Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

doi.org/10.51891/rease.v12i1.23728

DESENVOLVIMENTO DE UM MODELO PREDITIVO IN SILICO PARA ESTIMATIVA DE TOXICIDADE DE MOLÉCULAS UTILIZANDO DADOS PÚBLICOS E INTELIGÊNCIA ARTIFICIAL

DEVELOPMENT OF AN IN SILICO PREDICTIVE MODEL FOR ESTIMATING MOLECULAR TOXICITY USING PUBLIC DATA AND ARTIFICIAL INTELLIGENCE

DESARROLLO DE UN MODELO PREDICTIVO IN SILICO PARA LA ESTIMACIÓN DE LA TOXICIDAD DE MOLÉCULAS UTILIZANDO DATOS PÚBLICOS E INTELIGENCIA ARTIFICIAL

Casimiro Waete Agostinho[1]
Grazielly Honorio Rodrigues de Freitas[2]
John Henrique Soares Costa[3]
Maria Eduarda de Melo Pretes[4]
William Argolo Saliba[5]

RESUMO: Esse artigo buscou desenvolver um modelo preditivo *in silico*, em linguagem Python, para estimar a toxicidade de pequenas moléculas orgânicas utilizando dados públicos e técnicas de Inteligência Artificial. Para isso, foi construído um conjunto de dados com 200 moléculas contendo até 10 átomos de carbono, selecionadas no repositório PubChem, priorizando compostos halogenados e amínicos estruturalmente relacionados a cloroaminas e halometanos. Foram extraídos descritores estruturais e físico-químicos (como massa molar, tipo de cadeia, quantidade de halogênios, proporção halogênio/carbono, anéis alifáticos e aromáticos, carbonos quirais e função orgânica), além de uma variável-alvo binária de toxicidade. A modelagem foi conduzida em Google Colab, empregando *Random Forest* e regressão logística, com tratamento de desbalanceamento por SMOTENC e avaliação por *holdout* (70/30) e validação cruzada estratificada. O *Random Forest* apresentou desempenho global superior (*accuracy* 0,9333; *balanced_accuracy* 0,8693; ROC-AUC 0,9673), enquanto a regressão logística maximizou o *recall* (0,9804) e forneceu maior interpretabilidade, evidenciando maior risco associado à halogenação e à aromaticidade e efeito protetor de anéis alifáticos e de maior número de hidrogênios ligados ao nitrogênio. Conclui-se que o *pipeline* proposto é promissor para triagem toxicológica preliminar, embora a ampliação e a validação externa da base sejam essenciais para aumentar a robustez e a generalização dos modelos.

Palavras-chave: Toxicidade molecular. Inteligência artificial. Modelo preditivo *in sílico*.

[1] Discente do curso de Ciência de Dados e Inteligência Artificial do Centro Universitário Única.
[2] Discente do curso de Química do Centro Universitário Única
[3] Discente do curso de Farmácia do Centro Universitário Única
[4] Discente do curso de Farmácia do Centro Universitário Única.
[5] Docente do Centro Universitário Única - Prof. Orientador. Centro Universitário Única – UNIÚNICA.

ABSTRACT: This article aimed to develop an in silico predictive model, implemented in Python, to estimate the toxicity of small organic molecules using public data and Artificial Intelligence techniques. A dataset of 200 molecules containing up to 10 carbon atoms was constructed from the PubChem repository, prioritizing halogenated and amino compounds structurally related to chloramines and halomethanes. Structural and physicochemical descriptors were extracted, including molar mass, chain type, number of halogens, halogen/carbon ratio, aliphatic and aromatic rings, chiral carbons, and main organic function, in addition to a binary toxicity target variable. Modeling was performed in Google Colab using Random Forest and logistic regression, with class imbalance handled by SMOTENC and performance assessed via holdout (70/30) and stratified cross-validation. Random Forest showed superior overall performance (accuracy 0.9333; balanced accuracy 0.8693; ROC-AUC 0.9673), whereas logistic regression maximized recall (0.9804) and provided greater interpretability, indicating higher risk associated with halogenation and aromaticity and a protective effect of aliphatic rings and a greater number of hydrogens bound to nitrogen. It is concluded that the proposed pipeline is promising for preliminary toxicological screening, although expansion and external validation of the dataset are essential to increase model robustness and generalizability.

Keywords: Molecular toxicity. Artificial intelligence. In silico predictive model.

