um resultado condizente para os atributos ausentes nas bases primitivas sem exigir recursos computacionais e tempo excessivamente elevados. Assim, foi feita a média entre os atributos dos cinco objetos mais similares ao exemplar com ausência de certos atributos tratar a problemática.

2.1.3 Desenvolvimento de atributo alvo

Para que a base de dados tenha um atributo alvo que englobe as taxas de evasão em ambos os níveis de ensino analisados, médio e fundamental, foi feita a taxa total de evasão da escola com base na quantidade de turmas existentes na unidade de ensino, na média de alunos por turma e pela taxa de evasão escolar em cada nível de ensino. Portanto, o atributo alvo foi calculado pela equação:

$$\frac{Alunos/Turma_{EF} \cdot Turmas_{EF} \cdot Evasão_{EF} + Alunos/Turma_{EM} \cdot Turmas_{EM} \cdot Evasão_{EM}}{Alunos/Turma_{EF} \cdot Turmas_{EF} + Alunos/Turma_{EM} \cdot Turmas_{EM}}$$

Dessa maneira, as taxas de evasão total de cada instituição de ensino são categorizadas da seguinte forma:

$$\begin{cases} \text{Baixa} & \text{se } Evasão < 5 \\ \text{Moderada} & \text{se } 5 \leq Evasão \leq 10 \\ \text{Alta} & \text{se } Evasão > 10 \end{cases}$$

2.2 Análise exploratória de dados

Com o intuito de se obter uma maior noção sobre o conjunto de dados do problema, faz-se uma análise exploratória da base.

2.2.1 Distribuição da base

A fim de analisar a distribuição entre as classes de evasão escolar, lança-se mão do histograma de categorias do conjunto de dados analisados.

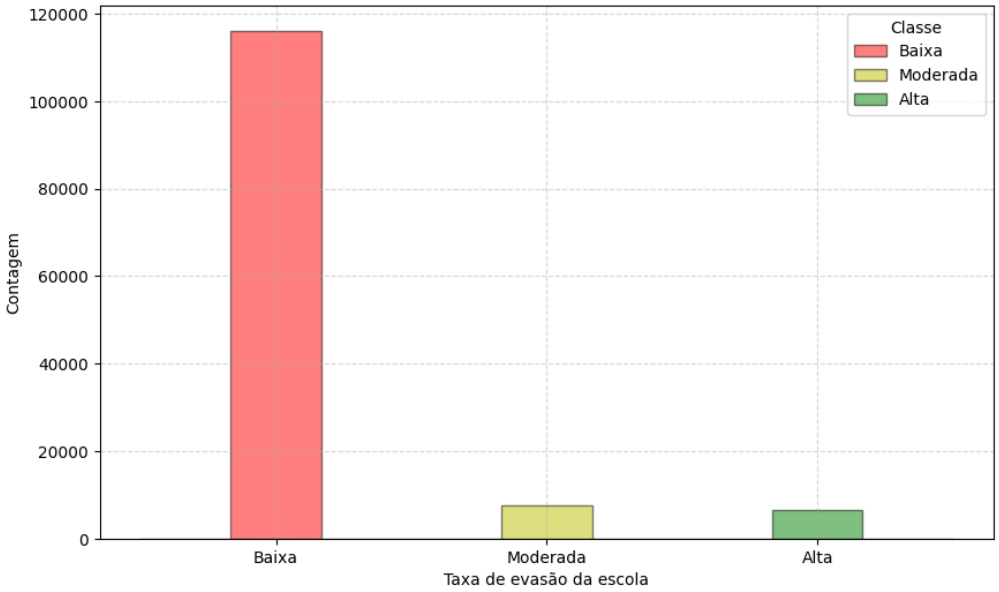

Figura 1: Histograma das categorias de evasão da base.

Vista a distribuição explicitada na **Figura 1**, vê-se que há um desbalanceamento significativo entre as categorias de evasão escolar, sendo a classe majoritária a classe Baixa. Assim, é razoável inferir que haverá um enviesamento no treinamento do modelo, beneficiando a classificação como a categoria com mais representantes, fator que faz necessário métodos de amostragem e de treinamento que contornem o problema.

A distribuição também mascara outra problemática do sistema educacional brasileiro, porém, sendo logístico, de que cerca de 79,5% das escolas que possuem taxa de evasão baixa, não apresentam turmas no ensino médio. Portanto, dado que a taxa de evasão no ensino fundamental é naturalmente mais baixa, a taxa de evasão da maioria das entidades se configura como baixa devido à falta de estrutura, outro problema do sistema educacional nacional.

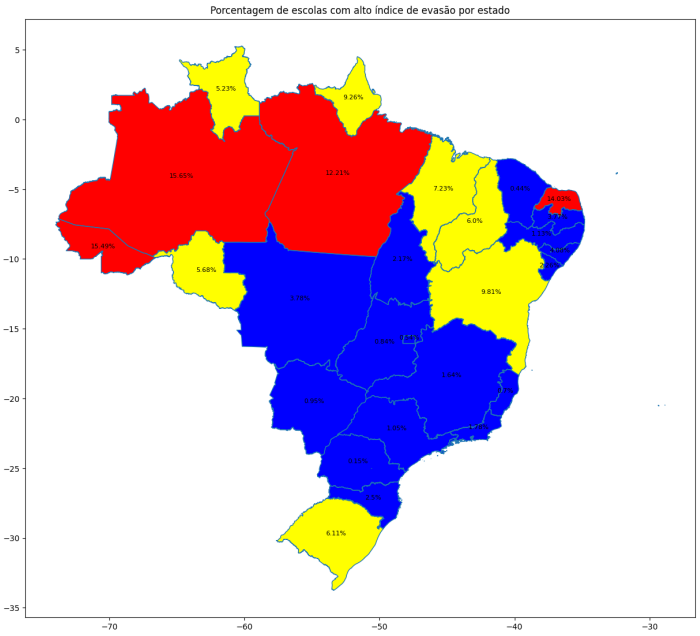

Figura 2: Mapa com taxa de escolas com índice de evasão alto.

2.2.2 Mapa de calor

Com o intuito de se observar a correlação das variáveis utilizadas no problema de aprendizado, faz-se uso do mapa de calor para as correlações com os coeficientes de Pearson e de Kendall.

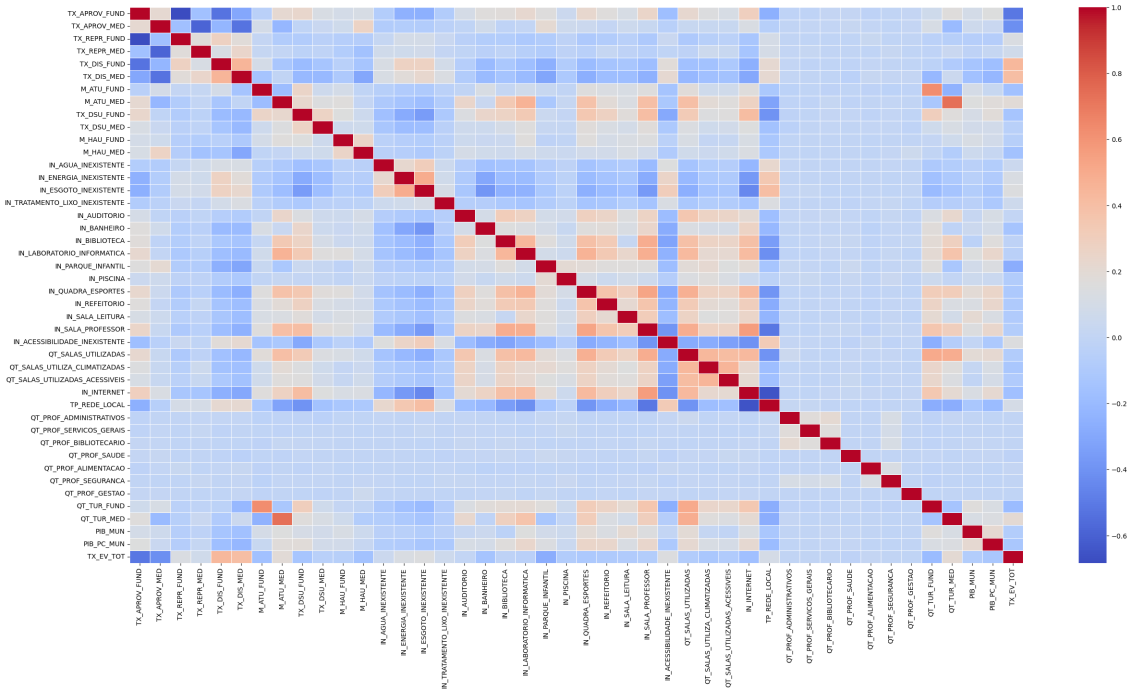

Figura 3: Mapa de correlação entre os atributos da base com coeficiente de Pearson.

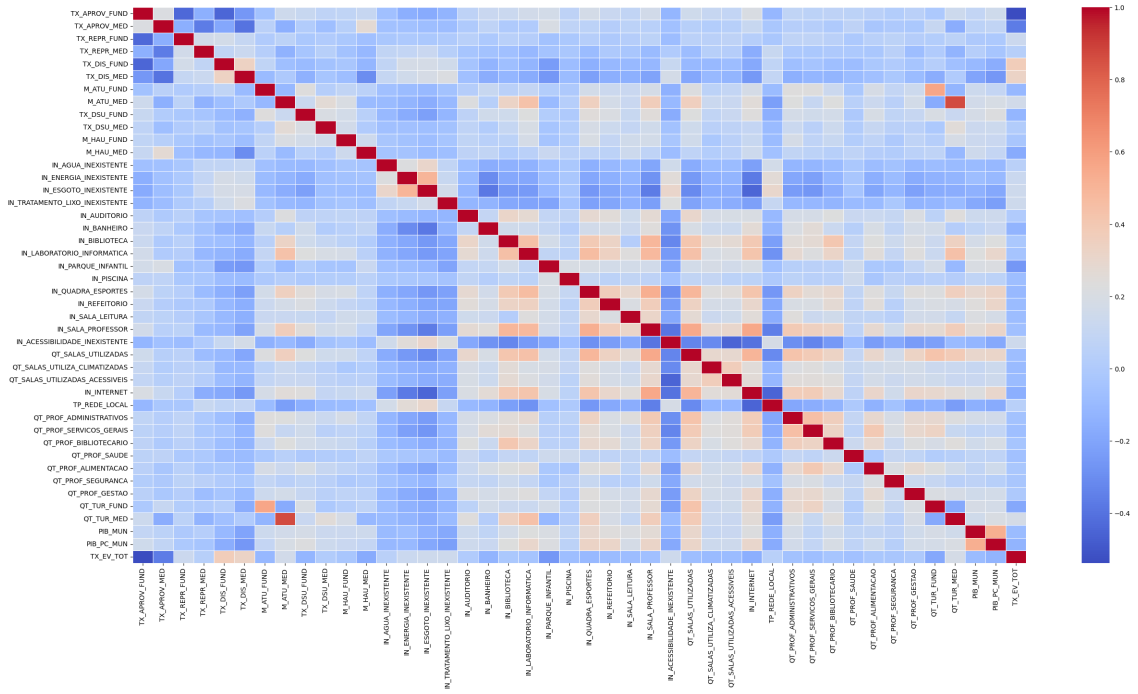

Figura 4: Mapa de correlação entre os atributos da base com coeficiente de Kendall.

A partir dos gráficos de correlação nos mapas de calor, é possível inferir que há poucos casos de variáveis que apresentam correlação significativamente forte, seja de maneira diretamente ou inversamente proporcional. Desse modo, é razoável concluir que os dados são pouco relacionados, fator que corrobora a hipótese de que não há uma redundância considerável entre os atributos e que não há urgência na redução da dimensionalidade da base, com cada característica representando um aspecto distinto de cada instituição de ensino.

2.2.3 Gráficos de dispersão

Com o intuito de se obter uma visualização mais precisa da correlação entre os atributos, faz-se uso dos gráficos de dispersão da base.

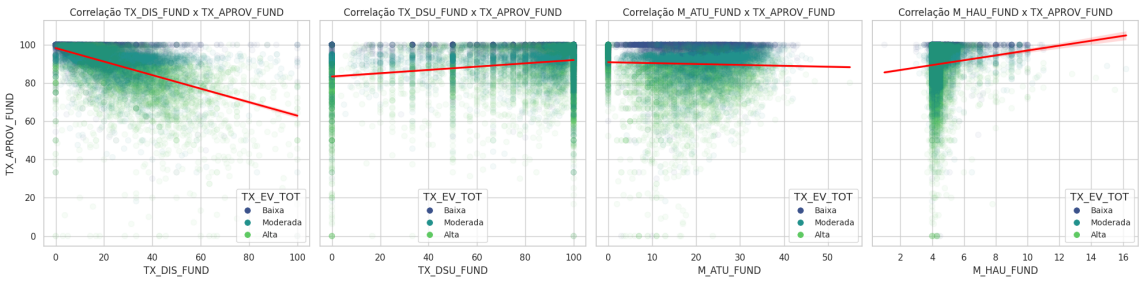

Figura 5: Gráficos de dispersão entre as taxas do ensino fundamental.

Entre as taxas de rendimento do ensino fundamental, é razoável inferir que apenas a taxa de

distorção idade-série apresenta uma correlação significativa com a taxa de aprovação no nível de ensino, fator que implica uma forte relação entre as características.

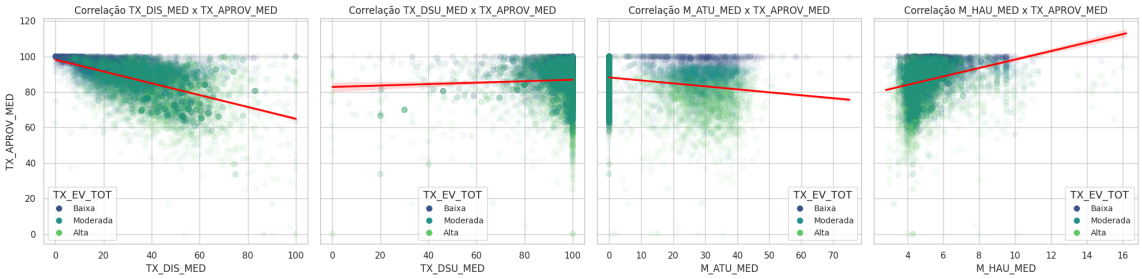

Figura 6: Gráficos de dispersão entre as taxas do ensino fundamental.

Entre as taxas de rendimento do ensino médio, é possível tomar uma conclusão análoga ao ensino fundamental em que a taxa de aprovação tem uma relação mais forte com a taxa de distorção idade-série em detrimento das outras características observadas.

Os gráficos de dispersão, em consonância com o mapa de calor, corrobora a ideia de que as variáveis não apresentam redundância significativa e que a dimensionalidade da base não evidencia problema gritante.

2.2.4 Gráficos de violino

A fim de observar a concentração de valores de alguns atributos entre as categorias da base, são utilizados gráficos violino para que seja possível visualizar as tendências de acúmulo entre os dados.

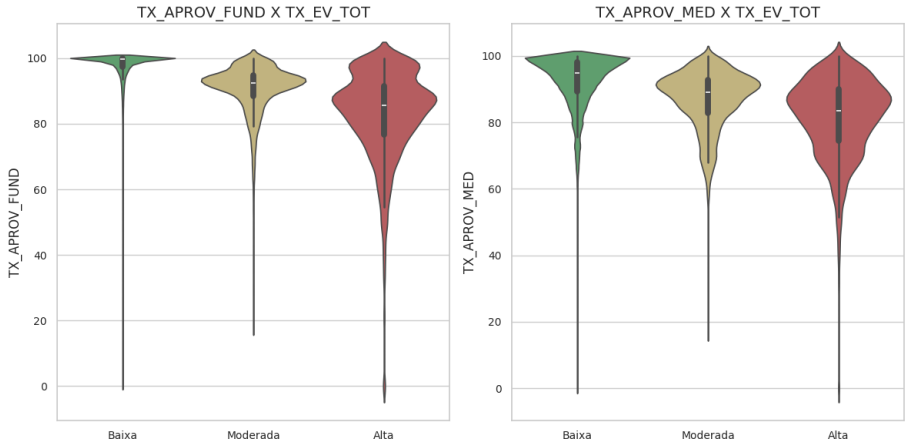

Figura 7: Gráficos de violino da taxa de aprovação para cada categoria.

A partir da **Figura 4**, vê-se que a distribuição da taxa de aprovação dentre as classes segue a intuição de que à medida de que a evasão escolar aumenta a taxa de reprovação é reduzido.

Ademais, analisando os gráficos individualmente, vê-se que a redução taxa de aprovação no ensino fundamental se mostrou ligeiramente mais significativa para a distinção entre as taxas mais altas.

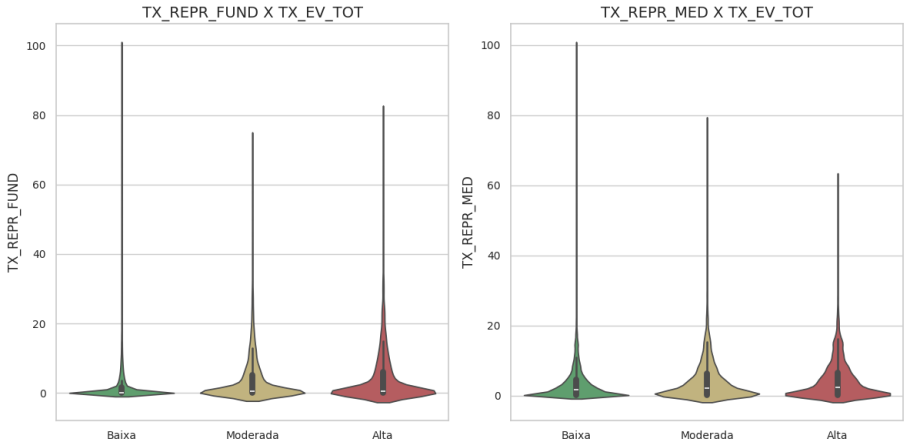

Figura 8: Gráficos de violino da taxa de reprovação para cada categoria.

Tomando como substrato de análise a **Figura 5**, não é tão razoável inferir uma proporcionalidade visual tão clara da reprovação nos níveis de ensino à medida que a taxa de evasão aumenta, porém, ainda é possível observar um ligeiro incremento da reprovação em categorias mais altas.

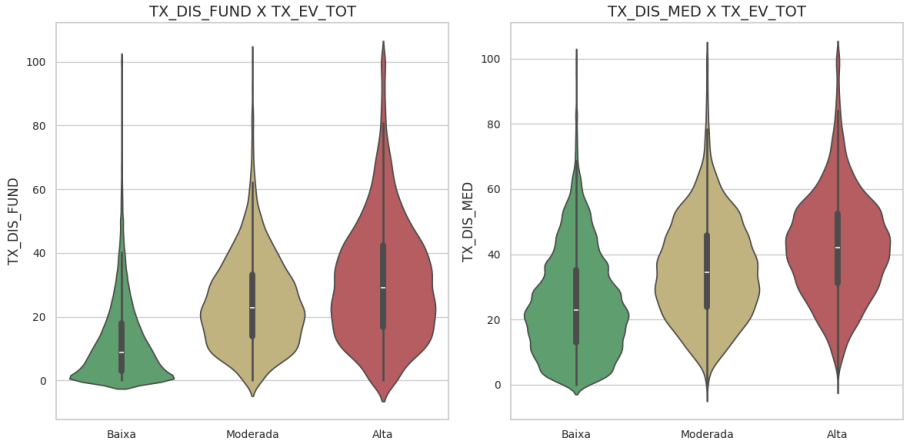

Figura 9: Gráficos de violino da taxa de distorção idade-série para cada categoria.

Por intermédio da **Figura 3**, percebe-se uma notável relação do indicador de distorção idade-série com a taxa de evasão, apresentando um aumento considerável proporcional ao atributo alvo. Além disso, analogamente aos índices de aprovação, infere-se que a evasão se mostrou mais sensível ao incremento da taxa de distorção idade-série do ensino fundamental em detrimento do ensino médio, fator que evidencia uma significativa relevância do nível de ensino no sistema educacional como um todo.

2.3 Modelagem

2.3.1 Amostragem

A fim de reduzir os custos computacionais e diminuir o tempo de execução durante a etapa de treinamento, a abordagem acatada foi utilizar uma amostra da base com cerca de 10% dos objetos totais do conjunto original.

Graças ao alto grau de desbalanceamento da base completa, a etapa de amostragem foi abordada de duas formas diferentes que serão exploradas: uma amostra será artificialmente balanceada, ou seja, os exemplos serão escolhidos dada uma proporção estabelecida de exemplares de cada classe, a proporção escolhida foi uniforme entre as categorias, um terço de cada classe, sendo que os exemplos dentro de cada segmento serão escolhidos aleatoriamente, e a outra amostra será desenvolvida com exemplos aleatórios, sendo assim, a proporção original será, teoricamente, mantida.

2.3.2 Random Forest

Random Forest é um algoritmo de aprendizado de máquina baseado em árvores de decisão que utiliza o conceito de ensemble learning para melhorar a precisão e robustez das previsões. Tal técnica agrega múltiplas árvores de decisão, técnica de *bagging*, para aprimorar a acurácia do modelo final e reduzir o risco de *overfitting*.

Suas vantagens incluem a redução do risco de overfitting, pois combina múltiplas árvores de decisão, o cálculo da importância dos atributos, fornecendo uma medida de quanto cada um contribui para a predição e sua versatilidade, já que pode ser usado tanto para classificação quanto para regressão.

No contexto do problema apresentado, o Random Forest pode ser utilizado tanto na classificação das taxas de evasão escolar quanto na identificação dos atributos mais importantes que influenciam essa taxa ajudando na tomada de decisões para mitigar o problema.

Para o treinamento do modelo com Random Forest, foi utilizado o método de divisão do K-Fold com K igual a dez.

2.3.2.0.1 Amostra balanceada

Primeiramente, treina-se o modelo com a Random Forest na amostra de dados balanceada.

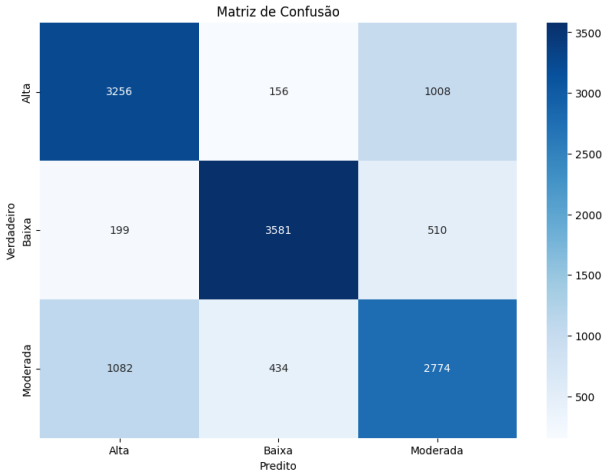

Figura 10: Matriz de confusão para o algoritmo da Random Forest na amostra balanceada.

Classe	**Precisão**	**Recall**
Alta	0.72	0.74
Baixa	0.86	0.83
Moderada	0.65	0.65
Acurácia	0.74	

Tabela 1: Tabela de precisão e recall para as classes alta, moderada e baixa para a Random Forest na amostra balanceada.

2.3.2.1 Amostra aleatória .

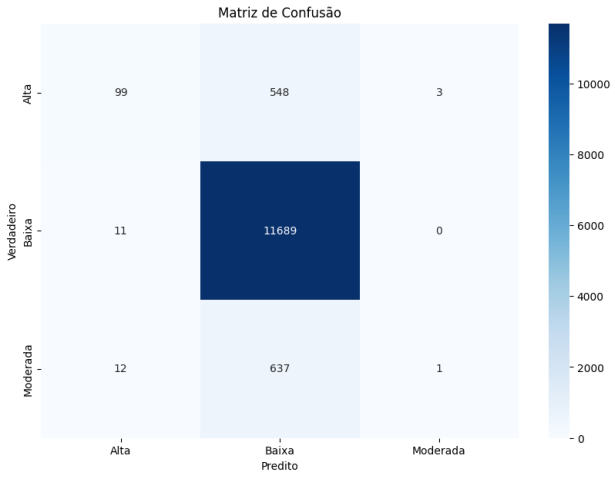

Figura 11: Matriz de confusão para o algoritmo da Random Forest na amostra balanceada.

Classe	Precisão	Recall
Alta	0.81	0.15
Baixa	0.91	1.00
Moderada	0.25	0.00
Acurácia	0.91	

Tabela 2: Tabela de precisão e recall para as classes alta, moderada e baixa para a Random Forest na amostra aleatória.

2.3.3 Redes Neurais

Redes Neurais são modelos inspirados no funcionamento do cérebro humano, capazes de identificar padrões complexos e não-lineares nos dados, tornando-os ideais para prever a evasão escolar com alta precisão. As estruturas são capazes de modelar relações complexas e não lineares nos dados, podem ser escaladas com camadas adicionais e neurônios para aumentar a capacidade de aprendizado e, quando bem treinadas, podem generalizar bem em dados não vistos, assumindo que não há overfitting.

Para o nosso problema, as redes neurais podem ser usadas para classificar as taxas de evasão escolar com alta precisão, especialmente quando os dados são complexos e não lineares, além da detecção de padrões complexos nos dados que podem não ser evidentes com métodos mais simples.

2.3.3.1 Amostra balanceada .

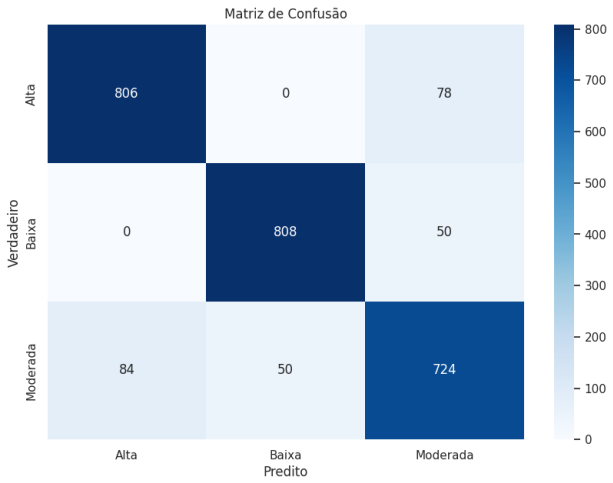

Figura 12: Matriz de confusão para o algoritmo da rede neural na amostra balanceada.

Classe	Precisão	Recall
Alta	0.91	0.91
Baixa	0.94	0.94
Moderada	0.85	0.85
Acurácia	0.90	

Tabela 3: Tabela de precisão e recall para as classes alta, moderada e baixa para a rede neural na amostra balanceada.

2.3.3.2 Amostra aleatória .

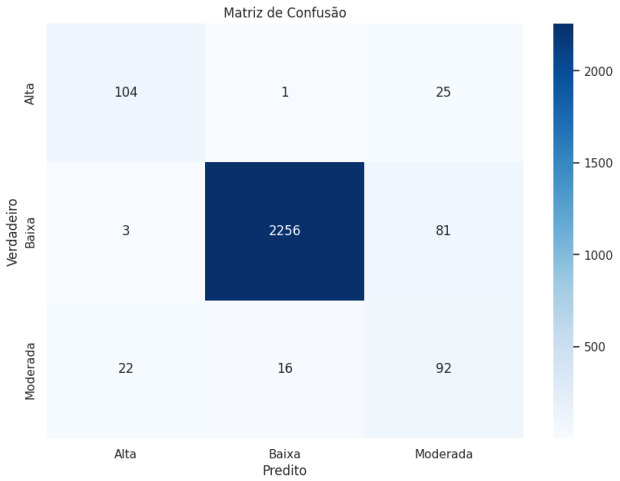

Figura 13: Matriz de confusão para o algoritmo da rede neural na amostra balanceada.

Classe	Precisão	Recall
Alta	0.81	0.80
Baixa	0.99	0.96
Moderada	0.46	0.71
Acurácia	0.94	

Tabela 4: Tabela de precisão e recall para as classes alta, moderada e baixa para a rede neural na amostra aleatória.

3 Análise dos Resultados

3.1 Análise geral

A partir dos resultados da modelagem, é possível inferir que o treinamento sobre uma amostra aleatória, apesar de elevar a acurácia geral do modelo, reduz significantemente a sensibilidade da estrutura preditiva em relação a instituições de ensino que apresentam uma taxa de evasão mais elevado, fator observado pelos baixos valores de recall das categorias de evasão mais altas nos resultados, as quais são as entidades alvo do estudo, portanto, as amostras apresentam menor eficiência para o propósito da tarefa de predição que é identificar os principais atributos que implicam fragilização de uma escola ou de um município no geral.

Além disso, é possível extrair as importâncias dadas para cada atributo por meio do algoritmo da Random Forest.

