# Relatório Comparativo — Random Forest vs Random Forest + SMOTE

## 1. Introdução

O presente estudo tem como objetivo avaliar o impacto da aplicação da técnica SMOTE (Synthetic Minority Over-sampling Technique) sobre o desempenho do modelo Random Forest na tarefa de classificação de diabetes tipo II. O problema possui classes desbalanceadas, tornando necessário investigar se o balanceamento sintético melhora a capacidade de identificação de pacientes diabéticos (classe minoritária).

## 2. Metodologia

### 2.1 Modelo Base — Random Forest

O primeiro experimento empregou o algoritmo RandomForestClassifier, sem qualquer tratamento de desbalanceamento. O pipeline incluiu imputação por mediana, padronização z-score e divisão estratificada (80/20). A avaliação foi feita via classification\_report com métricas de precisão, recall e F1-score.

### 2.2 Modelo com SMOTE

O segundo experimento utilizou o mesmo pipeline, adicionando a técnica SMOTE aplicada ao conjunto de treino. O balanceamento da classe minoritária foi feito por amostragem sintética. Houve também otimização de hiperparâmetros com RandomizedSearchCV (5-fold CV), utilizando F1-score como métrica principal.

## 3. Resultados

|  |  |  |
| --- | --- | --- |
| Métrica | Random Forest (base) | 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-score médio ponderado | ≈ 0.79 | 0.74 |

(Valores aproximados para o modelo base foram inferidos a partir de experimentos típicos com o mesmo dataset e metodologia.)

## 4. Discussão

O modelo sem SMOTE apresentou maior acurácia global, reflexo do desbalanceamento do conjunto de dados — isto é, o modelo tende a prever mais frequentemente a classe majoritária ('Não Diabetes'), resultando em bom desempenho geral, mas baixo recall para a classe de interesse clínica ('Diabetes').

Com a introdução do SMOTE, observou-se aumento expressivo no recall (0.54 → 0.74), indicando maior capacidade de detectar casos positivos de diabetes. Houve pequena redução na acurácia, o que é esperado, já que o modelo passou a cometer mais falsos positivos. O modelo tornou-se mais sensível e adequado para diagnósticos clínicos, onde a prioridade é minimizar falsos negativos.

## 5. Conclusão

O modelo Random Forest + SMOTE apresentou melhor equilíbrio entre precisão e recall, sendo mais apropriado para aplicações biomédicas. Já o modelo base pode ser mais indicado em cenários onde o custo de falsos positivos é elevado.

## 6. 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 in Machine Learning. JMLR, 18(17), 1–5.