RESUMEN: Este artículo buscó desarrollar un modelo predictivo in silico, implementado en Python, para estimar la toxicidad de pequeñas moléculas orgánicas utilizando datos públicos y técnicas de Inteligencia Artificial. Se construyó un conjunto de datos con 200 moléculas que contienen hasta 10 átomos de carbono a partir del repositorio PubChem, priorizando compuestos halogenados y amínicos estructuralmente relacionados con cloraminas y halometanos. Se extrajeron descriptores estructurales y fisicoquímicos, incluyendo masa molar, tipo de cadena, número de halógenos, proporción halógeno/carbono, anillos alifáticos y aromáticos, carbonos quirales y función orgánica principal, además de una variable objetivo binaria de toxicidad. El modelado se realizó en Google Colab utilizando Random Forest y regresión logística, con el desbalance de clases tratado mediante SMOTENC y el desempeño evaluado por holdout (70/30) y validación cruzada estratificada. El modelo Random Forest presentó un desempeño global superior (accuracy 0,9333; balanced accuracy 0,8693; ROC-AUC 0,9673), mientras que la regresión logística maximizó el recall (0,9804) y ofreció mayor interpretabilidad, indicando mayor riesgo asociado a la halogenación y a la aromaticidad y un efecto protector de los anillos alifáticos y de un mayor número de hidrógenos unidos al nitrógeno. Se concluye que el pipeline propuesto es prometedor para el cribado toxicológico preliminar, aunque la ampliación y la validación externa de la base de datos son esenciales para aumentar la robustez y la capacidad de generalización de los modelos.

Palabras clave: Toxicidad molecular. Inteligencia artificial. Modelo predictivo in sílico.

INTRODUÇÃO

No contexto contemporâneo da indústria farmacêutica e da avaliação de risco químico, a Inteligência Artificial (IA) e o Aprendizado de Máquina têm se consolidado como abordagens capazes de acelerar a caracterização de perfis toxicológicos e apoiar a descoberta e o desenvolvimento de novas moléculas. Essa evolução em Pesquisa e Desenvolvimento (P&D)

decorre, em grande parte, da disponibilidade crescente de bases químicas públicas e do avanço de algoritmos que integram descritores moleculares e dados toxicológicos para produzir estimativas rápidas, reprodutíveis e potencialmente escaláveis, reduzindo a dependência de ensaios laboratoriais extensos nas etapas iniciais de triagem (CAMPOS e VASCONCELOS, 2021). Nessa perspectiva, o presente estudo se insere ao propor um *pipeline in silico*, em Python, para predição binária de toxicidade com base em informações estruturais e físico-químicas obtidas em repositório público.

A necessidade de métodos computacionais torna-se particularmente evidente ao considerar moléculas halogenadas e N-halogenadas, de elevada relevância ambiental e sanitária, que incluem classes como cloroaminas (derivados N-clorados associados a processos de desinfecção) e haloalcanos, frequentemente relacionados a efeitos adversos em diferentes sistemas biológicos. Apesar dos avanços em ensaios *in vivo* e *in vitro*, a caracterização toxicológica de uma ampla diversidade dessas substâncias permanece incompleta, sobretudo em razão da heterogeneidade estrutural, das variações de grupos funcionais e de limitações práticas e éticas inerentes à experimentação, como custo, tempo e restrições no uso de modelos biológicos. Além disso, a compreensão do destino ambiental de haloalcanos envolve processos de biodegradação mediados por enzimas como haloalcano desalogenases, que clivam ligações carbono–halogênio e influenciam diretamente persistência e risco associado (GEORGAKIS N *et al.*, 2017), reforçando a importância de ferramentas que antecipem potenciais perigos a partir da estrutura molecular.

Diante desse cenário, evidencia-se uma lacuna central: a demanda por estratégias que permitam estimar toxicidade de modo transparente, reprodutível e em larga escala, utilizando informações estruturais acessíveis publicamente. Assim, formula-se o problema de pesquisa deste artigo: como estimar, de maneira reprodutível e escalável, a toxicidade de moléculas halogenadas e N-halogenadas — com ênfase em cloroaminas e haloalcanos — a partir de dados públicos, por meio de modelos preditivos in silico implementados em Python e técnicas de IA, reduzindo a dependência de testes laboratoriais nas etapas iniciais?

A justificativa científica e aplicada do estudo apoia-se em dois eixos complementares. Primeiro, modelos preditivos *in silico* viabilizam triagem de risco e priorização de substâncias, contribuindo para decisões regulatórias, ambientais e industriais ao antecipar potenciais perigos antes de exposições em larga escala. Segundo, a integração entre repositórios públicos (como o PubChem) e métodos de aprendizagem de máquina fortalece transparência e reprodutibilidade, além de otimizar recursos ao direcionar a experimentação para casos de maior criticidade.

Evidências recentes reforçam a necessidade de ferramentas que superem análises pontuais e captem tendências estruturais associadas a efeitos adversos, o que sustenta a relevância de abordagens preditivas aplicadas às classes investigadas neste trabalho.

