Aplicação de Aprendizado de Máquina para Predição de Toxicidade de Anti-Hipertensivos

Bruna Rodrigues Cardoso[1], Rodrigo Bonacin[2], Mariangela Dametto[1]

bruna.cardoso@cti.gov.br, rodrigo.bonacin@cti.gov.br,
mdametto@cti.gov.br

[1]Divisão de Metodologias da Computação – DIMEC
CTI/MCTI Renato Archer – Campinas/SP

[2]Pontifícia Universidade Católica de Campinas – Campinas/SP

Abstract. *Machine learning techniques allow creating models based on chemical structures to predict attributes of interest to the pharmaceutical industry. In this study, four models were created using Linear Regression, Support Vector Machines, Multilayer Perceptron and Random Forest techniques to predict the toxicity of antihypertensive drugs based on their chemical structures. The models were created and evaluated using DrugBank, an open database with chemical structures of drugs. Preliminary results indicate high coefficients of determination and low mean errors.*

Resumo. *Com técnicas de aprendizado de máquina é possível criar modelos baseados em estruturas químicas para predizer atributos de interesse para a indústria farmacêutica. Neste trabalho, foram criados quatro modelos com as técnicas de Linear Regression, Support Vector Machines, Multilayer Perceptron e Random Forest para predizer a toxicidade de fármacos anti-hipertensivos, com base nas estruturas químicas. Os modelos criados e avaliados utilizando o DrugBank, uma base de dados aberta com estruturas químicas de fármacos. Os resultados prelininares apontam para altos coeficientes de determinação e baixos erros médios.*

1. Introdução

A Inteligência Artificial é uma tecnologia utilizada em sistemas ou máquinas que imitam a inteligência humana e por isso, está se disseminando em diversas áreas do conhecimento [1]. O Aprendizado de Máquina (ou *Machine Learning*) é um ramo da Inteligência Artificial que visa o desenvolvimento de algoritmos capazes de aprender a partir de uma base de dados [2].

O uso destas tecnologias no campo da saúde auxilia em pesquisas como na indústria farmacêutica onde tem se mostrado uma ferramenta de suma importância [3]. Na indústria farmacêutica, a Inteligência Artificial e o Aprendizado de Máquina permitem a análise de grande volumes de dados moleculares, facilitando a descoberta de novos medicamentos e a otimização de processos de desenvolvimento de fármacos. Além disso, essas tecnologias têm o potencial de reduzir custos e acelerar a introdução

de novos medicamentos no mercado, melhorando a eficácia e a eficiência dos medicamentos disponíveis para o tratamento de doenças [4].

Este trabalho utiliza quatro técnicas de aprendizado de máquina *Linear Regression* (LR), *Support Vector Machines* (SVM), *Multilayer Perceptron* (MLP), *Random Forest* (RF) para criar modelos para predição de toxicidade de anti-hipertensivos. Esses modelos foram criados e avaliados a partir de estruturas químicas de anti-hipertensivos disponíveis na base de dados DrugBank (https://go.drugbank.com/). Este estudo avalia a eficácia de cada técnica e determina quais modelos são capazes de predizer valores de LD50 (dose letal média necessária para matar 50% de um grupo de animais), a partir da estrutura molecular descrita em SMILES (*Simplified Molecular Input Line Entry System*). Sendo assim, a análise detalhada dos resultados de cada modelo permite a comparação entre desempenhos e, consequentemente, a identificação de possíveis limitações e áreas de melhoria. Espera-se instigar a busca por novas estratégias para superar as dificuldades encontradas, potencializando o uso do aprendizado de máquina como uma ferramenta na pesquisa farmacêutica.

O restante deste documento está estruturado da seguinte maneira: a seção 2 apresenta a metodologia utilizada, a seção 3 apresenta os resultados obtidos, a seção 4 faz considerações e conclusões com base nos resultados obtidos e, por fim, a seção 5 apresenta os próximos passos desta pesquisa.

2. Metodologia

Para o desenvolvimento do trabalho, foi utilizado o banco de dados constituído com informações de características sobre o grau de toxicidade (LD50, mol/kg) e estrutura química de 382 fármacos anti-hipertensivos [5].

Inicialmente, a partir de um script desenvolvido na linguagem Python, as estruturas SMILES foram convertidas em objetos moleculares utilizando a biblioteca do RDKit e posteriormente gerou-se um vetor binário de 2048 bits representando as propriedades estruturais da molécula. Estes vetores de descritores foram armazenados em um DataFrame, o qual foi utilizado em quatro tipos de técnicas de aprendizado de máquina (LR, SVM, MLP e RF) para conseguir predizer o valor da característica de toxicidade de cada fármaco.

As métricas utilizadas para avaliar o desempenho das linguagens de aprendizado de máquina foram: R^2, R^2 Ajustado, *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE) e F statistics. Foi utilizado o modelo de treinamento 80/20, onde o conjunto de dados é dividido em duas partes sendo que 80% é utilizado para treinar e 20% para testes.

Inicialmente, foi utilizado o banco de dados completo obtido no DrugBank, para os testes realizados com as técnicas de aprendizado de máquina. No entanto, após resultados não satisfatórios optou-se pela remoção de *outliers*, definidos como pontos de dados que se desviam significativamente do padrão geral [6]. No conjunto de dados

original, composto por vetores binários de 382 fármacos, identificou-se e retirou-se 20 *outliers*. Essa etapa foi de grande importância para evitar que esses valores atípicos prejudicassem a generalização dos modelos permitindo uma análise mais representativa dos padrões presente no restante do conjunto de dados.

3. Resultados

A partir dos resultados de cada técnica de aprendizado de máquina e das métricas foram construídos um gráfico e uma tabela para melhor compreensão.

Na Figura 1 é possível observar que o modelo construído utilizando a LR aplicado aos dados apresentou alto desempenho. Conforme apresenta a Tabela 1, esse modelo obteve *0,999* na métrica R^2, o que indica que o modelo é capaz de explicar *99.9%* da variabilidade dos dados, refletindo um excelente resultado entre as variáveis preditoras e as reais. Além disso, o adjusted R^2 apresentou valor de *1,000*, sugerindo que o modelo se ajusta bem aos dados e mantém a precisão das variáveis preditas. O RMSE foi de *0,012*, indicando que as previsões do modelo estão próximas dos valores reais. O MAE também apresentou valor reduzido de *0,003*, indicando que o modelo captura a tendência geral dos dados e minimiza os desvios médios entre os valores reais e preditos. Entretanto, a F statistics apresentou um valor negativo de *-899,915* o que sugere um problema relacionado a multicolinearidade entre as variáveis preditoras sugerindo a necessidade de uma investigação mais aprofundada. Isso foi observado para os resultados de todos os modelos testados neste trabalho e estão descritos a seguir.

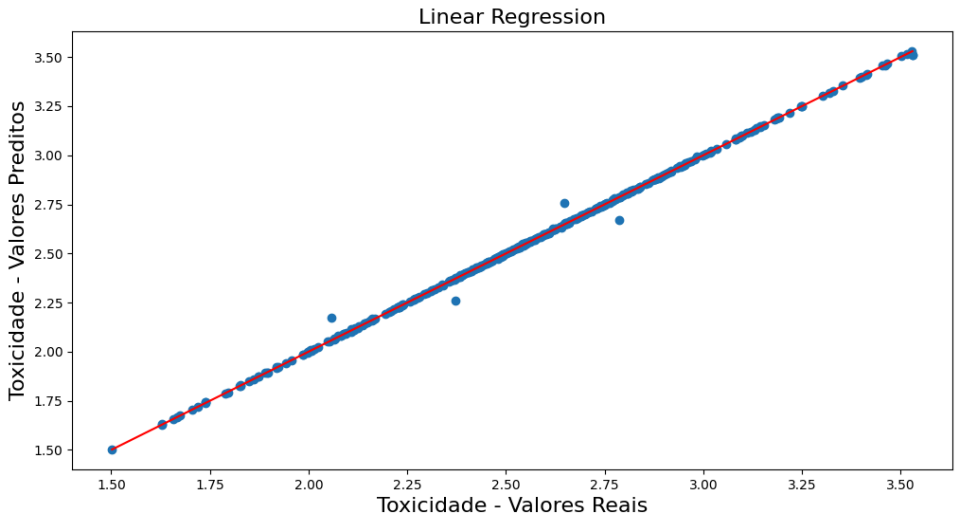

Figura 1. Gráfico de valores reais (eixo X) versus valores preditos (eixo Y) obtidos com modelo criado a partir da técnica LR

