Relatório Técnico Comparativo — Random Forest vs Random Forest + SMOTE

Este relatório técnico apresenta a comparação entre dois modelos de classificação aplicados ao conjunto de dados de diabetes tipo II: um modelo base utilizando **Random Forest** e outro aprimorado com a técnica **SMOTE (Synthetic Minority Over-sampling Technique)**. O objetivo é avaliar o impacto do balanceamento sintético no desempenho preditivo, com foco na identificação correta de casos de diabetes (classe minoritária).

1. Metodologia

Ambos os experimentos utilizaram o mesmo dataset, com imputação de valores ausentes por mediana e padronização z-score. A divisão de treino e teste foi estratificada (80/20). - **Modelo Base (Random Forest):** sem tratamento de desbalanceamento. - **Modelo SMOTE:** aplicou balanceamento da classe minoritária no conjunto de treino, seguido de otimização de hiperparâmetros via *RandomizedSearchCV* (5-fold CV) usando F1-score como métrica principal.

2. Resultados Comparativos

Métrica	Random Forest	Random Forest + SMOTE
Precision (classe 1)	0.73	0.61
Recall (classe 1)	0.54	0.74
F1-score (classe 1)	0.62	0.67
Acurácia global	0.82	0.74
F1 ponderado	0.79	0.74

3. Discussão

O modelo **Random Forest** apresentou maior acurácia geral (0.82), mas desempenho inferior em recall (0.54), indicando tendência a favorecer a classe majoritária ('Não Diabetes'). O modelo **Random Forest + SMOTE** aumentou o recall para 0.74, melhorando a detecção de casos positivos, ainda que à custa de uma leve redução na precisão e acurácia geral. Esse comportamento é esperado em contextos clínicos, onde o custo de um falso negativo é mais relevante que o de um falso positivo.

4. Conclusão

O modelo com SMOTE demonstrou **maior sensibilidade e equilíbrio entre precisão e recall**, sendo mais apropriado para aplicações médicas e biomédicas. O modelo base, por outro lado, pode ser preferível em cenários onde a prioridade é evitar falsos positivos. **Conclusão geral:** o uso do SMOTE melhora a capacidade de generalização para a classe minoritária e é recomendado para bases desbalanceadas em diagnóstico de doenças.

5. Referências

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR*, 12, 2825–2830.
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets. *JMLR*, 18(17), 1–5.