O objetivo geral deste artigo foi desenvolver, com técnicas de *Machine Learning*, um modelo preditivo *in silico* para estimar a toxicidade de moléculas a partir de dados públicos, utilizando linguagem Python e técnicas de IA.

Como objetivos específicos, pretendeu-se: (i) revisar a literatura sobre aprendizagem molecular aplicada à predição, com foco em cloroaminas e haloalcanos; (ii) obter e organizar dados estruturais, físico-químicos e toxicológicos no repositório público PubChem; (iii) estruturar um conjunto de variáveis/descritores coerente com o pipeline do estudo — incluindo fórmula molecular, nome, massa molar, tipo de cadeia (aberta/fechada), hidrogênios ligados a nitrogênio, quantidade de halogênios e proporção halogênio/carbono, presença de anéis alifáticos e aromáticos, função(ões) orgânica(s), número de carbonos quirais e representação em SMILES — e definir a variável-alvo binária de toxicidade; e (iv) treinar e avaliar modelos de classificação, com ênfase em robustez diante de desbalanceamento de classes, comparando desempenho e interpretabilidade para apoiar a estimativa de toxicidade em moléculas estruturalmente relacionadas às classes investigadas no conjunto analisado.

MÉTODOS

Foi construído um conjunto de dados contendo 200 moléculas orgânicas com até 10 átomos de carbono. As informações estruturais, físico-químicas e toxicológicas dessas moléculas foram obtidas a partir do repositório público PubChem (https://pubchem.ncbi.nlm.nih.gov/), uma fonte amplamente utilizada em química computacional e toxicologia preditiva. Os dados foram inicialmente estruturados em planilha eletrônica (Microsoft Excel), a partir da qual se construiu a base de análise.

Para cada molécula, foram extraídos e organizados descritores que contemplam aspectos estruturais e físico-químicos, incluindo fórmula molecular, nome da substância, massa molar, tipo de cadeia (variável binária: 0 = cadeia aberta; 1 = cadeia fechada), quantidade de hidrogênios ligados ao átomo de nitrogênio, quantidade total de halogênios, proporção halogênio/carbono, quantidade de anéis alifáticos e aromáticos, função(ões) orgânica(s) principal(is), quantidade de carbonos quirais e a representação estrutural em SMILES. A toxicidade foi definida como variável alvo binária, indicada pela presença de qualquer evidência de toxicidade (Não tóxica = 0 e tóxica = 1).

A etapa de mineração, tratamento e análise foi conduzida em ambiente Google Colab, utilizando linguagem Python e bibliotecas apropriadas para ciência de dados e aprendizado de máquina. Realizou-se limpeza e padronização das variáveis, bem como codificação de atributos categóricos quando necessário, com preparação do conjunto de dados para modelagem preditiva. Considerando a possibilidade de desbalanceamento entre as classes da variável alvo e a presença de preditores de natureza mista (numéricos e categóricos, como "Função Orgânica"), foi empregada a técnica SMOTENC (*Synthetic Minority Over-sampling Technique for Nominal and Continuous*) para reamostragem da classe minoritária durante o treinamento. O SMOTENC foi aplicado exclusivamente ao conjunto de treinamento e no interior do pipeline de validação, após a separação treino–teste, a fim de reduzir viés do aprendizado sem introduzir vazamento de informação do conjunto de teste; adicionalmente, essa escolha preserva a coerência de variáveis categóricas (evitando a geração de codificações "fracionárias" típicas do SMOTE quando aplicado diretamente após *one-hot encoding*).

Para a classificação da toxicidade, foi adotado um modelo de Random Forest (Floresta Randômica), por se tratar de um algoritmo supervisionado robusto a não linearidades e interações entre descritores. O desempenho preditivo foi avaliado pelo método holdout, com particionamento de 70% das observações para treinamento e 30% para teste, visando estimar a capacidade de generalização em dados não vistos. A avaliação do modelo utilizou um conjunto abrangente de métricas, incluindo *accuracy*, *balanced accuracy*, *precision*, *recall*, *F1-score*, área sob a curva ROC (ROC-AUC), área sob a curva precisão–revocação (PR-AUC) e coeficiente de correlação de Matthews (MCC), permitindo análise mais robusta sob desbalanceamento de classes.

Adicionalmente, foi ajustado um modelo de regressão logística com o objetivo de obter uma função probabilística explícita para estimar a probabilidade de toxicidade em função dos descritores selecionados. Essa abordagem complementou o modelo baseado em árvores ao favorecer interpretabilidade, permitindo avaliar a contribuição relativa dos preditores e fornecer uma expressão matemática para predição probabilística de toxicidade em novas moléculas com estruturas semelhantes a cloroaminas e halometanos.