Tabela 1. Resultados das métricas para o modelo construído com LR

Linear Regression	Valores das métricas
R^2	0,9990856598610436
adjusted R^2	1,0001950014514362
RMSE	0,012147116822559723
MAE	0,00309277352203622
F statistics	-899,9149899899343

Na Figura 2 é possível observar que os resultados para o modelo criado utilizando SVM indicam bom desempenho na predição da toxicidade. Conforme apresenta a Tabela 2, o coeficiente de determinação R^2 igual a *0,925* indica que o modelo é capaz de explicar *92.5%* da variação observada nos dados. O adjusted R^2 com valor superior a *1,016* aponta robustez do modelo. O RMSE de *0,1110* e o MAE de *0,094* indicam que as previsões do modelo estão próximas aos valores reais, demonstrando precisão. Porém, o valor negativo de F statistics novamente indica multicolinearidade entre as variáveis, sugerindo investigação mais precisa, como mencionado anteriormente.

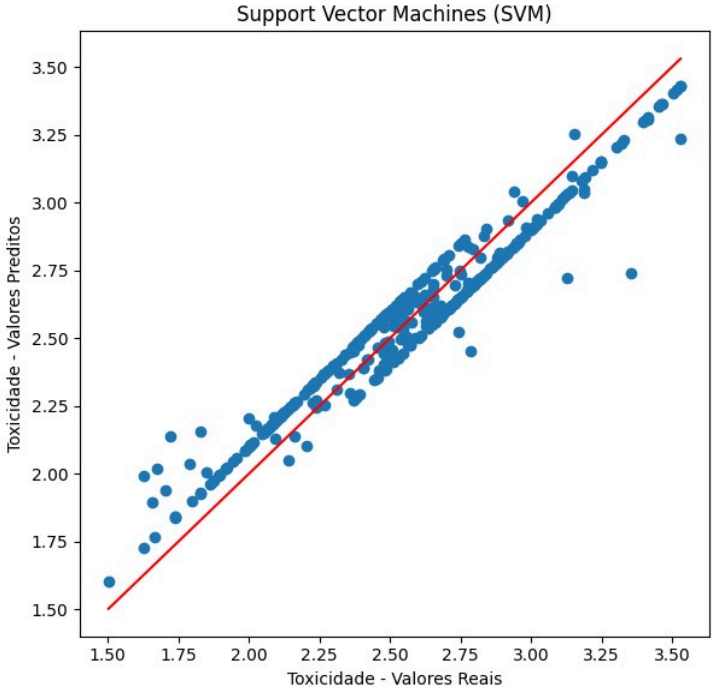

Figura 2. Gráfico de valores reais (eixo X) versus valores preditos (eixo Y) obtidos com modelo criado a partir da técnica SVM

Tabela 2. Resultados das métricas para o modelo construído com SVM

SVM	Valores das métricas
R^2	0,9245190151049666
adjusted R^2	1,0160978403804575
RMSE	0,11036670952705839
MAE	0,09443868213206504
F statistics	-10,095336037820912

Na Figura 3 é possível observar os resultados obtidos para o modelo MLP. Conforme apresenta a Tabela 3, o valor para o R^2 foi de *0,979*, indicando a explicação de *97.9%* da variabilidade nos dados de toxicidade, o que sugere excelente capacidade preditiva. Além disso, o adjusted R^2 apresentou valor de *1,004*, sugerindo que o modelo se ajusta bem aos dados e mantém a precisão das variáveis preditas. Já os valores RMSE igual a *0,058* e MAE *0,030* por serem baixos demonstram que o modelo também possui um erro de previsão relativamente pequeno e por isso as previsões são bastante próximas dos valores reais. Contudo, novamente, a F statistics mostra-se negativa com valor de *-38,423*.

