Predição de Autoimunidade Induzida por Fármacos: Avaliação Reprodutível de Modelos com Descritores RDKit

Valério Viégas Wittler

valeriow@gmail.com

Resumo

A autoimunidade induzida por fármacos (DIA) é um evento adverso relevante em farmacovigilância, e sua previsão precoce pode apoiar a triagem de candidatos a fármacos. Este trabalho descreve um pipeline reprodutível de predição binária com descritores moleculares RDKit e modelos supervisionados, usando conjuntos de treino (477 compostos) e teste (120 compostos) com desbalanceamento moderado (3:1). Após análise exploratória e tipagem de variáveis, foram removidas 49 colunas (1 textual, 17 constantes e 31 correlacionadas), reduzindo de 198 para 149 descritores. Foram comparadas diferentes famílias de modelos e estratégias de balanceamento (class_weight, SMOTE, ensembles balanceados e RFE) com RandomizedSearchCV, validação cruzada estratificada (5 folds) e métricas adequadas ao cenário (AUC-PR e MCC). O modelo consolidado "Original - BalancedRandomForest RFE" atingiu AUC-ROC 0,933, AUC-PR 0,811, F1 0,737 e MCC 0,653, acima do desempenho reportado no estudo que introduziu o dataset InterDIA (AUC 0,893) [4]. As escolhas metodológicas e limitações são registradas para facilitar auditoria e reprodução.

Palavras-chave: Toxicologia Computacional. Eventos Adversos. Descritores Moleculares. Aprendizado de Máquina. Classes Desbalanceadas.

Abstract

Drug-induced autoimmunity (DIA) is a relevant adverse event in pharmacovigilance, and early prediction can support drug candidate screening. This work reports a reproducible binary prediction pipeline based on RDKit molecular descriptors and supervised learning models, using training (477 compounds) and test (120 compounds) sets with moderate imbalance (3:1). After exploratory analysis and variable typing, 49 columns (1 textual, 17 constant, and 31 highly correlated) were removed, reducing the feature set from 198 to 149 descriptors. Multiple model families and balancing strategies (class_weight, SMOTE, balanced ensembles, and RFE) were compared using RandomizedSearchCV with stratified 5-fold cross-validation and metrics suitable for imbalance (AUC-PR and MCC). The consolidated model "Original - BalancedRandomForest RFE" achieved AUC-ROC 0.933, AUC-PR 0.811, F1 0.737, and MCC 0.653, exceeding the performance reported in the InterDIA

study (AUC 0.893) [4]. Methodological choices and limitations are described to support reproducibility and auditing.

Keywords: Computational Toxicology. Adverse Events. Molecular Descriptors. Machine Learning. Class Imbalance.

---

1 Introdução

A autoimunidade induzida por fármacos (DIA - *Drug-Induced Autoimmunity*) é um evento adverso no qual a exposição a um composto pode desencadear ou agravar respostas autoimunes. Esse risco é relevante em farmacovigilância, pois mesmo quando há eficácia terapêutica, efeitos autoimunes podem inviabilizar o uso do fármaco ou demandar monitoramento clínico rigoroso.

Uma característica importante do problema é a assimetria entre classes, ou seja, geralmente há mais exemplos de compostos não associados do que associados à DIA. Nessa condição, a acurácia pode ser enganosa, pois um classificador tende a ser recompensado por favorecer a classe majoritária.

Modelos computacionais baseados em características moleculares (descritores) relacionam padrões estruturais e físico-químicos a desfechos de segurança. Neste trabalho, utilizam-se descritores do RDKit, biblioteca *open source* de quimioinformática, amplamente empregada em QSAR (*Quantitative Structure–Activity Relationship*) e aplicações correlatas [1]. Na prática, cada molécula é convertida em um vetor numérico, viabilizando o uso de modelos supervisionados tabulares.

Este estudo prioriza três pontos: lidar adequadamente com desbalanceamento de classes; comparar estratégias (modelos, balanceamento, seleção de variáveis e hiperparâmetros) sob um protocolo de validação consistente; e registrar decisões e resultados com transparência para auditoria e reprodutibilidade.