RESULTADOS

Foi construído um conjunto de dados com 200 moléculas orgânicas (até 10 átomos de carbono), cujos descritores estruturais, físico-químicos e toxicológicos foram inicialmente organizados em planilha eletrônica (Microsoft Excel) e posteriormente processados em

ambiente Google Colab (Python). Optou-se por compostos orgânicos com até 10 carbonos, possuindo halogênios e grupos aminas, pois essas estruturas se assemelham às moléculas de cloroaminas e halometanos.

Após limpeza e padronização, não foram identificados valores nulos nas variáveis utilizadas para modelagem, e os tipos de dados mostraram-se compatíveis com o problema: massa molar e proporção halogênios/carbonos como variáveis contínuas (*float64*), descritores de contagem/indicadores como inteiros (*int64*) e "Função Orgânica" como variável categórica (*object*). Persistiram 10 observações duplicadas no *dataframe* de modelagem, aspecto metodologicamente relevante porque duplicatas (ou compostos extremamente semelhantes) podem introduzir dependência entre amostras e inflar estimativas de desempenho caso sejam distribuídas entre treino e teste ou entre dobras de validação cruzada.

A variável alvo "Tóxico (0/1)" foi definida como binária (0 = não tóxico; 1 = presença de qualquer evidência de toxicidade). A distribuição reportada foi de 169 moléculas tóxicas (classe 1) e 30 não tóxicas (classe 0), caracterizando desbalanceamento acentuado em favor da classe positiva (TAB.1). Observa-se ainda que a soma (199) não totaliza 200, sugerindo que uma observação foi removida ou filtrada em alguma etapa do fluxo (por inconsistência, deduplicação parcial, ou outro critério).

Tabela 1: Amostragem utilizada para o desenvolvimento do modelo.

Classe	Código binário	contagem
Tóxica	1	169
Não Tóxica	0	30

Fonte: Saliba *et al.*, 2026.

A estratificação por "Função Orgânica" evidenciou heterogeneidade marcada na proporção de toxicidade entre classes químicas, com grupos altamente prevalentes e fortemente associados ao rótulo. "Haleto orgânico" foi a categoria mais numerosa (69 moléculas) e apresentou 94,2% de tóxicos (65/69). Diversas categorias relacionadas a halogenação e/ou aromaticidade apresentaram proporção de toxicidade igual a 1,0 (100%), como "Aromático halogenado" (13/13) e "Amina aromática halogenada" (15/15), enquanto classes como "Aminoácido" (16/16 não tóxicos) e "Alcano" (2/2 não tóxicos) exibiram proporção nula de toxicidade no conjunto analisado. Embora isso sugira elevado poder discriminante dessa variável categórica, deve-se considerar que várias categorias têm suporte amostral muito baixo

(frequentemente n=1), o que pode favorecer aprendizado de padrões idiossincráticos e reduzir capacidade de generalização (FIG.1).

Figura 1 – Proporção de compostos orgânicos tóxicos segmentado por função orgânica.

Fonte: Saliba *et al.*, 2026.

Para mitigar efeitos do desbalanceamento durante o treinamento, empregou-se SMOTENC (*Synthetic Minority Over-sampling Technique for Nominal and Continuous*), técnica indicada quando há mistura de variáveis numéricas e categóricas. Diferentemente do SMOTE aplicado após codificação *one-hot* (que pode gerar combinações fracionárias artificiais em dimensões binárias), o SMOTENC preserva a natureza categórica das variáveis nominativas ao sintetizar novos exemplos da classe minoritária de modo consistente com essas restrições. Do ponto de vista metodológico, a reamostragem foi aplicada exclusivamente no conjunto de treino e dentro do *pipeline* de validação, de modo a evitar vazamento de informação para o conjunto de teste; assim, a distribuição 169/30 descreve o conjunto original, enquanto o balanceamento atua apenas na etapa de ajuste do modelo.

A tabela 2 sintetiza a ausência de nulos e os tipos das variáveis utilizadas no *dataframe* de modelagem.

Tabela 2 - Estrutura do *dataframe* de modelagem (nulos e tipos de dados).

Variável	n_nulos	dtype
Massa Mol.	0	float64
Ciclo (0=aberto,1=fechado)	0	int64
H ligados a N	0	int64
Qtd Halogênios	0	int64

Variável	n_nulos	dtype
Hal/C proporção	0	float64
Anéis Alif.	0	int64
Anéis Arom.	0	int64
Função Orgânica	0	object
C quirais	0	int64
Tóxico (0/1)	0	int64

Fonte: Saliba *et al.*, 2026.

O desempenho do classificador *Random Forest*, avaliado por *holdout* (70% treino; 30% teste) com n=60 amostras, demonstrou resultados superiores aos encontrados por Freitas *et al.*, 2021. Enquanto estudos de doenças cardiometabólicas com este mesmo algoritmo alcançaram uma acurácia de 0,869, o presente modelo atingiu 0,9333. A acurácia balanceada de 0,8693 e o MCC de 0,7386 evidenciam um equilíbrio preditivo robusto, superando o desafio comum de desbalanceamento de classes (TAB.3). Essa preocupação com o equilíbrio é corroborada pelo estudo realizado por Freitas *et al.*, 2021, que destacou que métricas como a acurácia isolada podem mascarar falhas na detecção da classe minoritária, o que em saúde pode representar falsos negativos críticos para o tratamento

Tabela 3 - Métricas no conjunto de teste (*holdout* 70/30) com aplicação da técnica SMOTENC.

Métricas	Valor
accuracy	0,9333
balanced_accuracy	0,8693
precision	0,9608
recall	0,9608
F1	0,9608
roc_auc	0,9673
pr_auc	0,9941
mcc	0,7386
n_teste	60

Fonte: Saliba *et al.*, 2026.

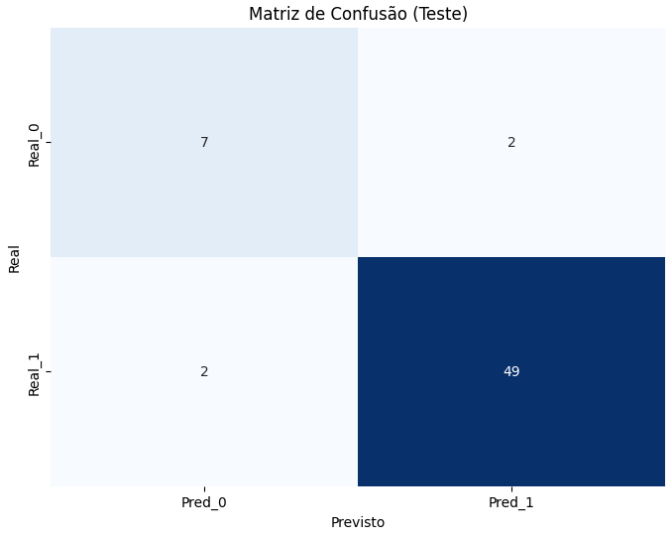
No que tange à capacidade de discriminação, os valores de ROC-AUC (0,9673) e PR-AUC (0,9941) indicam uma separação quase ideal entre compostos tóxicos e não tóxicos. Esse

desempenho é notavelmente superior à média de 93% de AUC ROC obtida nos modelos de predição testada por Freitas *et al.*, 2021.

Além disso, os resultados obtidos contrastam positivamente com o cenário atual dos modelos de ecotoxicidade em abelhas (QSTR), nos quais muitos modelos publicados enfrentam limitações de preditividade devido ao uso de bases de dados reduzidas e falta de padronização (Lemes *et al.*, 2019). Portanto, a boa performance aqui observada sugere que o modelo não apenas é robusto, mas também atende à necessidade urgente de ferramentas computacionais mais eficientes para a avaliação de risco ambiental e biológico de moléculas semelhantes às cloroaminas e halometanos.

A matriz de confusão (FIG.2) mostra que, dentre as 9 amostras reais da classe 0 (Não tóxica) no teste, 7 foram corretamente classificadas como não tóxicas, enquanto 2 foram incorretamente classificadas como tóxicas (falsos positivos). Para a classe 1(Tóxica) (51 amostras), 49 foram corretamente classificadas como tóxicas e 2 foram incorretamente classificadas como não tóxicas (falsos negativos). Em triagens toxicológicas conservadoras, os falsos negativos tendem a ser o erro de maior custo; portanto, a ocorrência de apenas 2 casos nessa categoria indica um perfil mais adequado para reduzir o risco de liberar compostos potencialmente tóxicos como "seguros", embora permaneça a necessidade de considerar ajuste de limiar e calibração probabilística conforme o objetivo aplicado (*screening* vs. confirmação).

Figura 2 — Matriz de confusão no conjunto de teste.



Fonte: Saliba *et al.*, 2026.

O detalhamento por classe (TAB.4) evidencia que a classe minoritária (0) apresentou *precision* = 0,7778, *recall* = 0,7778 e F1 = 0,7778, desempenho que contribui diretamente para a elevação da *balanced accuracy*. Para a classe majoritária (1), observou-se *precision* = 0,9608 e *recall* = 0,9608, indicando manutenção de alta capacidade de identificação de tóxicos. Ainda assim, o suporte reduzido da classe 0 (apenas 9 amostras no teste) implica maior variância amostral e exige cautela na extrapolação desses valores.

