822141543 - Danny Machado
821222653 - Eric Nunes
823212026 - Guilherme Fontes
822150064 - Julia Caroline de Paiva Silva
822139364 - Madox Shibata Oliveira
822147596 - Thais Aires Paiva

GEOGRAFIA DO ABANDONO

Analisando o Papel do Território na Evasão Escolar em São Paulo

São Paulo
2024

SUMÁRIO

1. INTRODUÇÃO

A evasão escolar é um dos grandes desafios enfrentados pelo sistema educacional, impactando diretamente o desenvolvimento social e econômico das regiões afetadas. Em um estado como São Paulo, com marcantes diversidades regionais, compreender os padrões de abandono escolar exige uma análise que integre o papel do território na formação desses indicadores e considere os efeitos de eventos disruptivos, como a pandemia de COVID-19.

Nosso objetivo é explorar a relação entre as mesorregiões do estado e os índices de evasão escolar, classificando como de baixa, média ou alta evasão. Utilizando ferramentas de análise de dados e inteligência artificial, buscamos desvendar como fatores regionais contribuem para o aumento ou redução do abandono escolar, além de avaliar se a pandemia agravou desigualdades já existentes.

2. DADOS

A base de dados utilizada neste projeto abrange informações da Secretaria da Educação do Estado de São Paulo referentes ao ensino médio no período de 2013 a 2023. Os dados incluem diretoria de ensino (NM_DIRETORIA), taxas de aprovação (APR_3) e reprovação (REP_3), e o ano letivo (ANO_LETIVO), permitindo uma compreensão detalhada das dinâmicas que envolvem o abandono escolar (ABA_3). A análise foca em identificar padrões, variações entre regiões e possíveis relações entre os indicadores.

2.1. Tratamento

O tratamento dos dados foi realizado inteiramente por meio de código, assegurando precisão, consistência e replicabilidade em todas as etapas do processo.

Inicialmente, as planilhas dos anos de 2013 e 2014 foram ajustadas para que seu formato ficasse alinhado ao dos demais períodos. Durante essa etapa, os nomes das colunas foram padronizados, e a coluna "ano" foi renomeada para "ano letivo", sendo também convertida para o tipo inteiro (int), garantindo uniformidade na manipulação dos dados.

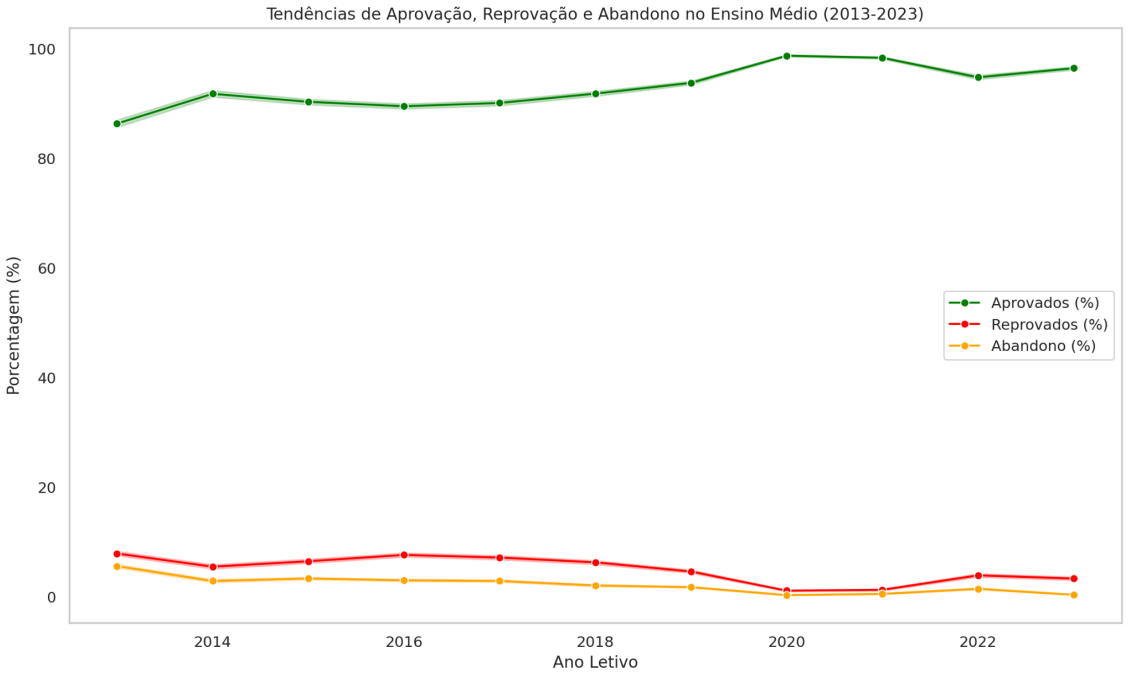
Para facilitar a análise regional e gerenciar o grande volume de diretorias de ensino, foi adicionada uma nova coluna chamada "MESORREGIAO", categorizando as diretorias por mesorregião. Paralelamente, as diretorias de ensino foram organizadas em ordem alfabética, otimizando a visualização e a navegação pelos dados.

