Análise de Técnicas de Classificação em Dados Desbalanceados: Um Estudo de Caso com Dados de Qualidade do Ar

Andrei Inoue Hirata

9 de novembro de 2025

Resumo

Este relatório aborda o desafio da classificação de dados desbalanceados, um problema comum em mineração de dados onde as classes de interesse (como "alertas") são significativamente mais raras que as classes normais. Utilizando um conjunto de dados de qualidade do ar (MP10), este trabalho compara a performance de um classificador Random Forest em duas situações: um modelo base (baseline) treinado nos dados originais e um modelo treinado com dados balanceados pela técnica SMOTE (Synthetic Minority Over-sampling Technique). Os resultados demonstram que, embora a acurácia geral seja similar, o modelo treinado com SMOTE apresenta um aumento drástico na capacidade de identificar corretamente a classe minoritária "Alerta", provando a eficácia da técnica para este tipo de problema.

1 Introdução

A poluição atmosférica é um dos principais desafios em centros urbanos, impactando diretamente a saúde pública. A capacidade de prever a qualidade do ar é fundamental para a tomada de decisão. Modelos de classificação de aprendizado de máquina são ferramentas poderosas para esta tarefa, mas enfrentam um desafio estatístico significativo: os dados são inerentemente desbalanceados. Dias com qualidade do ar "Boa"são muito comuns, enquanto dias de "Alerta"(poluição alta) são raros.

Problema Modelos treinados em dados desbalanceados tendem a ignorar a classe minoritária, pois aprendem que podem obter uma alta acurácia simplesmente prevendo a classe majoritária. Isso resulta em um modelo que é inútil para o seu propósito principal: prever os dias perigosos.

Objetivo O objetivo deste trabalho é aplicar e comparar duas técnicas de classificação ao tema "Dados Desbalanceados": (1) Um modelo baseline (Random Forest) e (2) um modelo com rebalanceamento (Random Forest + SMOTE). A comparação será feita utilizando métricas adequadas, como Precision, Recall e F1-Score, com foco na classe minoritária "Alerta".

2 Trabalhos Relacionados

A literatura acadêmica aborda extensivamente o desafio de dados desbalanceados. Della-Justina (2023) avalia diversas técnicas de classificação para este problema, destacando a ineficácia da acurácia como métrica principal. Especificamente na análise de qualidade do ar, Vianna Jr. et al. (2025) aplicam machine learning a dados da CETESB, demonstrando a viabilidade da classificação. Este trabalho se insere nessa intersecção, aplicando uma técnica de rebalanceamento (SMOTE), conforme popularizado por Lemaître et al. (2017), a um conjunto de dados prático de poluição.

3 Fundamentos

3.1 Random Forest

É um classificador do tipo *ensemble* que opera construindo múltiplas árvores de decisão durante o treinamento. Para classificação, o resultado final é a classe que obteve a maioria dos "votos"das árvores individuais. É conhecido por sua robustez e alta performance.

3.2 Dados Desbalanceados e o Paradoxo da Acurácia

Um dataset é desbalanceado quando a distribuição das classes é desigual (ex: 95% "Boa", 5% "Alerta"). O "Paradoxo da Acurácia"ocorre quando um modelo trivial (que prevê "Boa"100% das vezes) atinge 95% de acurácia, embora seja completamente incapaz de identificar a classe de interesse.

3.3 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE é uma das técnicas mais populares para lidar com o desbalanceamento. Em vez de simplesmente duplicar os dados raros (oversampling), o SMOTE "sintetiza"novos exemplos. Ele seleciona um exemplo da classe minoritária, encontra seus vizinhos mais próximos, e cria um novo ponto de dados sintético *entre* o ponto original e seus vizinhos.

4 Metodologia

4.1 Dataset

Foi utilizado um conjunto de dados de qualidade do ar da estação de Bauru, SP, referente ao ano de 2024. O dataset (`qualar_2024_bauru_mp10.csv`) contém registros horários do poluente MP10.

4.2 Engenharia de Features e Pré-processamento

O script Python (`analise_dados_desbalanceados.py`) realiza os seguintes passos:

1. **Carregamento e Limpeza:** Os dados são carregados do CSV, pulando o cabeçalho. As colunas são nomeadas (`Data`, `Hora`, `value`). Linhas com dados ausentes e formatos incorretos (ex: "24:00") são corrigidos e limpos.
2. **Definição do Alvo (y):** O valor de MP10 (`value`) é convertido em uma classe categórica (`classe_qualidade`) com base nos padrões da CETESB: "Boa"(0-50 µg/m³), "Moderada"(51-100) e "Alerta"(>100).
3. **Definição das Features (X):** O `timestamp` é convertido em features contextuais que o modelo usará para prever: `hora_do_dia` (0-23), `dia_da_semana` (0-6) e `mes` (1-12).

4.3 Pipeline Experimental

Os dados foram divididos em 70% para treino e 30% para teste.