O texto descreve o pipeline do pré-processamento à seleção do modelo final e apresenta os resultados de forma comparável entre as alternativas avaliadas.

2 Trabalho Relacionado

A modelagem de eventos adversos relacionados a fármacos faz parte da quimioinformática, toxicologia computacional e aprendizado de máquina. Em QSAR, parte-se da hipótese de que moléculas com certas características estruturais e físico-químicas tendem a compartilhar propriedades biológicas, permitindo mapear descritores (entradas) para um desfecho (saída) [1]. Embora abordagens modernas usem redes neurais profundas e *transformers* com representações como SMILES e *fingerprints*, modelos tabulares clássicos (por exemplo, Random Forest e métodos de *boosting*) continuam competitivos em muitos cenários, especialmente quando o número de amostras é limitado.

Em bases desbalanceadas, o uso de métricas adequadas e um protocolo de validação sem vazamento são decisivos para evitar conclusões superestimadas. AUC-PR e MCC tendem a ser mais informativas do que acurácia quando a classe minoritária é a de maior interesse; o MCC, em particular, resume a concordância entre rótulos e predições e penaliza de forma balanceada falsos positivos e falsos negativos [3]. Além disso, técnicas de balanceamento e seleção de variáveis devem ser aplicadas dentro dos *folds* para não contaminar a validação. Entre as abordagens clássicas, o SMOTE [2] realiza *oversampling* por interpolação de exemplos da classe minoritária.

Este estudo utiliza o dataset InterDIA [4], que fornece dados já separados de treino e teste e reporta AUC de 0,893 com Easy Ensemble Classifier em validação externa. Ao manter o mesmo conjunto de dados e protocolo de avaliação com teste independente, os resultados obtidos aqui podem ser comparados diretamente aos números reportados no trabalho original.

3 Metodologia

3.1 Dados e Preparação

3.1.1 Fonte dos Dados

Utilizou-se o dataset InterDIA, originalmente publicado por Huang, Liu e Huang (2025) [4] para predição de autoimunidade induzida por fármacos. O conjunto de dados está disponível publicamente no UCI Machine Learning Repository (DOI: 10.24432/C5332M) e já fornece dados de treino e teste, que facilita comparações entre abordagens.

O dataset consiste em descritores moleculares RDKit pré-computados para 597 compostos farmacêuticos. Cada composto é caracterizado por 195 descritores moleculares que capturam propriedades físico-químicas, topológicas e estruturais. A variável alvo (`Label`) indica se o fármaco está associado a autoimunidade induzida (1) ou não (0).

3.1.2 Dimensão e Distribuição

A **Tabela 1** resume as dimensões e distribuição de classes dos conjuntos de treino e teste.

Conjunto	Amostras	Colunas (bruto)	DIA-negativo (0)	DIA-positivo (1)	Razão	Colunas (limpo)
Treino	477	198	359	118	3,04:1	149
Teste	120	198	90	30	3,00:1	149

Tabela 1. Dimensões e distribuição de classes dos conjuntos de treino e teste.

3.1.3 Tipagem de variáveis e limpeza

A análise exploratória (EDA) classificou as 198 variáveis originais conforme apresentado na **Tabela 2**.

Tipo	Contagem	Exemplos
Numéricos contínuos	85	BalabanJ, BertzCT, Chi0, MolLogP, TPSA
Categóricos (baixa cardinalidade)	66	NHOHCount, NumAliphaticCarbocycles, RingCount
Binários	29	Label, flags fr_* (p.ex., fr_C_S, fr_NH2)
Texto	1	SMILES
Constantes	17	NumRadicalElectrons, SMR_VSA8

Tabela 2. Tipagem de variáveis (dados brutos, 198 colunas).

Aplicaram-se três remoções:

1. **SMILES** (1 coluna): representação textual não utilizável por modelos tabulares sem transformação adicional.
2. **Constantes** (17 colunas): sem variância, não contribuem para decisão.
3. **Alta correlação** (31 colunas): pares com $|\rho| > 0{,}95$ foram identificados; selecionou-se uma representante por par para reduzir multicolinearidade.