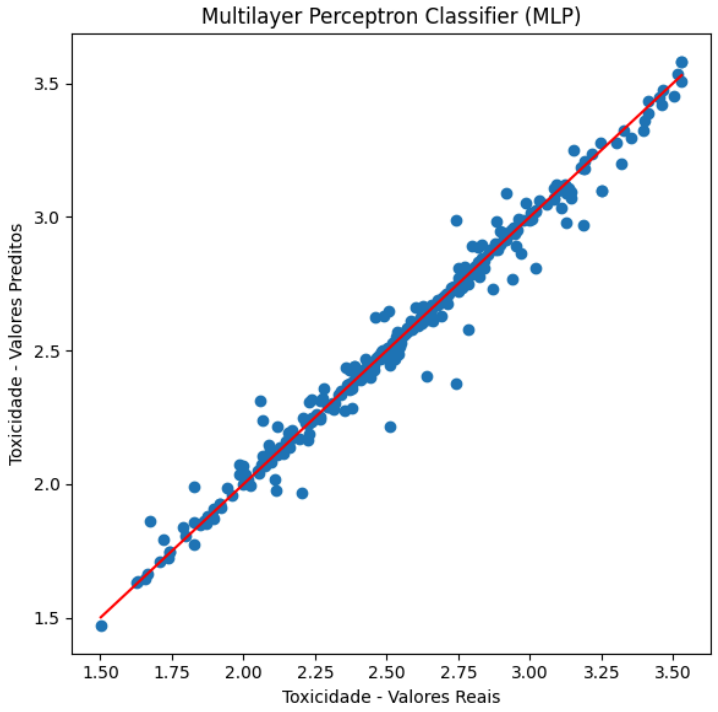

Figura 3. Gráfico de valores reais (eixo X) versus valores preditos (eixo Y) obtidos com modelo criado a partir da técnica MLP

Tabela 3. Resultados das métricas para MLP

MLP Regressor	Valores das métricas
R^2	0,978999183464529
adjusted R^2	1,004478847128418
RMSE	0,05821533567601785
MAE	0,030478962207489177
F statistics	-38,42276712828092

Na Figura 4 é possível observar os resultados obtidos com o modelo RF. não indicam uma performance tão satisfatória como os já descritos acima. Conforme apresenta a Tabela 4, o coeficiente R^2 de *0,02897* sugere que apenas *28,96%* da variabilidade nos dados pode ser explicada pelo modelo, evidenciando baixa capacidade preditiva. Além disso, o valor de *0,3282* do RMSE e *0,2430* do MAE indicam que as previsões estão distantes dos valores reais. Por fim, o resultado de *-0,3935* de F statistics indica mais uma vez que multicolinearidade dos dados, sendo necessário futuros ajustes.

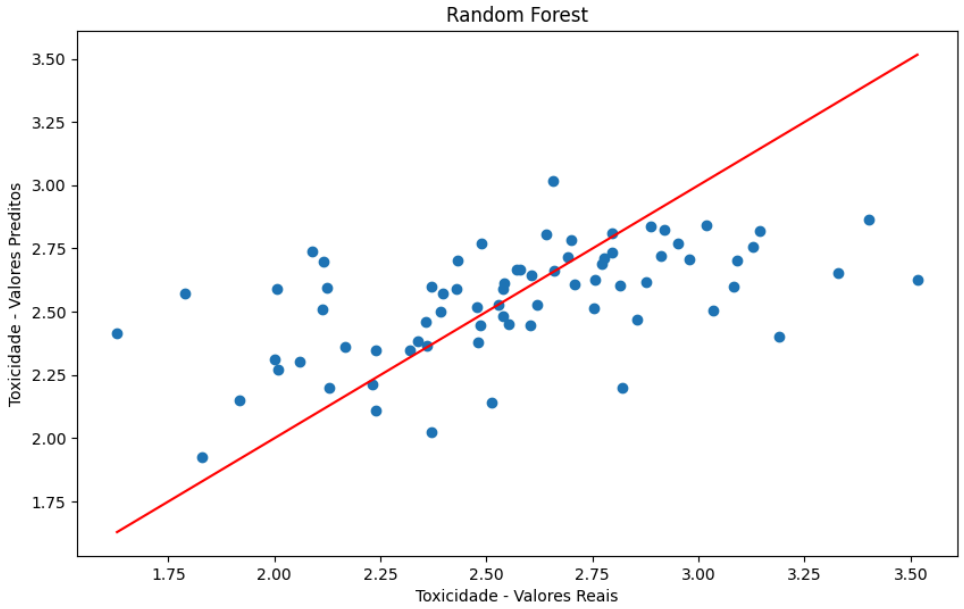

Figura 4. Gráfico de valores reais (eixo X) versus valores preditos (eixo Y) obtidos com modelo criado a partir da técnica RF

Tabela 4. Resultados das métricas para RF

Random Forest	Valores das métricas
R^2	0,2897229060297323
adjusted R^2	1,0258805418855563
RMSE	0,3281644433816679
MAE	0,24301246575342458
F statistics	-0,393560960205097

4. Considerações e Conclusões

Este estudo analisou a eficácia de diferentes modelos de aprendizagem de máquina na predição da toxicidade de fármacos anti-hipertensivos, com base nas estruturas moleculares.

Os resultados indicam que os modelos de LR e MLP se destacam em termos de precisão e capacidade preditiva, devido aos altos valores de R^2 e baixos erros médios, RMSE e MAE. Devido a isso, pode-se afirmar que esses modelos, entre os estudados, são mais adequados para realizar a predição do valor da toxicidade de fármacos a partir da estrutura química SMILES transformada em vetor binário de 2048 bits. Esses resultados são promissores, pois indicam que modelos podem ser utilizados para prever com alta precisão a toxicidade de novos compostos, o que é de extrema importância na indústria farmacêutica ao considerar a redução de riscos e custos associados ao desenvolvimento de novos medicamentos.

No entanto, valores negativos na métrica de F statistics em ambos os modelos sugere problemas de multicolinearidade, uma condição em que duas ou mais variáveis independentes em um modelo de regressão encontram-se altamente correlacionadas. Essa alta correlação pode interferir na qualidade dos resultados e dificultar a interpretação. Portanto, é necessário realizar uma investigação aprofundada.

O modelo RF apresentou desempenho mais baixo, com R^2 de *0,02897*, o que indica que foi o menos eficaz para predizer a toxicidade para o conjunto de moléculas analisadas. Isso indica que apesar de ser um modelo eficaz em muitos outros contextos, não foi o mais adequado para este tipo específico de análise, o que pode ser devido a configuração dos dados utilizados nesta pesquisa.

5. Perspectivas e trabalho futuro

Dada a importância de prever corretamente a toxicidade dos fármacos anti-hipertensivos, a próxima etapa do trabalho visa integrar técnicas de clusterização

aos modelos de aprendizado de máquina aplicados neste estudo. A finalidade de agrupar compostos com perfis estruturais e de toxicidade semelhantes. Além disso, a clusterização pode auxiliar na identificação de novos padrões e apontar outras relações entre as propriedades moleculares dos fármacos e sua toxicidade.

6. Referências

[1] Russel, S., and Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th Edition). Pearson.

[2] Jordan, M. I and Mitchell, T. M (2015). Machine learning: Trend, perspectives and prospects. Science, 349(6245), 255-260.

[3] Patel, Veer and *et al* (2022). Artificial intelligence and machine learning in drug discovery and development. Intelligent Medicine, India, p. 134-140.

[4] KOLLURI, S. *et al*. Machine learning and artificial intelligence in pharmaceutical research. Journal of Pharmaceutical Sciences, v. 111, n. 4, p. 1234-1249, 2022.

[5] Cardoso, Bruna and Bonacin, R. and Dametto, M. (2023). Análise da Toxicidade dos Fármacos Anti-Hipertensivos Relacionada às suas Estruturas Químicas. XXV Jornada de Iniciação Científica do Centro de Tecnologia da Informação Renato Archer - JICC´2023. PIBIC/CNPq/CTI - Outubro de 2023 – Campinas – São Paulo.

[6] Aggarwal, C. C. (2017). Outlier Analysis. (2nd Edition). Springer.