Por fim, todas as planilhas tratadas foram consolidadas em um único arquivo integrado, resultando em uma base de dados unificada e estruturada, pronta para análises subsequentes.

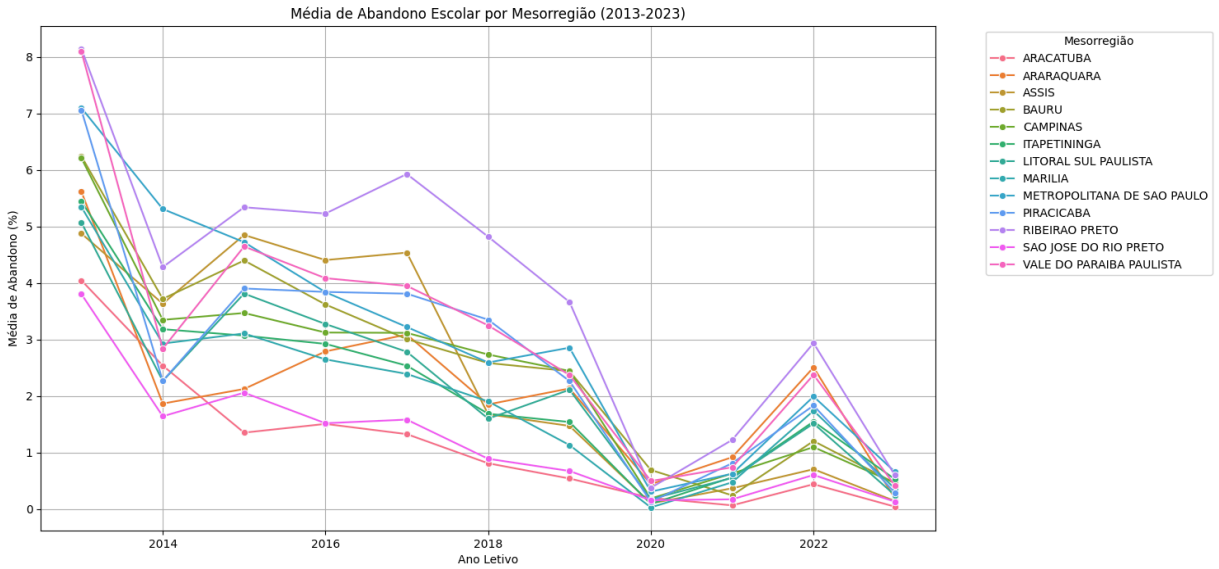
2.2. Análise

A análise dos dados revelou tendências importantes nas taxas de aprovação, reprovação e abandono escolar. A taxa média de aprovação foi de 92,88%, indicando um desempenho positivo na maioria das escolas, com a maior parte das diretorias registrando valores acima de 88%. No entanto, a reprovação apresentou maior variabilidade, com uma média de 4,95% e alguns picos significativos, chegando a 38,81% em determinados casos.

O abandono escolar, embora tenha mostrado uma média relativamente baixa de 2,13%, também apresentou casos extremos, com taxas que chegaram a 23%. Ao analisar as tendências ao longo dos anos, observou-se que as taxas de aprovação se mantiveram altas e consistentes, enquanto a reprovação e o abandono exibiram oscilações pontuais, sugerindo possíveis influências de eventos externos ou mudanças nas políticas educacionais.

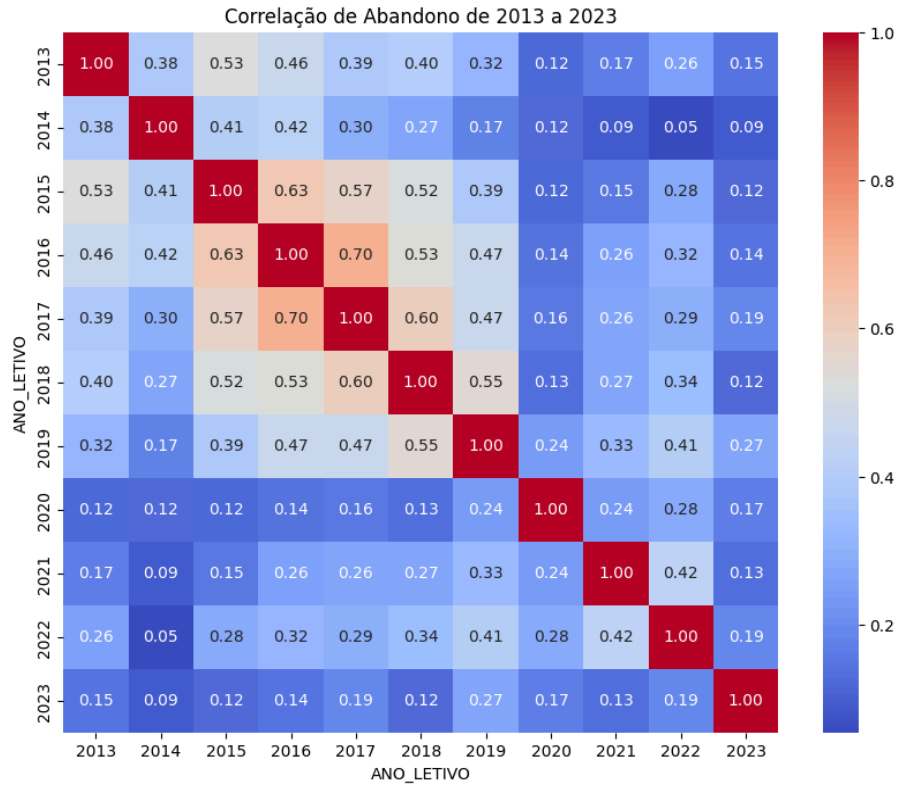


A análise regional, realizada por mesorregião, evidenciou diferenças marcantes. Algumas apresentaram uma dispersão maior nas taxas de reprovação e abandono, indicando disparidades regionais que podem estar ligadas a fatores socioeconômicos ou à infraestrutura educacional.



2.3. Correlação

Este mapa de calor apresenta a matriz de correlação entre os índices de abandono escolar ao longo dos anos de 2013 a 2023. Cada célula da matriz representa o grau de correlação entre dois anos, com valores variando de -1 (correlação negativa forte) a 1 (correlação positiva forte).



É possível observar que anos próximos apresentam correlações mais altas, refletindo continuidade nas tendências educacionais. Por outro lado, anos mais afastados, como 2013 e 2023, apresentam correlações mais baixas, refletindo mudanças significativas ao longo do tempo. Correlações moderadas em anos como 2016 e 2017 sugerem períodos de transição nos padrões de abandono, o que pode indicar momentos de adaptação ou resposta a intervenções.

3. PREVISÕES

3.1. Verificação do melhor modelo

Foi desenvolvido uma função nomeada *evaluate_model* para identificar o melhor modelo preditivo para o conjunto de dados. Essa função avalia o desempenho de modelos de regressão e classificação, utilizando os seguintes parâmetros:

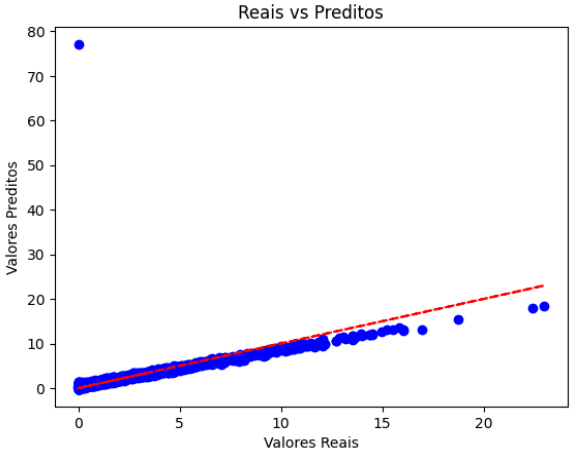
- **X**: DataFrame com as variáveis independentes (*features*).
- **y**: Vetor contendo a variável dependente (*target*).
- **model**: Modelo de machine learning a ser avaliado.
- **model_type**: Define o tipo de problema, podendo ser *regression* (regressão) ou *classification* (classificação).

O processo começa com a normalização dos dados, utilizando o método *StandardScaler* para escalonar variáveis numéricas principais. Em seguida, os dados são divididos em conjuntos de Treinamento (70%) e Teste (30%) e é realizado o treinamento do modelo. Conforme o tipo de modelo especificado, a função realiza as seguintes análises:

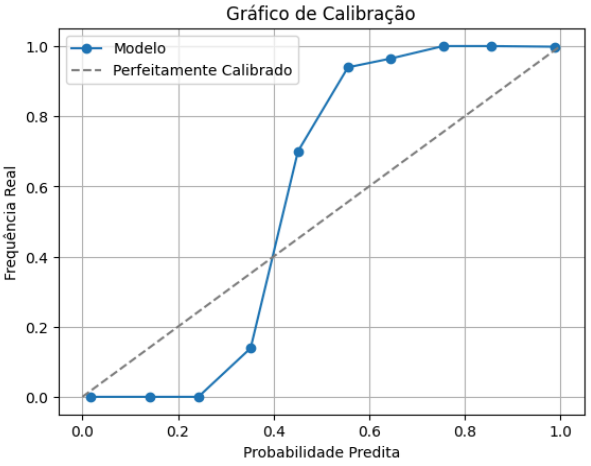
- **Modelos de regressão:** São calculados o R^2, o *MAE* (Erro Absoluto Médio) e o *MSE* (Erro Quadrático Médio). Além disso, é gerado um gráfico de dispersão comparando os valores reais (*y_test*) com os valores preditos pelo modelo (*y_pred*).
- **Modelos de classificação:** São apresentados o *classification_report*, *recall* e *F1-score* por classe. Também são calculados o *AUC-ROC* e gerada a *curva ROC*, além de um gráfico de calibração, que compara os valores reais com os preditos.

A seguir, são apresentados os gráficos gerados pela função *evaluate_model*:

Regressão Linear



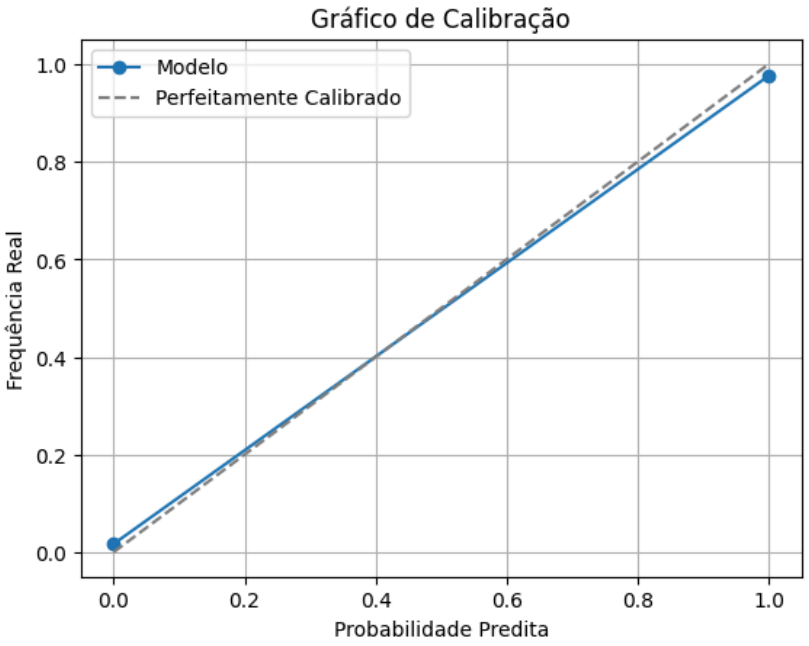
Regressão Logística



Como mostrado, o modelo de regressão linear não foi adequado, pois assume uma relação linear entre as variáveis, o que não reflete a complexidade dos dados, sugerindo que o

modelo não consegue capturar adequadamente padrões mais complexos ou variações nos dados. Quanto à regressão logística, o gráfico de calibração revela que o modelo não está bem calibrado, ou seja, as probabilidades previstas não correspondem às frequências reais. Isso indica que a confiança do modelo nas previsões pode ser imprecisa.

Árvore de Decisão



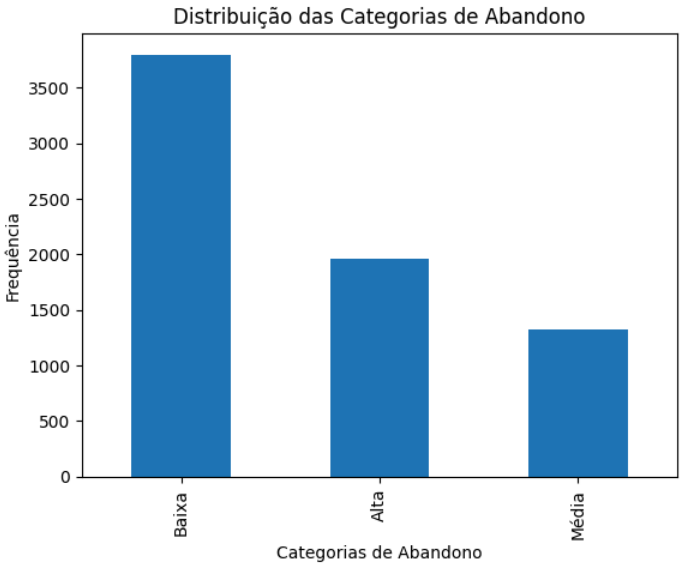
A análise do gráfico de calibração da árvore de decisão mostra que o modelo está bem calibrado. Isso indica que as probabilidades previstas pelo modelo correspondem com precisão às frequências reais observadas nos dados. Em comparação com os outros modelos analisados, a árvore de decisão apresenta maior confiabilidade nas previsões, sugerindo que é capaz de capturar a complexidade dos padrões nos dados de forma mais eficaz.

3.2. Balanceamento

Nesta etapa, foi realizada por meio de código a definição de intervalos e rótulos para categorizar os valores da coluna ABA_3. Esses intervalos foram criados para classificar os dados em três categorias interpretativas: "Baixa", "Média" e "Alta". O intervalo correspondente a "Baixa" abrange valores entre zero e um, enquanto "Média" inclui valores entre um e três. Valores acima de três foram categorizados como "Alta".

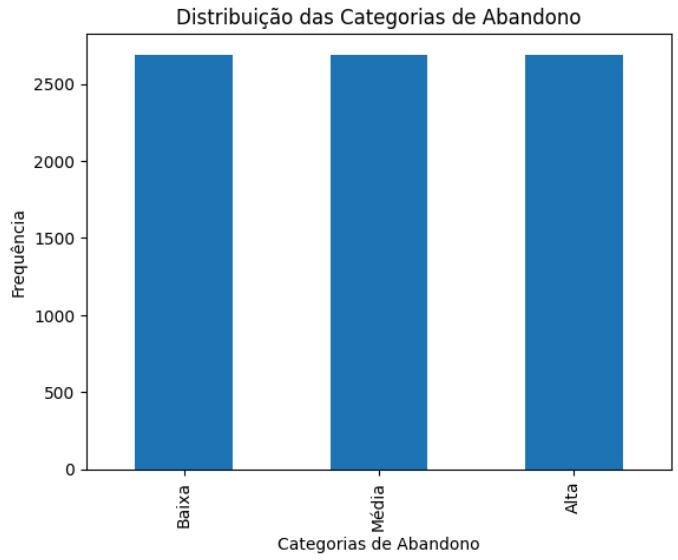
Após a criação dos intervalos, uma nova coluna foi adicionada à base de dados, atribuindo cada valor de ABA_3 à sua respectiva categoria, com base nos limites definidos. Para evitar inconsistências, os valores que não se enquadrassem nos intervalos estabelecidos ou fossem nulos foram automaticamente preenchidos com a categoria "Média", utilizada como padrão.

Com os dados categorizados, foi realizada a análise da distribuição das categorias, apresentando o número de ocorrências de cada uma (Baixa, Média e Alta).



Antes do balanceamento

Ao perceber que as classes estavam desbalanceadas, com 3.796 registros na categoria "Baixa", 1.961 na categoria "Alta" e apenas 1.328 na categoria "Média", foi aplicada a técnica SMOTE. Essa abordagem foi utilizada para equilibrar a distribuição das classes, criando exemplos sintéticos para as categorias minoritárias. O objetivo foi reduzir o impacto do desbalanceamento nos modelos de aprendizado de máquina, garantindo que todas as categorias fossem devidamente representadas durante o treinamento.



Após o balanceamento

3.3. Criação do modelo

O desenvolvimento do modelo de previsão de abandono escolar foi realizado através da função *prever_abandono*, que permite ao usuário fornecer dados como o ano letivo e a mesorregião de interesse. Inicialmente, o nome da mesorregião é codificado para valores numéricos utilizando um *label_encoder*. Caso a mesorregião fornecida não exista no conjunto de dados, uma mensagem de erro informativa é retornada. Em seguida, os dados são filtrados com base no ano e na mesorregião codificada. Se não houver registros correspondentes, uma mensagem é exibida, indicando a falta de dados para os critérios fornecidos. Caso os dados estejam disponíveis, os valores de aprovação (*APR_3*) e reprovação (*REP_3*) são extraídos e utilizados, juntamente com os dados codificados, como entrada para o modelo, que então realiza a previsão da categoria de abandono escolar (Baixa, Média ou Alta).

```python
# Função para prever a categoria (Baixa, Média ou Alta)
def prever_abandono(ano, mesorregiao):
    try:
        # Codificar a mesorregião fornecida
        mesorregiao_encoded = label_encoder.transform([mesorregiao])[0]
    except ValueError:
        return f"Mesorregião '{mesorregiao}' não encontrada no dataset!"

    filtro = (data['ANO_LETIVO'] == ano) & (data['MESORREGIAO_ENCODED'] == mesorregiao_encoded)
    dados_filtrados = data[filtro]

    # Verificar se existem dados correspondentes
    if dados_filtrados.empty:
        return f"Nenhum dado encontrado para o ano {ano} e a mesorregião '{mesorregiao}'!"

    # Obter os valores de APR_3 e REP_3 do banco de dados
    apr_3 = dados_filtrados['APR_3'].iloc[0]
    rep_3 = dados_filtrados['REP_3'].iloc[0]

    # Criar os dados de entrada para a previsão
    input_data = pd.DataFrame({
        'MESORREGIAO_ENCODED': [mesorregiao_encoded],
        'APR_3': [apr_3],
        'REP_3': [rep_3],
        'ANO_LETIVO': [ano]
    })

    # Fazer a previsão
    prediction = model.predict(input_data)[0]
    return f"A categoria de abandono escolar prevista é: {prediction}"
```

