XVI Colóquio Técnico Científico de Saúde Única,
Ciências Agrárias e Meio Ambiente

COMPARAÇÃO ENTRE *RANDOM FOREST* E *SUPPORT VECTOR MACHINE* EM ALGORITMOS DE *QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP(QSAR)*

Vítor Ferreira Cançado[1]*, Maria Aparecida Scatamburlo Moreira[2]

[1] *Mestrando no Programa de Pós Graduação em Medicina Veterinária da UFV (PPGMV-UFV) – Universidade Federal de Viçosa– Viçosa/MG – Brasil – *Contato: vitor.cancado@ufv.br*

[2] *Docente e orientadora do Pós Graduação em Medicina Veterinária da UFV (PPGMV-UFV) – Universidade Federal de Viçosa– Viçosa/MG – Brasil*

INTRODUÇÃO

O aumento da resistência bacteriana a antimicrobianos constitui um desafio crítico para a saúde pública global, exigindo soluções rápidas e emergenciais. A Concentração Inibitória Mínima (*MIC*) é um parâmetro essencial para validar a eficácia real dos compostos antimicrobianos frente a adaptação crescente dos patógenos. Nesse contexto, os testes *in silico* que utilizam tecnologias de *Machine learning(* aprendizado de máquina) configuram-se como alternativas promissoras para a otimização do processo de triagem de novas moléculas de interesse terapêutico, permitindo a redução significativa do tempo e recursos envolvidos. O projeto busca identificar padrões entre estrutura molecular dos compostos em SMILES (*inputs*) e sua atividade biológica(*output*), no caso o valor de MIC, também conhecido como *QSAR*(Relação Estrutura-Atividade Quantitativa). Permitindo, portanto, uma análise massiva de dados(*large data bank*) e triagem de padrões moleculares de interesse.

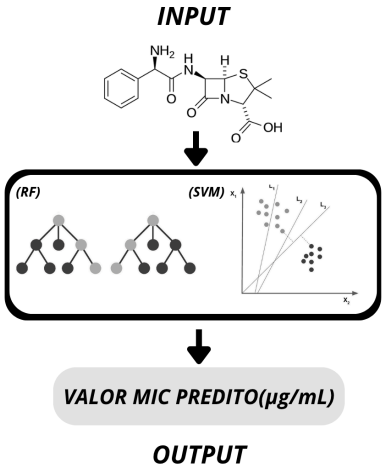

Figura 1: Esquema de processamento de dados moleculares. Entrada nos algoritmos e saída de valores, lógica válida para o treinamento e predições futuras.

Com isso almeja-se desenvolver um modelo de Inteligência Artificial que seja capaz de predizer o valor do MIC de novas moléculas, com base na identificação de farmacóforos(porção funcional da molécula que se liga a proteína *target*). Essa abordagem multidisciplinar inovadora pode se tornar uma das frentes de combate à resistência antimicrobiana, com potencial de impactar a saúde pública de forma significativa.

O objetivo geral da pesquisa é aplicar técnicas de aprendizado de máquina (machine learning) para predizer a Concentração Inibitória Mínima (MIC) de compostos antimicrobianos. Com o objetivo específico de validar e comparar diferentes modelos de regressão, no presente, testar a eficácia de duas metodologias distintas, Random forest(RF) e Support Vector Machine(SVM).

MATERIAL E MÉTODOS

A implementação foi realizada na linguagem de programação *Python 3* por meio do ambiente virtual/notebook *Google CoLab*, o qual opera em sistema de *Cloud Computing*. Essa abordagem facilita a integração entre pesquisadores parceiros. Os dados foram minerados no repositório público do *Chemical European Molecular Biology Laboratory* (*ChEMBL*), onde reúne mais de 2,5 milhões de compostos químicos e mais de 21,1 milhões de ensaios biológicos. Os dados alvo do presente estudo, incluíram o *SMILES* da molécula e o valor de MIC para a bactéria *Escherichia coli.* A escolha da *E. coli* como bactéria modelo é justificada pela extensa disponibilidade de dados sobre sua atividade biológica. Além disso, tais dados apresentam heterogeneidade significativa entre si, o que eleva a complexidade para treinamento e avaliação do modelo, uma vez validado e otimizado, a extrapolação a outras bactérias se torna fácil. Após a extração bruta de mais de 20.000 linhas de informação, foi realizada a filtragem e tabulação dos dados com auxílio do pacote *pandas*, salvando o banco de dados gerado em formato .csv(*Comma-Separated Values*) para modelagem.