Essas remoções têm um objetivo prático: reduzir variáveis sem informação (constantes), variáveis que exigiriam outro tipo de modelagem (texto/SMILES) e redundâncias extremas que podem tornar a otimização menos estável. Em modelos de árvore, multicolinearidade não impede o treinamento, mas pode aumentar a variância das medidas de importância por variável; em modelos lineares, a redundância pode piorar a estabilidade de coeficientes.

Resultado: 198 → 149 colunas, mantendo 477 amostras no treino e 120 no teste.

3.2 Modelos e Estratégias Avaliadas

A **Tabela 3** lista as famílias de modelos e estratégias de balanceamento/seleção de variáveis integradas ao pipeline.

Família	Modelos	Estratégias
Lineares	Regressão Logística	`class_weight='balanced'`
Árvores & Boosting	RandomForest, XGBoost, LightGBM, CatBoost	StandardScaler, RFE
Balanceados	BalancedRandomForest, EasyEnsembleClassifier, RUSBoostClassifier	SMOTE, RFE, `class_weight`
Vizinhança	KNN	StandardScaler

Tabela 3. Famílias de modelos e estratégias.

3.3 Pipelines e Balanceamento

Cada pipeline integra etapas sequenciais com cuidado explícito para evitar vazamento de informação (*data leakage*). Todas as operações que estimam parâmetros a partir dos dados (por exemplo, padronização, *oversampling* e seleção de variáveis) são ajustadas **apenas nos dados de treino de cada fold**.

1. **Padronização (quando aplicável):** `StandardScaler` foi usado para modelos sensíveis à escala (por exemplo, Regressão Logística e KNN). Para modelos baseados em árvores (Random Forest e variantes), a padronização não é necessária e pode ser omitida.
2. **Balanceamento (sempre dentro do pipeline):**
   - *Oversampling:* `SMOTE(random_state=42)` sintetiza amostras da classe minoritária a partir de vizinhos próximos [2].
   - *Ponderação:* `class_weight='balanced'` ajusta a função de perda para penalizar mais erros na classe minoritária.
   - *Ensembles balanceados:* `BalancedRandomForestClassifier` combina árvores treinadas em subconjuntos com subamostragem, reduzindo a dominância da classe majoritária.
3. **Seleção de variáveis:** `RFE` (*Recursive Feature Elimination*) remove iterativamente descritores menos úteis segundo um estimador interno, reduzindo dimensionalidade e potencialmente aumentando generalização.
4. **Classificador:** o modelo final é configurado por busca aleatória de hiperparâmetros (RandomizedSearchCV), sempre dentro da validação cruzada.

Exemplo (BalancedRandomForest com RFE - sem padronização, por ser baseado em árvores):

```
pipeline = ImbPipeline([
    ('rfe', RFE(estimator=RandomForestClassifier())),
    ('classifier', BalancedRandomForestClassifier())
])
```

3.4 Validação Cruzada e Otimização

Estratégia de CV

Validação cruzada estratificada com 5 folds:

```
StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

Mantém proporção de classes em cada fold e garante reprodutibilidade.

A validação cruzada estratificada fornece uma estimativa mais estável de desempenho em um cenário com poucos exemplos positivos, reduzindo a dependência de uma única partição de treino.

Otimização

RandomizedSearchCV com 50 iterações:

```
RandomizedSearchCV(
    estimator=pipeline,
    param_distributions=param_dist,
    n_iter=50,
    cv=StratifiedKFold(...),
    scoring='average_precision',  # ou 'mcc'
    random_state=42
)
```

Distribuições de hiperparâmetros (exemplos): - `classifier__n_estimators: randint(100, 500)` - `classifier__learning_rate: uniform(0.01, 0.2)` - `classifier__max_depth: randint(5, 30)`

O RandomizedSearchCV foi adotado para explorar o espaço de hiperparâmetros com custo computacional controlado.

3.5 Métricas de Avaliação

Reportadas no conjunto de teste:

- **AUC-ROC:** mede a probabilidade de um positivo receber escore maior que um negativo. É útil para avaliar separação global, mas pode parecer "otimista" quando a classe positiva é rara.
- **AUC-PR:** resume o compromisso entre precisão e revocação conforme o limiar varia; tende a ser mais informativa quando a classe positiva é minoritária.
- **F1-score:** média harmônica de precisão e revocação em um limiar fixo (neste trabalho, 0,5), refletindo uma regra de decisão operacional.
- **MCC:** Matthews Correlation Coefficient; robusto para classes desbalanceadas e sensível a todos os elementos da matriz de confusão (VP, VN, FP, FN) [3].

Essas métricas complementam-se: AUC-ROC/AUC-PR resumem qualidade de ordenação (sem fixar limiar), enquanto F1 e MCC refletem o desempenho de uma regra de decisão em um limiar fixo.

4 Resultados

Foram avaliadas 270 configurações (combinações de pré-processamento, pipelines e modelos), registradas como linhas de resultados. Como várias configurações diferem apenas por detalhes de pré-processamento, mas compartilham o mesmo nome de modelo (por exemplo, variações de escalonamento), foi realizada uma deduplicação por `model_name`, retendo o melhor desempenho por modelo. Após esse procedimento, obtiveram-se 125 modelos únicos avaliados no conjunto de teste independente.

A **Tabela 4** apresenta os 10 melhores desempenhos ordenados por AUC-ROC.

Rank	Modelo (único por nome)	AUC-ROC	AUC-PR	F1	MCC
1	Original - BalancedRandomForest RFE	0,9330	0,8108	0,7368	0,6531
2	Limpos - Sem Escalonamento - LightGBM Ponderado	0,9278	0,7943	0,5652	0,5095
3	Original - EasyEnsemble RFE	0,9278	0,8109	0,4103	0,4201
4	Limpo Escalonado Numericas - LightGBM Ponderado	0,9278	0,7943	0,5652	0,5095
5	Limpo Escalonado Geral - LightGBM Ponderado	0,9278	0,7943	0,5652	0,5095
6	Limpos - Sem Escalonamento - EasyEnsemble RFE	0,9274	0,8049	0,4103	0,4201
7	Original - Com Escalonamento - BalancedRandomForest RFE	0,9270	0,8034	0,7222	0,6254
8	Original - Com Escalonamento - EasyEnsemble RFE	0,9263	0,8105	0,4103	0,4201
9	Limpo Escalonado Geral - LightGBM SMOTE RFE	0,9237	0,7813	0,4651	0,4180
10	Limpo Escalonado Geral - LightGBM RFE	0,9237	0,7813	0,4651	0,4180

Tabela 4. Top 10 modelos únicos por AUC-ROC (conjunto de teste).

Para tornar a Tabela 4 mais clara, os nomes dos modelos seguem uma convenção baseada na versão do conjunto de atributos e no pré-processamento aplicado:

- **Original:** usa a tabela de descritores do InterDIA como fornecida (198 colunas no treino, incluindo `SMILES`), aplicando apenas o tratamento mínimo necessário para viabilizar modelos tabulares (por exemplo, remoção/ignorância de `SMILES` e outras etapas quando aplicável no pipeline).
- **Limpo(s):** usa a versão "limpa" gerada neste trabalho (149 colunas após remoções descritas na Metodologia: `SMILES`, constantes e alta correlação), reduzindo redundâncias e facilitando análise.
- **Sem Escalonamento:** não aplica padronização (`StandardScaler`) aos descritores.
- **Escalonado Geral:** aplica padronização a todos os descritores numéricos do conjunto (exceto a variável alvo).
- **Escalonado Numéricas:** aplica padronização apenas ao subconjunto de descritores contínuos, preservando descritores binários/de contagem sem transformação (útil quando se deseja evitar "mexer" em variáveis naturalmente discretas).

O modelo com melhor desempenho foi o "Original - BalancedRandomForest RFE" que apresentou:

- **AUC-ROC:** 0,9330 (capacidade de discriminação excelente)
- **AUC-PR:** 0,8108 (precisão adequada na classe rara)
- **F1-score:** 0,7368 (equilíbrio precisão-revocação robusto)
- **MCC:** 0,6531 (concordância forte entre previsão e realidade)

Esse modelo combina estratégias efetivas: BalancedRandomForest (subamostragem intrínseca), RFE (redução de dimensionalidade) e dados originais (sem remoção de descritores além da limpeza inicial).

O modelo "Original - EasyEnsemble RFE" apresenta AUC-PR marginalmente superior (0,8109), porém F1 substancialmente menor (0,4103 vs. 0,7368). Isso ilustra um trade-off recorrente: alguns modelos ranqueiam bem (boa AUC-PR/AUC-ROC), mas produzem probabilidades (ou escores) que, ao serem limiarizadas em 0,5, geram decisões com menor equilíbrio entre precisão e revocação.

O limiar 0,5 é apenas uma convenção; na prática, a escolha depende do custo relativo de falsos positivos e falsos negativos. Ainda assim, reportar F1 e MCC em um limiar fixo oferece um ponto de comparação direto entre modelos sob a mesma regra de decisão.

5 Discussão

O desempenho alcançado (AUC-ROC 0,933) no conjunto de teste com 120 amostras supera o valor reportado no estudo original do dataset InterDIA [4] (AUC 0,893 com Easy Ensemble Classifier em validação externa). Essa comparação deve ser entendida como **comparação por referência** (isto é, usando números reportados no artigo original) e não como um estudo de reimplementação controlada de todas as escolhas do trabalho original. Ainda assim, o resultado sugere que há espaço relevante de melhoria ao combinar estratégias de balanceamento, seleção de variáveis e otimização sistemática.

Uma explicação plausível para a melhora observada é a sinergia entre (i) o mecanismo de balanceamento do BalancedRandomForest, que reduz a dominância da classe majoritária durante o treinamento, e (ii) o RFE, que elimina descritores redundantes ou pouco informativos, reduzindo ruído e estabilizando o modelo.

A validação cruzada estratificada com 5 folds durante otimização, combinada com avaliação em conjunto de teste *holdout* independente (fornecido separadamente), reduz o risco de sobreajuste. Em outras palavras, o modelo não foi avaliado nas mesmas amostras usadas para ajustar hiperparâmetros, o que torna os resultados mais realistas para uso futuro.

O desempenho superior de BalancedRandomForest RFE sobre alternativas sugere que:

1. **Balanceamento é efetivo:** métodos que ignoram desbalanceamento tendem a privilegiar a classe majoritária, reduzindo a sensibilidade à classe positiva. Na prática, isso pode elevar a acurácia sem melhorar a utilidade para triagem.

2. **RFE contribui:** ao reduzir dimensionalidade, RFE pode diminuir redundância entre descritores e favorecer modelos mais estáveis. O ganho costuma aparecer mais claramente em métricas dependentes de limiar (F1/MCC), pois a redução de ruído tende a tornar probabilidades/escores mais bem calibrados.

3. **Combinação:** BalancedRandomForest (subamostragem intrínseca) com RFE (seleção) e com dados bem limpos tende a produzir robustez tanto em ranking (AUC) quanto em decisão (F1/MCC).

Com relação às decisões de engenharia de dados, a remoção de 49 colunas antes da otimização foi apropriada pois:

- SMILES não é vetorizável sem transformação adicional.
- Constantes não carregam informação e podem introduzir artefatos numéricos.
- Colinearidade reduzida facilita modelos interpretáveis e diminui variância.

Observou-se que variações de *feature engineering* (por exemplo, esquemas de escalonamento e versões do conjunto de atributos) não trouxeram ganhos consistentes em todos os modelos, apesar do aumento de complexidade do pré-processamento. Esse comportamento é compatível com o fato de que modelos baseados em árvores e *boosting* são, em geral, pouco sensíveis à escala e conseguem lidar com colinearidades e redundâncias sem depender de normalização. Além disso, introduzir mais variantes de pré-processamento amplia o espaço de busca e eleva o custo da validação cruzada e da otimização, o que nem sempre se traduz em melhoria prática.