Tabela 4 — Relatório de classificação por classe aplicando a técnica SMOTENC.

Classe	precision	recall	f1-score	support
0	0,7778	0,7778	0,7778	9
1	0,9608	0,9608	0,9608	51
Macro avg	0,8693	0,8693	0,8693	60
Weighted avg	0,9333	0,9333	0,9333	60

Fonte: Saliba *et al.*, 2026.

Na validação cruzada estratificada com 5 dobras (TAB.5), as métricas médias ± desvio-padrão foram: acurácia (*test_acc*) = 0,9206 ± 0,0472; acurácia balanceada (*test_bal_acc*) = 0,8340 ± 0,1149; precisão (*test_prec*) = 0,9500 ± 0,0349; sensibilidade (*test_rec*) = 0,9580 ± 0,0295; e F1(*test_f1*) = 0,9537 ± 0,0272. Observaram-se AUC-ROC (*test_roc_auc*) = 0,8773 ± 0,1126 e AP (*test_ap*) = 0,9532 ± 0,0520, indicando desempenho médio elevado, especialmente nas métricas orientadas à classe positiva, porém com variabilidade relevante entre dobras, compatível com o tamanho limitado do conjunto de dados, o desequilíbrio de classes e a possível sensibilidade da ordenação probabilística à composição dos subconjuntos.

Tabela 5 — Validação cruzada (média ± desvio padrão) após SMOTENC.

Métrica	mean	std
test_acc	0,9206	0,0472
test_bal_acc	0,8340	0,1149
test_prec	0,9500	0,0349
test_rec	0,9580	0,0295
test_f1	0,9537	0,0272

Métrica	mean	std
test_roc_auc	0,8773	0,1126
test_ap	0,9532	0,0520

Fonte: Saliba *et al.*, 2026.

Em consonância com as recomendações de Castro e Ferreira (2022) para desfechos categóricos, modelou-se a probabilidade de toxicidade por regressão logística binária, $p_{(tóxico)} = 1/(1+\exp(-z))$, com $z = \beta_0 + \sum\beta_i X_i$. Para tornar comparáveis preditores em diferentes escalas, as variáveis numéricas foram padronizadas com a média e o desvio-padrão do conjunto de treino, resultando em $z = \beta_{0\_pad} + \sum\beta_{i\_pad} \cdot X_{i\_pad}$, onde $X_{i\_pad} = (X_i - \text{média\_treino})/\text{desvio\_treino}$; a variável categórica "Função Orgânica" não foi incluída. Nessa parametrização, β_{i_pad} representa a variação no log das chances (log-odds) associada a um aumento de 1 desvio-padrão na variável i. Para interpretação aplicada, também apresentamos coeficientes reescalados na unidade original, de modo que $\exp(\beta_{i\_unid})$ corresponda ao *odds ratio* (OR) para um incremento unitário (ou para a mudança 0⟶1 em variáveis binárias). Em regressão logística, a direção e a força da associação são expressas por OR (OR = 1 indica ausência de associação; OR > 1, associação positiva/fator de risco; OR < 1, associação negativa/fator de proteção) (CASTRO e FERREIRA, 2022).

Comparando o desempenho em teste (n = 60) (TAB.6), a regressão logística apresentou *accuracy* = 0,9167, *precision* = 0,9259, *recall* = 0,9804, F1 = 0,9524, ROC-AUC = 0,9172, PR-AUC = 0,9823 e *balanced_accuracy* = 0,7680, enquanto o *Random Forest* treinado com SMOTENC atingiu *accuracy* = 0,9333, *balanced_accuracy* = 0,8693, *precision* = 0,9608, *recall* = 0,9608, F1 = 0,9608, ROC-AUC = 0,9673, PR-AUC = 0,9941 e MCC = 0,7386 (Tab. 6). Em termos comparativos, o *Random Forest* superou a regressão logística na maioria das métricas de discriminação e desempenho global — ROC-AUC (+0,050), PR-AUC (+0,012), *accuracy* (+0,017), F1 (+0,008) e, sobretudo, *balanced_accuracy* (+0,101), refletindo maior especificidade (estimada ≈ 0,78 no RF vs. 0,56 na logística). Por outro lado, a regressão logística maximizou o *recall* (0,9804 vs. 0,9608), reduzindo falsos negativos, característica desejável em triagem toxicológica. Essas diferenças podem decorrer tanto da capacidade do algoritmo quanto da estratégia de reamostragem (SMOTENC) e do ponto de corte adotado; assim, a escolha entre modelos deve considerar o custo relativo de falsos negativos e positivos, bem como análises complementares de calibração e intervalos de confiança para julgar a significância prática dos ganhos observados.