A conversão do formato molecular de SMILES para *Morgan fingerprint* é realizada pelo *RDKit*, um pacote de ferramentas *open source* voltado à quimioinformática. Durante a conversão cada subgrupo molecular é representado por um bit específico, uma unidade que pode assumir dois estados, pode estar ativado na presença do subgrupo(1) ou desligado em sua ausência(0). Considerando duas possibilidades para cada um dos 2048 bits do vetor molecular, o número total de combinações possíveis resulta em 2^{2048}, permitindo na prática, exprimir a complexidade molecular em ordem infinita. O *split,* divisão, do banco de dados em *train* e *test* seguiu uma simples divisão de 20%, com a função `train_test_split`. Como o objetivo do experimento é testar a eficácia de dois modelos e por se tratar de um extenso banco de dados não se justifica o uso de *cross-validation*, como o pipeline de *K-fold,* o aumento de custo computacional neste caso não se traduz em melhores métricas.

Como descrito anteriormente foram avaliados dois algoritmos de regressão: Random Forest e Support Vector Machine. O primeiro baseia-se em uma "floresta de árvores" de tomada de decisão agregadas por votação final, oferecendo robustez a ruído, outliers e variabilidade nos dados. Em contraponto o SVM utiliza funções de núcleo (*kernel functions*) para mapear dados em espaços de maior dimensionalidade, capturando relações não lineares. *StandardScaler* foi utilizado para normalização dos dados.

A avaliação do desempenho de cada modelo se deu após a conclusão do treinamento utilizando dados da repartição de *test*. Os seguintes testes estatísticos foram realizados: Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE), Coeficiente de Determinação (R^2), e os coeficientes de correlação de Pearson e de Spearman. Os resultados foram salvos em formato tabular e representados de maneira gráfica utilizando a biblioteca Matplotlib, possibilitando uma visualização clara do desempenho relativo dos dois modelos quanto às diferentes métricas avaliadas.

RESULTADOS E DISCUSSÃO

A análise comparativa entre os modelos de *Machine learning,* revelou desempenhos consistentes na predição dos valores de MIC para *Escherichia coli.* As diferentes métricas utilizadas permitiram avaliar os modelos, treinamento e predição, sob diferentes prismas.

Os resultados obtidos pela pesquisa *in silico* demonstraram que o Random Forest apresentou um desempenho ligeiramente superior nas métricas aplicadas. O modelo atingiu valores de MAE(0.5362) e RMSE(0.7957), enquanto o SVM registrou MAE(0.5596) e RMSE(0.8420). Dados que corroboram para erros médios e desvios menores nesse modelo, sugerindo menor erro nas estimativas individuais de MIC. O valor de R^2 foi superior para o Random Forest (0.7947) em comparação ao SVM (0.7701), indicando o potencial superior em explicar os dados com precisão.

Em relação ao coeficientes de correlação, Pearson e Spearman, ambos modelos apresentaram performances elevadas e semelhantes, reforçando a coerência entre previsões e valores experimentais reais. O modelo de Random Forest obteve coeficiente de Pearson igual a 0.892 e de Spearman igual a 0.858, enquanto o SVM apresentou valores de 0.878 e 0.847, respectivamente. Correlações altas evidenciaram que, embora a diferença entre predição e valores reais seja pequena, os dois modelos testados captaram adequadamente os padrões de variação da atividade antimicrobiana ao longo da variação de estrutura molecular. Os dados mostram que sim, ambos modelos conseguiram traçar paralelos entre

estrutura molecular e atividade microbiológica. O resultado já era esperado visto a excelente capacidades de identificação de padrões em distintos dados pelos algoritmos de *Machine learning(*Tabela-1)

Tabela 1: Resultado da comparação entre *Random Forest* e *Support Vector Machine*

Modelo	MAE	RMSE	R^2	Pearson	Spearman
Random Forest	0.536	0.796	0.795	0.892	0.858
Support Vector Machine	0.560	0.842	0.770	0.878	0.847