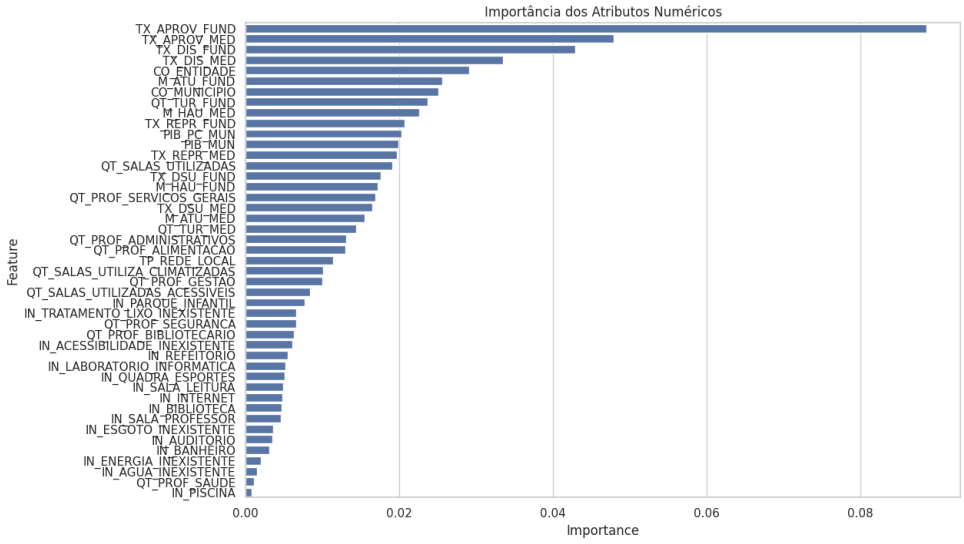

Figura 14: Importância dos atributos atribuídas pela Random Forest na amostragem balanceada.

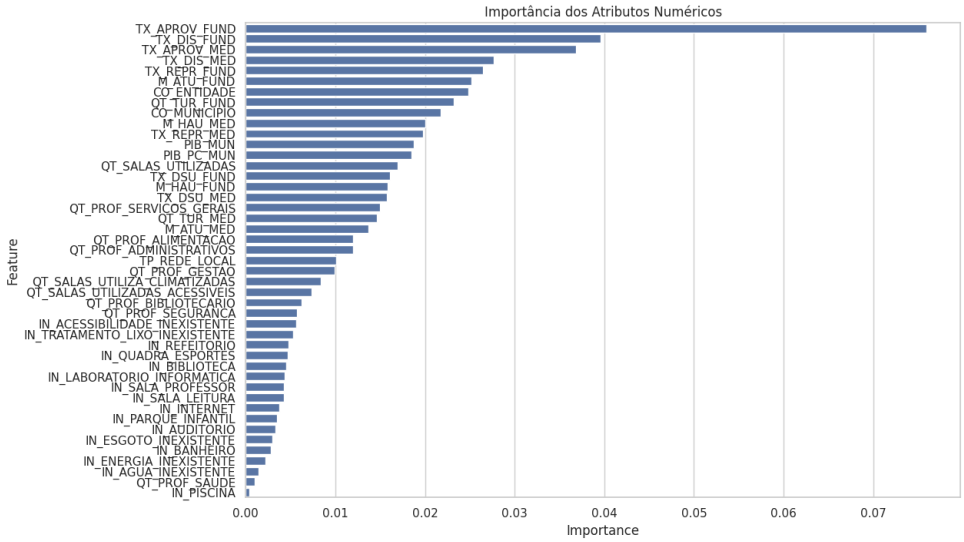

Figura 15: Importância dos atributos atribuídas pela Random Forest na amostragem aleatória.

Assim, a partir das figuras, infere-se que as taxas de rendimento no ensino fundamental apresentaram uma relevância consideravelmente elevada na identificação de entidades escolares vulneráveis. Portanto, a fim de se combater os problemas de evasão escolar no território nacional, uma medida de suma importância é o investimento na educação basilar, ou seja, voltar as atenções para o ensino fundamental pois, a partir dele, um aluno pode adquirir condições de progredir educacionalmente.

3.2 Comparação entre algoritmos

3.2.1 Random Forest

O algoritmo da Random Forest, apesar de fornecer uma interpretação mais simples e intuitiva do resultado por meio da importância de cada atributo, resultou em um modelo com uma acurácia ligeiramente menor que o método de redes neurais. Desse modo, é razoável inferir que o modelo pode oferecer uma abordagem mais interpretável do modelo, porém, pode ser limitado pelo algoritmo e não oferecer um poder de generalização tão elevado.

3.2.2 Redes neurais

A partir das matrizes de confusão e das medidas do modelo, é possível concluir que o algoritmo apresentou um poder de generalização e de extração de conhecimento a partir dos dados maior, possibilitando obter recall e precisão maiores na modelagem.

3.2.3 Comparação de métodos de amostragem

3.2.4 Amostra balanceada

Com substrato lógico nos resultados dos modelos treinados a partir das amostras de dados balanceadas, é possível concluir que a etapa de treinamento a partir dessa amostra, justamente pela ausência de enviesamento por meio de uma classe majoritária no conjunto, proporcionou modelos mais sensíveis para identificar escolas com taxa de evasão alta e moderadas, aumentando o recall dessas categorias de escola, o que é o foco principal da problemática. Portanto, vê-se que, para que seja viável uma distinção mais bem definida de categorias de taxa de evasão escolar, priorizando as unidades educacionais realmente carentes, o treinamento em amostra balanceada se mostrou mais eficiente.