Tabela 6 — Comparação entre as métricas do modelo *Forest Random* (SMOTENC) e do modelo Regressão Logística.

Métricas	Forest Random	Regressão Logística	Variação
accuracy	0,9333	0,9167	0,0166
balanced_accuracy	0,8693	0,768	0,1013
precision	0,9608	0,9259	0,0349
recall	0,9608	0,9804	-0,0196
F1	0,9608	0,9524	0,0084
roc_auc	0,9673	0,9172	0,0501
pr_auc	0,9941	0,9823	0,0118

Fonte: Saliba *et al.*, 2026.

Na técnica de regressão logística, a interpretação conjunta dos coeficientes do modelo evidencia elevada coerência com hipóteses químicas consolidadas para toxicidade molecular, tanto em relação à direção quanto à magnitude dos efeitos estimados (TAB. 7). A variável mais influente foi a quantidade de halogênios, que apresentou coeficiente padronizado elevado (β_{pad} = 1,561), correspondente a um OR ≈ 4,76 por aumento de um desvio padrão, indicando que moléculas mais halogenadas possuem risco substancialmente maior. Na escala original, cada halogênio adicional resultou em β = 1,360, com OR ≈ 3,90, reforçando o papel central da halogenação no aumento da toxicidade. De forma consistente, a proporção Hal/C também apresentou efeito positivo relevante (β_{pad} = 0,598; OR ≈ 1,82), indicando que a densidade relativa de halogênios na estrutura, e não apenas sua contagem absoluta, contribui para o risco estimado; na escala original, um aumento unitário nessa razão resultou em OR ≈ 1,85.

Tabela 7 — Coeficientes e *Odds Ratio* referentes à regressão logística.

variavel	beta_padronizado	OR_exp(beta_padronizado)	beta_unidade_original	OR_exp(beta_unidade_original)
Intercepto	3,0318	20,7353	0,5488	1,7312
Massa Mol.	-0,2380	0,7882	0,0000	1,0000
Ciclo (0=aberto,1=fechado)	0,2066	1,2295	0,4416	1,5553
H ligados a N	-0,4867	0,6146	-0,4230	0,6551
Qtd Halogênios	1,5610	4,7638	1,3604	3,8978
Hal/C proporção	0,5977	1,8180	0,6138	1,8475
Anéis Alif.	-0,3538	0,7020	-1,6180	0,1983

Anéis Arom.	0,6139	1,8476	1,3664	3,9212
C quirais	0,3416	1,4072	1,8346	6,2626

Fonte: Saliba *et al.*, 2026.

A presença de anéis aromáticos destacou-se igualmente como fator de risco importante, com β_pad = 0,614 (OR ≈ 1,85) e β = 1,366 na escala original (OR ≈ 3,92 por anel aromático adicional), resultado compatível com a maior estabilidade, planaridade e capacidade de interação com sistemas biológicos frequentemente associadas a estruturas aromáticas. Em contraste, os anéis alifáticos apresentaram efeito protetor, com β_pad = −0,354 (OR ≈ 0,70) e β = −1,618 na escala original (OR ≈ 0,20 por anel adicional), evidenciando diferenças estruturais relevantes entre ciclos saturados e aromáticos no contexto toxicológico.

Outros descritores estruturais apresentaram efeitos de menor magnitude, porém quimicamente plausíveis. O número de hidrogênios ligados ao nitrogênio mostrou associação negativa com o risco (β_pad = −0,487; OR ≈ 0,62), indicando que maior protonabilidade e polaridade tendem a reduzir a toxicidade estimada; na escala original, cada hidrogênio adicional ligado ao nitrogênio resultou em β = −0,423 (OR ≈ 0,66). A variável ciclo (estrutura fechada versus aberta) apresentou efeito positivo moderado (β_pad = 0,207; OR ≈ 1,23), com β = 0,442 na escala original (OR ≈ 1,56), sugerindo que a ciclicidade contribui para o risco, ainda que de forma menos pronunciada do que a halogenação ou a aromaticidade. O número de centros quirais também exibiu associação positiva (β_pad = 0,342; OR ≈ 1,41), embora sua interpretação por unidade original (β = 1,835; OR ≈ 6,26) deva ser feita com cautela devido à baixa média e ao pequeno desvio padrão da variável (DP = 0,186), tornando a análise padronizada mais estável e informativa.

A massa molecular apresentou efeito protetor discreto quando considerada por desvio padrão (β_pad = −0,238; OR ≈ 0,79), enquanto o coeficiente na escala original foi muito pequeno ($\beta = -4{,}404 \times 10^{-5}$; OR ≈ 0,99996 por dalton, aproximadamente 0,957 para um aumento de 1000 Da), comportamento esperado dada a escala dessa variável. Esse resultado sugere que o tamanho molecular isolado exerce influência limitada quando comparado a descritores estruturais e composicionais mais específicos.

Em termos de importância relativa, os valores absolutos dos coeficientes padronizados indicam clara dominância da quantidade de halogênios (|β_pad| = 1,561), seguida por anéis aromáticos (|β_pad| = 0,614) e proporção Hal/C (|β_pad| = 0,598). Esses preditores são quimicamente relacionados e potencialmente correlacionados entre si, o que pode levar ao compartilhamento de efeitos e ao aumento da incerteza dos coeficientes individuais. Apesar

disso, a estabilidade dos sinais e a coerência química dos resultados reforçam a validade do modelo, embora análises adicionais de multicolinearidade e o uso de técnicas de regularização sejam recomendáveis para maior robustez inferencial.

CONCLUSÃO

Concluiu-se que o *pipeline* proposto — construído a partir de dados do PubChem e combinando pré-processamento, reamostragem SMOTENC e modelos complementares — mostrou-se eficaz para predição de toxicidade de pequenas moléculas, com desempenho superior do *Random Forest* na maioria das métricas (*accuracy* 0,9333; *balanced_accuracy* 0,8693; ROC-AUC 0,9673; PR-AUC 0,9941; MCC 0,7386) e maior sensibilidade da regressão logística (*recall* 0,9804), perfil adequado a triagens conservadoras. A análise dos coeficientes logísticos, mesmo sem incluir "Função Orgânica", foi coerente com hipóteses químicas consolidadas, apontando maior risco associado à halogenação e à aromaticidade e efeito protetor de anéis alifáticos e de maior número de hidrogênios ligados ao nitrogênio, reforçando a utilidade do modelo para interpretação mecanística. Apesar desses avanços e do desempenho competitivo frente à literatura, o tamanho amostral reduzido, o desbalanceamento acentuado, a presença de duplicatas e o baixo suporte de várias categorias de "Função Orgânica" — além da inconsistência de contagem (169+30=199) — limitam a generalização e exigem cautela. Recomenda-se, portanto, validação externa em bases independentes, particionamento mais restritivo para mitigar vazamento por analogia, calibração probabilística e ajuste de limiar conforme o uso (*screening* vs. confirmação), além de checagens de multicolinearidade, regularização e estratégias hierárquicas para incorporar "Função Orgânica". Em conjunto, os resultados indicam que o *Random Forest* com SMOTENC é uma solução robusta e de alta capacidade discriminativa para apoiar a avaliação de risco de moléculas similares a cloroaminas e halometanos, enquanto a regressão logística agrega transparência e quantificação de efeitos, sendo uma ferramenta promissora para aplicações toxicológicas.

.REFERÊNCIAS

CAMPOS, T. C. de, & VASCONCELOS, T. C. L. de. (2021). Aplicação de algoritmos de machine learning na área farmacêutica: uma revisão. *Research, Society and Development*, 2021; 10(15): e140101522862. https://doi.org/10.33448/rsd-v10i15.22862

CASTRO, H. M., & FERREIRA, J. C. (2022). Modelos de regressão linear e logística: quando utilizá-los e como interpretá-los? Jornal Brasileiro de Pneumologia, 48(6): e20220439.

FREITAS, A. L. de SOUSA; IEKER, A. S. D.; TEIXEIRA, H. M. P.; PINHEIRO, J. M.; RINALDI, W. (2021). Aprendizado de Máquina Aplicado à Predição de Doenças Cardiometabólicas com Utilização de Indicadores Metabólicos e Comportamentais de Risco à Saúde. XII Computer on the Beach, 2021; 7–9 abr. 2021.

GEORGAKIS, N., SKLIROS, D., GAD, M. A., EFROSE, R., LABROU, N. E., FLEMETAKIS, E., & CHRONOPOULOU, E. (2017). Functional and Catalytic Characterization of the Detoxifying Enzyme Haloalkane Dehalogenase from Rhizobium leguminosarum. Protein and Peptide Letters, 2017; 24(7)

LEMES, J. A.; SOUSA, J. E. F.; PEREIRA, K. S.; LACERDA, B. F. Cardoso de; ARAÚJO, K. C. B.; PEIXOTO, J. de Castro; ROSSETO, L.P.; NAPOLITANO, H. B.; NEVES, B. J. (2019). Desenvolvimento de Modelos Computacionais para Avaliação Preditiva de Ecotoxicidade em Abelhas: Desafios Atuais. *Fronteiras: Journal of Social, Technological and Environmental Science*, 2019; 8(2):132-146. DOI: 10.21664/2238-8869.2019v8i1.p132-146.