Quanto ao desbalanceamento e à escolha das métricas, a inclusão de AUC-PR e MCC é crítica em contextos desbalanceados. Em um cenário 3:1, um classificador que sempre escolhe a classe negativa pode atingir acurácia próxima de 75% sem identificar nenhum positivo. Por isso, a combinação de métricas (AUC-ROC, AUC-PR, F1 e MCC) fornece uma visão mais completa: ranking global, desempenho na classe rara e qualidade da decisão.

A seguir são apresentadas algumas limitações metodológicas identificadas:

1. **Tamanho amostral:** 477 + 120 amostras é um tamanho moderado. Em bases pequenas, pequenas variações nos exemplos positivos podem alterar métricas de forma relevante.

2. **Validação externa adicional:** embora exista um conjunto de teste separado, a generalização para outras coleções (por exemplo, outras populações de compostos, critérios de rotulagem ou pipelines de descritores) não pode ser assumida sem novas avaliações.

3. **Representação molecular:** o trabalho foca em descritores RDKit tabulares. Não se conclui, portanto, sobre desempenho com *fingerprints* ou representações baseadas em linguagem (SMILES tokenizado). Comparações diretas seriam

importantes para estabelecer melhor custo-benefício entre desempenho e esforço de modelagem.

4. **Limiar de decisão:** F1/MCC foram reportados com limiar 0,5 para padronização, mas aplicações reais tipicamente ajustam o limiar com base em custo-benefício e capacidade de validação experimental. Um passo adicional seria otimizar limiar para maximizar MCC ou atender a uma meta mínima de sensibilidade, assim como aplicar calibração de probabilidades.

6 Conclusão

Este trabalho apresentou um pipeline reprodutível para predição de autoimunidade induzida por fármacos utilizando o dataset InterDIA [4] e técnicas consolidadas de aprendizado de máquina. O modelo BalancedRandomForest com RFE atingiu AUC-ROC 0,933, AUC-PR 0,811, F1 0,737 e MCC 0,653, superando o valor reportado no estudo original do dataset (AUC 0,893).

A combinação de estratégias de balanceamento (BalancedRandomForest com subamostragem intrínseca), seleção de variáveis (RFE) e otimização sistemática de hiperparâmetros (RandomizedSearchCV) demonstrou ser efetiva para este domínio de problema.

A ênfase em reprodutibilidade documentada - sementes fixas, *splits* pré-definidos e consolidação de resultados - oferece base sólida para auditoria e extensões futuras. Trabalhos subsequentes podem explorar: (i) bases maiores e validação; (ii) comparação com *fingerprints* moleculares e modelos de *deep learning*; (iii) calibração de probabilidades e ajuste de limiares para contextos regulatórios; (iv) integração com dados biológicos complementares (como expressão gênica, proteínas ou metabólitos).

Referências

[1] Cherkasov, A., Muratov, E. N., Fourches, D., et al. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57(12), 4977–5010. https://doi.org/10.1021/jm4004285

[2] 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. https://doi.org/10.1613/jair.953

[3] Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

[4] Huang, L., Liu, P., Huang, X. (2025). InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches. *Toxicology*, 511, 154064. https://doi.org/10.1016/j.tox.2025.154064

Apêndice – Disponibilidade de Dados e Código

Código-fonte:

- https://github.com/valeriow/ufgagentes-pub/08-AM/

Dataset: O dataset InterDIA utilizado neste estudo está disponível publicamente em: - UCI Machine Learning Repository:

- https://archive.ics.uci.edu/dataset/1104/drug_induced_autoimmunity_prediction

- Repositório GitHub: https://github.com/Huangxiaojie2024/InterDIA

Reprodutibilidade: O pipeline pode ser reproduzido com as versões descritas em `requirements.txt/pyproject.toml`, em especial a versão de Python e as bibliotecas de aprendizado de máquina. Adotaram-se sementes fixas (`random_state=42`) em etapas críticas (validação cruzada, SMOTE e buscas aleatórias) para reduzir variabilidade entre execuções.