3.2.5 Amostra aleatória

A partir da saída dos modelos treinados em amostras aleatória, é razoável perceber que o enviesamento ocasionado pela classe majoritária, por mais que o modelo apresente uma acurácia geral maior, que tais estruturas não são suficientemente refinados no quesito de identificar escolas que realmente apresentam carência de ensino. Desse modo, conclui-se que as amostras não são tão efetivas para uma extração de conhecimento e reconhecimento de unidades de ensino vulneráveis, não sendo recomendada para ser base de uma pesquisa pública.

3.3 Pesquisas relacionadas

A respeito de pesquisas relacionadas à evasão escolar abordada com mineração de dados, no contexto brasileiro, podem ser encontradas tendo foco de análise a evasão escolar no ensino básico, a qual é estudada por Sales et al., 2019, em que é utilizado o algoritmo WRF ou *Weighted Random Forest* para prever a evasão escolar na cidade de Juiz de Fora em Minas Gerais. Os resultados dessa pesquisa foram obtidos após o algoritmo ser aplicado algumas vezes obtendo uma precisão média de 0.70 e um *recall* médio de 0.97 nos seus atributo mais relevante.

Essa vertente de pesquisa também é desenvolvida em outras instâncias da educação brasileira como é o artigo apresentado por Barbosa et al., 2023 em que são utilizados dados de um dos campi do Instituto Federal de Pernambuco para fazer uma análise preditiva com base em cerca de 13 atributos dos alunos. Nessa pesquisa os melhores resultados obtidos foram utilizando do algoritmo *Random Forest* obtendo 0.83 tanto na precisão quanto no *recall* com uso de atributos

sociodemográficos e acadêmicos.

4 Discussões e decisões

4.1 Definição de intervalos de categorização do atributo alvo

Também foram adotados como índices de evasão escolar os seguintes intervalos de valores: menor que 5% como baixo, entre 5% e 10% como moderado e maior que 10% como alto. Após pesquisar sobre o assunto, analisamos que esses valores como parâmetros de classificação são aceitáveis e condizem com a realidade.

4.2 Imputação de dados

A respeito do tratamento de dados faltantes, foi definido que a imputação por intermédio do algoritmo do KNN, com K igual a 5, seria uma estratégia válida devido à similaridade intraestadual dos municípios, preservando, assim, a fidelidade à realidade das unidades de ensino de uma mesma unidade federativa nacional.

4.3 Proporção utilizada no método de amostragem

Foi decidido que a utilização dos dados balanceados em detrimento dos desbalanceados seria mais benéfica dado que: para os dados balanceados, embora exista algumas predições incorretas quanto às evasões altas e moderadas, quando na realidade são baixas, o modelo prediz corretamente a grande maioria das evasões baixas. O problema surge nos dados desbalanceados, nos quais existem muitas predições baixas para evasões que são moderadas ou altas. Isso é grave, pois é muito importante que as evasões altas, principalmente, sejam preditas corretamente. Além disso, embora os modelos treinados na amostra aleatória apresentem maior acurácia geral, o recall da categoria de evasão alta foi mais elevado em modelos treinados na amostragem balanceada, o que prioriza as entidades alvo do estudo, as instituições de ensino vulneráveis e os fatores causadores de fragilidade.

4.4 Algoritmo mais eficiente

No estudo, foram analisados dois algoritmos para modelagem e, por fim, foi possível concluir que, apesar de uma interpretação e extração de parâmetros mais complexa, a rede neurais se

mostrou mais eficiente em relação ao algoritmo da Random Forest graças ao seu poder de generalização mais refinado e uma maior sensibilidade a escolas mais fragilizadas.

5 Conclusão

Portanto, tomando como base de raciocínio o estudo desenvolvido, é razoável concluir que a evasão escolar é um problema que apresenta múltiplas causas, sendo necessária uma visão holística sobre o problema e ações de várias entidades de múltiplos setores para que seja sanada. Além disso, foi possível concluir que as amostras balanceadas, para o propósito de identificar as fragilidades do sistema educacional e para que os modelos apresentem maior sensibilidade para identificação de vulnerabilidades, são mais indicadas para o treinamento de modelos devido ao menor impacto do enviesamento ocasionado por uma classe majoritária.

Ademais, fica evidente o poder de generalização do algoritmo de redes neurais que, apesar de apresentar interpretação mais complexa, é capaz de gerar regiões delimitadoras melhores e pode resultar em um modelo mais preciso e uma relevância mais elevada no ensino fundamental para a determinação da taxa de evasão escolar de uma instituição de ensino.

Referências

Barbosa, D., Cabral, L., Dwan, F., Feitas, E., & Mello, R. F. (2023). Previsão da Evasão Escolar através da Análise de Dados e Aprendizagem de Máquina: Um estudo de caso. *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, 42–50.

Sales, F., Mendes, Y., Dembogurski, B., Semaan, G., Silva, E., & Ferreira, F. (2019). Evasão no ensino básico da rede pública municipal de juiz de fora: uma análise com mineração de dados. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, *30*(1), 1371.

Censo Escolar - INEP

Evasão Escolar - Toda Matéria

Brasil tem 9 milhões de jovens fora da escola, mostra pesquisa - Correio Braziliense

Abandono e Evasão Escolar no Brasil - IMDS Brasil

Produto Interno Bruto dos Municípios - IBGE