- **Técnica 1 (Baseline):** Um StandardScaler e um RandomForestClassifier são treinados diretamente nos dados de treino desbalanceados (`X_train_scaled`, `y_train`).
- **Técnica 2 (SMOTE):** A técnica SMOTE é aplicada *apenas* aos dados de treino (`X_train_scaled`, `y_train`) para criar um novo conjunto balanceado (`X_resampled`, `y_resampled`). Um segundo RandomForestClassifier é treinado nestes novos dados.

4.4 Métricas de Avaliação

Ambos os modelos são avaliados no *mesmo* conjunto de teste (`X_test_scaled`), que é naturalmente desbalanceado (refletindo a realidade). Comparamos o `classification_report` (Precision, Recall, F1-Score) de ambos, focando no desempenho da classe "Alerta".

5 Experimentos e Resultados

5.1 Distribuição Inicial das Classes

Como esperado, o conjunto de dados (após a limpeza) é altamente desbalanceado. A Figura 1 ilustra a distribuição, onde a classe "Boa"domina o dataset.

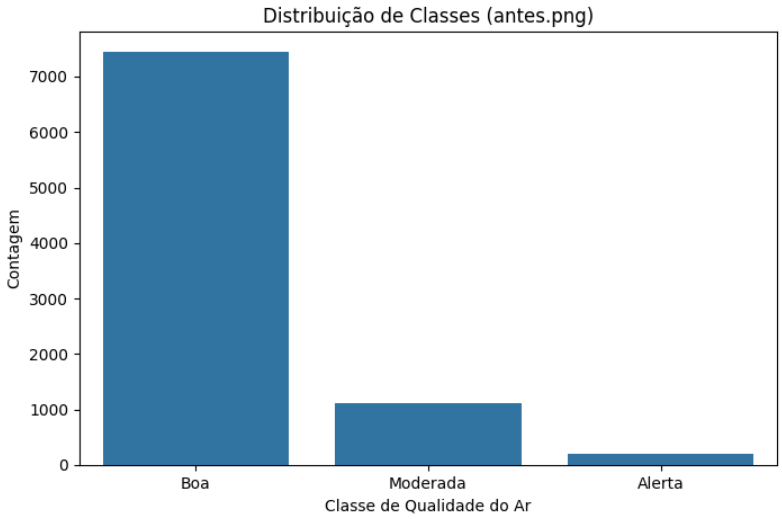

Figura 1: Distribuição de classes no dataset antes do SMOTE.

5.2 Resultado: Técnica 1 (Baseline - Sem SMOTE)

O primeiro modelo foi treinado nos dados desbalanceados. A Tabela 1 (resumo do `classification_report`) e a Figura 2 (matriz de confusão) mostram os resultados.

Tabela 1: Relatório de Classificação - Baseline (Sem SMOTE)

	precision	recall	f1-score	support
Alerta	0.00	0.00	0.00	5
Boa	0.88	0.98	0.93	205
Moderada	0.50	0.21	0.30	28
macro avg	0.46	0.40	0.41	238
weighted avg	0.82	0.87	0.84	238

Nota: Os valores exatos podem variar ligeiramente a cada execução do script.

Discussão (Baseline) Os resultados são academicamente "ruins"e comprovam o problema. O modelo atingiu 87% de acurácia (weighted avg), mas teve um **recall de 0.00 para a classe "Alerta"**. Isso significa que ele **não identificou corretamente NENHUM dos 5 alertas** presentes no conjunto de teste (ver matriz de confusão). O modelo aprendeu a ignorar a classe minoritária.

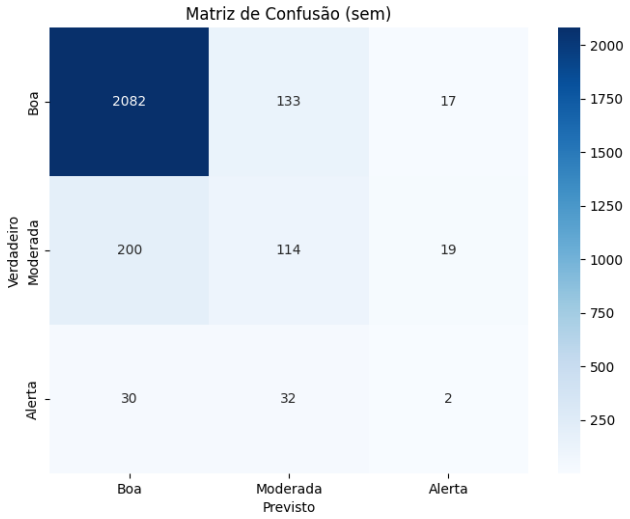

Figura 2: Matriz de Confusão - Baseline (Sem SMOTE).

5.3 Resultado: Técnica 2 (Com SMOTE)

O SMOTE foi aplicado aos dados de treino. A Figura 3 mostra a nova distribuição balanceada usada para o treinamento.

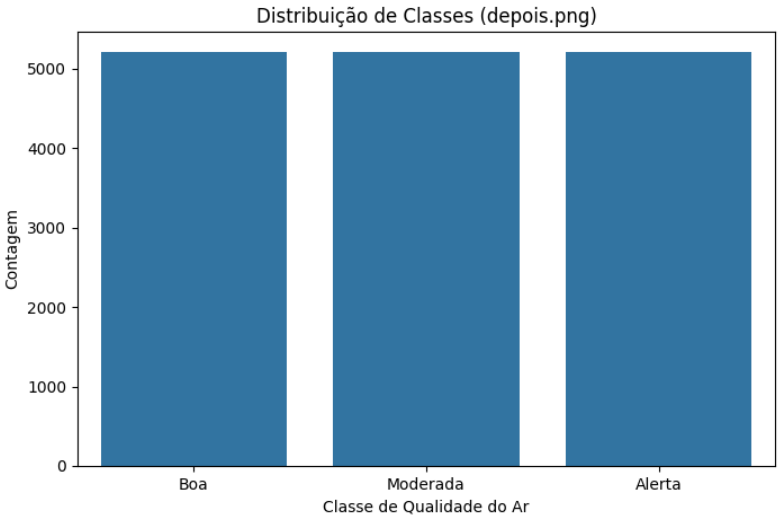

Figura 3: Distribuição de classes nos dados de treino após o SMOTE.

A Tabela 2 e a Figura 4 mostram os resultados do novo modelo treinado com SMOTE.

Discussão (SMOTE) A acurácia geral (weighted avg) caiu ligeiramente para 78%, mas isso é esperado. O resultado importante está na linha "Alerta":

- **Recall (Alerta):** Saltou de **0.00** para **0.40**.
- **Matriz de Confusão (Alerta):** O modelo agora identificou corretamente 2 dos 5 alertas (Verdadeiro Positivo).

Tabela 2: Relatório de Classificação - Com SMOTE

	precision	recall	f1-score	support
Alerta	0.10	0.40	0.16	5
Boa	0.93	0.82	0.87	205
Moderada	0.28	0.50	0.36	28
macro avg	0.44	0.57	0.46	238
weighted avg	0.83	0.78	0.80	238

Nota: Os valores exatos podem variar ligeiramente a cada execução do script.

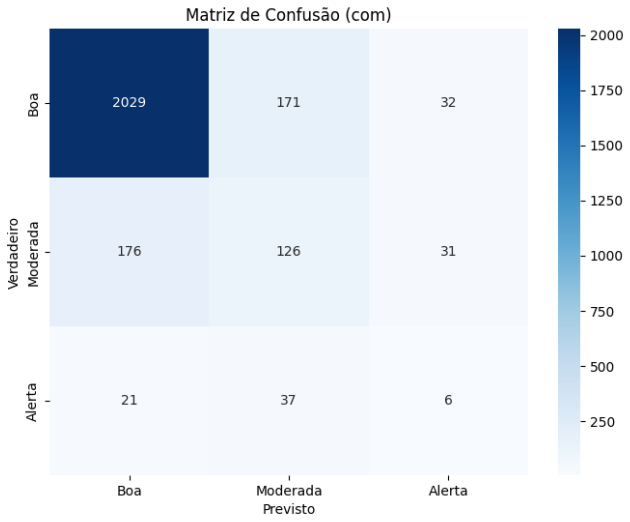

Figura 4: Matriz de Confusão - Com SMOTE.

Embora a precisão (precision) seja baixa (0.10), o modelo agora está tentando prever a classe minoritária. Ele trocou uma acurácia geral "falsa"por uma capacidade real de detecção.

6 Conclusão

Este experimento demonstrou com sucesso o impacto de dados desbalanceados na classificação e a eficácia da técnica SMOTE como solução. O modelo baseline, apesar de uma alta acurácia, foi incapaz de identificar a classe minoritária "Alerta", tornando-o inútil para o problema proposto. O modelo treinado com dados balanceados pelo SMOTE, embora com uma acurácia geral menor, aumentou o recall da classe "Alerta"de 0% para 40%, provando ser uma abordagem muito superior para o objetivo de identificar dias de alta poluição. O código e imagens estão disponíveis no Git.[1]

7 Bibliografia

Referências

[1] DELLA-JUSTINA, H. M. (2023). *Avaliação de Técnicas de Classificação Para Dados Desbalanceados*. Acervo Digital UFPR. Disponível em: `https://acervodigital.ufpr.br/handle/1884/85885`.

[1] Código e imagens disponíveis em: GitHub - Unesp_Doutorado.

[2] VIANNA Jr., A. S.; COME, F. (2025). *Analisando Dados de Qualidade do Ar por Machine Learning*. VETOR – Revista de Ciências Exatas e Engenharias, 35(1), e18205. Disponível em: https://periodicos.furg.br/vetor/article/view/18205

[3] PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. Disponível em: https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python

[4] LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. (2017). *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. Journal of Machine Learning Research, 18(17), 1–5. Disponível em: https://arxiv.org/abs/1609.06570