Para avaliar o desempenho do modelo, foram utilizadas métricas amplamente reconhecidas em problemas de classificação, como acurácia, precisão, revocação e F1-Score. A acurácia mede a proporção de previsões corretas realizadas pelo modelo. A precisão avalia a proporção de verdadeiros positivos entre os casos classificados como positivos, enquanto a revocação reflete a capacidade do modelo de identificar corretamente todos os casos positivos. O F1-Score combina essas duas métricas em uma única medida, representando sua média harmônica e equilibrando o desempenho em cenários onde classes desbalanceadas são

comuns. Todas essas métricas foram organizadas em uma função chamada *obter_metricas*, que formata os resultados de maneira clara para consulta.

```
# Calcular a acurácia
acuracia = accuracy_score(y_test, y_pred)

# Calcular precision, recall e f1-score
precision = precision_score(y_test, y_pred, average='weighted', zero_division=0)
recall = recall_score(y_test, y_pred, average='weighted', zero_division=0)
f1 = f1_score(y_test, y_pred, average='weighted', zero_division=0)

# Função para exibir as métricas do modelo
def obter_metricas():
    return (f"Acurácia: {acuracia:.2f}\n"
            f"Precisão (Precision): {precision:.2f}\n"
            f"Revocação (Recall): {recall:.2f}\n"
            f"F1-Score: {f1:.2f}")
```

A interface foi desenvolvida utilizando o *Gradio*. O usuário pode inserir o ano letivo e selecionar a mesorregião de interesse a partir de uma lista suspensa com opções disponíveis no conjunto de dados. Ao clicar no botão "Prever", o sistema executa a previsão e exibe o resultado de forma imediata. Além disso, outro botão, "Exibir Métricas", permite que o usuário visualize os indicadores de desempenho do modelo em uma caixa de texto.

```
# Criar a interface com Gradio
with gr.Blocks() as demo:
    gr.Markdown("# Previsão de Evasão Escolar")
    gr.Markdown("Insira o ano letivo e a mesorregião para prever a
    categoria de abandono escolar (Baixa, Média ou Alta).")

    ano = gr.Number(label="Ano letivo (2013-2024)", value=2023)
    mesorregiao = gr.Dropdown(
        label="Mesorregião",
        choices=data['MESORREGIAO'].unique().tolist()
    )
    output = gr.Textbox(label="Resultado da Previsão")
    metricas_output = gr.Textbox(label="Métricas do Modelo")

    btn_prever = gr.Button("Prever")
    btn_prever.click(prever_abandono, inputs=[ano, mesorregiao], outputs=output)

    btn_metricas = gr.Button("Exibir Métricas")
    btn_metricas.click(obter_metricas, inputs=[], outputs=metricas_output)

demo.launch()
```

Por fim, a aplicação é lançada com o comando *demo.launch()*, disponibilizando a solução em um navegador.

4. RELATÓRIO FINAL

A análise realizada confirmou que a evasão escolar no estado de São Paulo apresenta uma forte relação com fatores regionais, como condições socioeconômicas, infraestrutura educacional e níveis de urbanização. Durante a pandemia de COVID-19, os índices de abandono escolar apresentaram uma redução aparente em comparação a períodos anteriores. Esse fenômeno pode ser atribuído a mudanças na dinâmica educacional, como a suspensão temporária da exigência de presença física e a flexibilização de critérios de acompanhamento escolar. Contudo, essa queda nos números não necessariamente reflete uma melhora estrutural, mas possivelmente uma subnotificação ou adiamento de abandonos formais, sugerindo que os impactos reais da pandemia sobre a evasão escolar podem se manifestar no médio e longo prazo.

Após testar diversos modelos de regressão e classificação, optamos por utilizar o modelo de árvore de decisão, que apresentou as melhores métricas de desempenho. Nosso modelo conseguiu prever os índices de abandono escolar de forma muito eficaz, destacando os fatores regionais mais relevantes e proporcionando uma análise precisa das condições que contribuem para o fenômeno. Essas previsões podem servir como base para futuras pesquisas e decisões estratégicas no combate à evasão escolar.