De modo geral e simplista, sem entrar a fundo na explicação matemática de cada teste utilizado, o *Random Forest* demonstrou desempenho mais robusto e estável, possivelmente pela sua capacidade de lidar com dados discrepantes, característica comum em dados de origem microbiológica. Outro ponto que beneficiou esse modelo de regressão foi a forma com que o vetor molecular é gerado, binário, se adequando melhor a metodologia de processamento em árvore de tomada de decisão. O SVM também apresentou resultados expressivos, embora ligeiramente inferiores, entretanto não descarta sua aplicação visto que esse modelo necessita de maior calibração de hiperparâmetros e não a utilização de parâmetros em *Standard*(Figura-1). O *hyperparameter tuning* se faz necessário nos próximos passos da pesquisa, descobrindo valores ótimos para cada configuração e design, resultando em desempenho elevado

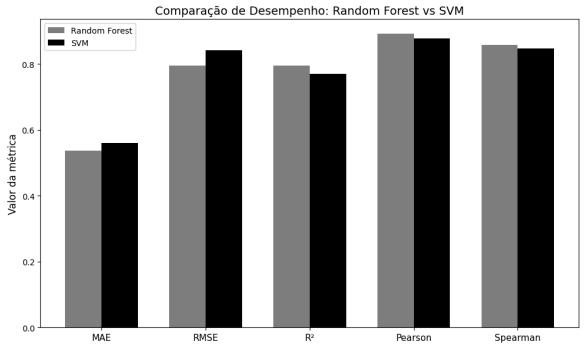

Figura 1: Gráfico comparativo de desempenho

Em suma, ambos modelos são adequados para a tarefa proposta de predição de atividade antimicrobiana a partir de descritores moleculares. Contudo, o desempenho superior do *Random Forest*, aliado à simplicidade de sua arquitetura e custo de processamento computacional reduzido, o torna uma alternativa de maior potencial para triagens virtuais e estudo de relação estrutura-atividade(SAR). Os resultados reforçam de maneira categórica a aplicabilidade de técnicas de *machine learning* em quimioinformática e *Drug Discovery*, abrindo caminho para aplicações futuras voltadas à descoberta assistida *in silico* de novos agentes antimicrobianos de interesse veterinário.

CONSIDERAÇÕES FINAIS

Os resultados indicam que o *Random Forest* apresentou bom desempenho para predição do MIC especificamente para *Escherichia coli,* validando a hipótese inicial e pavimentando o caminho das pesquisas no ramo do *Drug Discovery.* Contudo, é essencial que as predições sejam validadas e confrontadas experimentalmente em ensaios *in vitro,* permitindo além de confirmar o resultado ajustar o modelo. Melhorias como a implementação de validação cruzada, teste de outros algoritmos, otimização de hiperparâmetros tem a capacidade de potencializar os resultados e dar maior confiabilidade ao projeto

REFERÊNCIAS BIBLIOGRÁFICAS

1-BOSC, N. et al. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. **Journal of Cheminformatics**, v. 11, n. 1, p. 1–16, 10 jan. 2019.

2-DIÉGUEZ-SANTANA, K. et al. Machine Learning Study of Metabolic Networks vs ChEMBL Data of Antibacterial Compounds. **Molecular Pharmaceutics**, v. 19, n. 7, p. 2151–2163, 4 jul. 2022.

3-EKOSPUTRI, A. J. et al. Random Forest Based QSAR Model Analysis for Predicting Drug Effectiveness Against Mycobacterium Tuberculosis Bacteria. **11th International Conference on ICT for Smart Society: Integrating Data and Artificial Intelligence for a Resilient and Sustainable Future Living, ICISS 2024 - Proceeding**, 2024.

4-KROMER-EDWARDS, C. et al. Predicting Antibiotic Resistance Using Machine Learning. 2023.

5-KWON, S. et al. Comprehensive ensemble in QSAR prediction for drug discovery. **BMC Bioinformatics**, v. 20, n. 1, p. 1–12, 26 out. 2019.

7-LEE, K.; LEE, M.; KIM, D. Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server. **BMC Bioinformatics**, v. 18, n. 16, p. 75–86, 28 dez. 2017.

8-MALHEIRO, V. et al. The Potential of Artificial Intelligence in Pharmaceutical Innovation: From Drug Discovery to Clinical Trials. **Pharmaceuticals 2025, Vol. 18, Page 788**, v. 18, n. 6, p. 788, 25 maio 2025.

9-MITCHELL B.O., J. B. O. Machine learning methods in chemoinformatics. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 4, n. 5, p. 468–481, 1 set. 2014.

APOIO:

UFV Universidade Federal de Viçosa

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico