UNIVERSIDADE FEDERAL DE OURO PRETO

INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS

DEPARTAMENTO DE COMPUTAÇÃO

CAIO SILAS DE ARAUJO AMARO

ALGORITMOS SEMISSUPERVISIONADOS PARA ANÁLISE DE RELAÇÕES ENTRE ESTRUTURA QUÍMICA E ATIVIDADE BIOLÓGICA

Ouro Preto, MG
2025

CAIO SILAS DE ARAUJO AMARO

ALGORITMOS SEMISSUPERVISIONADOS PARA ANÁLISE DE RELAÇÕES ENTRE ESTRUTURA QUÍMICA E ATIVIDADE BIOLÓGICA

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Jadson Castro Gertrudes

Ouro Preto, MG
2025

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

A485a Amaro, Caio Silas de Araujo.
Algoritmos semissupervisionados para análise de relações entre estrutura química e atividade biológica. [manuscrito] / Caio Silas de Araujo Amaro. - 2025.
34 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Jadson Castro Gertrudes.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Ciência da Computação .

1. Inteligência artificial. 2. Aprendizado de máquina. 3. Descoberta de fármacos. I. Gertrudes, Jadson Castro. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.8:615.011

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Caio Silas de Araujo Amaro

Algoritmos semissupervisionados para análise de relações entre estrutura química e atividade biológica

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 26 de Agosto de 2025.

Membros da banca

Jadson Castro Gertrudes (Orientador) - Doutor - Universidade Federal de Ouro Preto
Luan Patrik Silva Pinto (Examinador) - Bacharel - Programa de Pós-Graduação em Ciência da Computação - UFOP
Hugo Eduardo Ziviani (Examinador) - Mestre - Programa de Pós-Graduação em Ciência da Computação - UFOP

Jadson Castro Gertrudes, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 26/08/2025.

Documento assinado eletronicamente por **Jadson Castro Gertrudes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 31/08/2025, às 20:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0965543** e o código CRC **9E7BD8D4**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Resumo

Este trabalho investiga o uso de algoritmos semissupervisionados para a análise de Relações Estrutura-Atividade (SAR), um campo crucial na descoberta de fármacos. A pesquisa aborda o desafio da escassez de dados rotulados, que é comum na Química Medicinal. O estudo avalia e compara o desempenho de duas técnicas semissupervisionadas de propagação de rótulos baseadas em grafos: o Gaussian Field Harmonic Function (GFHF) e o k-Nearest Neighbor Label Distribution Propagation (kNN-LDP). Além disso, a pesquisa explora a integração do algoritmo HDBScanSS* como uma etapa de pré-processamento para identificar e tratar ruídos e instâncias atípicas nos dados. A metodologia utiliza técnicas de aprendizado de máquina para otimizar a acurácia dos modelos. Os resultados demonstram que os métodos semissupervisionados, especialmente quando combinados com o HDBScanSS*, podem superar abordagens puramente supervisionadas como Random Forest e Support Vector Machines (SVM) em cenários com dados limitados. Essas descobertas reforçam o potencial desses algoritmos para a descoberta de novos compostos bioativos e para o desenvolvimento de fármacos.

Palavras-chave: Algoritmos semissupervisionados. Estrutura química. Atividade biológica. Aprendizado de máquina. Descoberta de fármacos.

Abstract

This study investigates the use of semi-supervised algorithms for the analysis of Structure-Activity Relationships (SAR), a crucial field in drug discovery. The research addresses the challenge of data scarcity, which is a common issue in Medicinal Chemistry. The work evaluates and compares the performance of two graph-based semi-supervised label propagation techniques: the Gaussian Field Harmonic Function (GFHF) and the k-Nearest Neighbor Label Distribution Propagation (kNN-LDP). Furthermore, the research explores the integration of the HDBScanSS* algorithm as a preprocessing step to identify and handle noise and outlier instances in the data. The methodology employs machine learning techniques to optimize the models' accuracy. The results demonstrate that semi-supervised methods, particularly when combined with HDBScanSS*, can outperform purely supervised approaches such as Random Forest and Support Vector Machines (SVM) in scenarios with limited data. These findings highlight the potential of these algorithms for the discovery of new bioactive compounds and for drug development.

Keywords: Semi-supervised algorithms. Chemical structure. Biological activity. Machine learning. Drug discovery.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

1.1 Introdução e Justificativa

Um campo de estudo crucial na Química Medicinal é a Relação Estrutura-Atividade (SAR), que investiga como as características moleculares de um composto influenciam suas interações com receptores biológicos(ARROIO; HONÓRIO; SILVA, 2010). O estudo do reconhecimento molecular é fundamental para o planejamento de novas entidades químicas capazes de modular alvos biológicos específicos (VERLI; BARREIRO, 2005; MONTANARI; BOLZANI, 2001). Esse processo baseia-se em interações intermoleculares entre ligantes e seus receptores proteicos (BISSANTZ; KUHN; STAHL, 2010). Desde o modelo chave-fechadura de Fischer (1894) até o conceito de encaixe induzido de Koshland (KOSHLAND, 1995), contínuos esforços têm sido dedicados à compreensão desses mecanismos (BROOIJMANS; KUNTZ, 2003).

No início, as pessoas dependiam apenas da natureza para encontrar substâncias que curassem doenças. Era uma busca limitada aos recursos naturais disponíveis. Com o passar do tempo, o conhecimento científico se desenvolveu. Isso permitiu que os pesquisadores não dependessem mais só da natureza, mas começassem a criar moléculas ativas em laboratório, de forma racional e planejada. Essa capacidade de "projetar"e sintetizar novos medicamentos foi uma revolução na medicina.(COME et al., 2019).

A incorporação de novas tecnologias, como a Inteligência Artificial (IA), tem ganhado destaque na Química Medicinal. A Organização Mundial da Saúde (OMS), em seu relatório sobre inteligência artificial na saúde, ressaltou os benefícios dessa abordagem (OPAS/OMS, 2021). Contudo, a análise de SAR permanece complexa e onerosa, com desafios como a obtenção de matérias-primas e a escassez de dados rotulados. Neste cenário, técnicas de aprendizado de máquina emergem como ferramentas promissoras para otimizar a identificação de relações SAR (GERTRUDES, 2019).

O Aprendizado de Máquina (AM), como subárea da Inteligência Artificial (IA), desenvolve métodos computacionais para aquisição autônoma de conhecimento (MITCHELL, 1997). Essas técnicas derivam padrões gerais a partir de dados específicos, construindo modelos preditivos versáteis (BARRETO, 2023). Conforme Flach (2012), o AM consiste no estudo sistemático de algoritmos que melhoram seu desempenho através da experiência.

Na Química Medicinal, os algoritmos de Aprendizado de Máquina (AM) têm se mostrado ferramentas valiosas para diversas etapas do desenvolvimento farmacêutico, desde a investigação de mecanismos patogênicos até a proposição de novos candidatos a fármacos, validação de alvos biológicos e otimização de testes *in vitro* e *in vivo* (YOUNG, 2009). Particularmente na análise de Relação Estrutura-Atividade (SAR), as técnicas de AM e mineração de dados vêm

revolucionando o processo de descoberta de medicamentos, com métodos de classificação e regressão demonstrando potencial para reduzir tanto custos quanto prazos de desenvolvimento (YOUNG, 2009; MALTAROLLO, 2013).

Entretanto, um desafio persistente limita o pleno potencial dessas abordagens: a drástica escassez de moléculas adequadamente rotuladas. Enquanto bancos de dados químicos contêm milhões de compostos, apenas uma fração ínfima, muitas vezes insignificante para o treinamento de modelos robustos, possui anotações confiáveis sobre atividade biológica – um gargalo que compromete a construção de modelos preditivos eficazes. Paradoxalmente, essa limitação coexiste com uma abundância de dados não rotulados que, se adequadamente aproveitados, poderiam potencializar o aprendizado (LEVATIĆ et al., 2013). É neste contexto que os métodos semissupervisionados surgem como possível solução, capazes de extrair conhecimento tanto dos poucos exemplos rotulados quanto das relações implícitas na massa de dados não anotados.

Nesse contexto, o aprendizado de máquina (AM) emerge como uma possível solução promissora para o estudo das Relações Estrutura-Atividade (SAR), dada sua capacidade de processar grandes conjuntos de dados de forma eficiente e identificar padrões complexos que abordagens tradicionais não conseguem (SILVERIO, 2021). Exemplos incluem a aplicação de redes neurais e máquinas de vetor de suporte no desenvolvimento de modelos Relação Quantitativa Estrutura-Atividade(QSAR), que melhoram a precisão das previsões e reduzem a complexidade computacional(MENEZES; SCOTTI; SCOTTI, 2024).

A relevância do AM torna-se ainda mais evidente em cenários onde a disparidade entre compostos caracterizados e não caracterizados é crescente, impulsionando a necessidade de métodos semissupervisionados. Este trabalho propõe uma avaliação estruturada de algoritmos de propagação de rótulos baseados em grafos para problemas SAR, hipotetizando que tais estratégias, ao explorarem a similaridade molecular através de representações em grafos, podem superar abordagens convencionais em acurácia preditiva (especialmente com poucos rótulos) e estabilidade frente a ruídos nos dados experimentais. Essa melhoria de desempenho é atribuída à habilidade desses métodos em capturar relações não lineares e hierárquicas, transformando dados não rotulados em participantes ativos do processo de aprendizado.

1.2 Objetivos

O objetivo geral deste projeto consiste em investigar o potencial de algoritmos semissupervisionados baseados em propagação de rótulos em grafos, como o Gaussian Field Harmonic Function (GFHF) e o k-Nearest Neighbor Label Distribution Propagation (kNN-LDP), bem como o método HDBScanSS* como forma de aprimorar a análise de Relação Estrutura-Atividade (SAR) em Química Medicinal, superando as limitações impostas pela escassez de dados rotulados. Para alcançar esse objetivo, foram definidos os seguintes objetivos específicos:

- Avaliar comparativamente o desempenho dos algoritmos GFHF e kNN-LDP em relação a métodos supervisionados tradicionais, como Random Forest e SVM, em tarefas de classificação SAR, utilizando a métrica de acurácia como referência.
- Investigar a sensibilidade dos métodos a parâmetros como a proporção de rótulos disponíveis (por exemplo, 5%, 10% e 15% dos dados de treinamento) e a robustez frente a ruídos nos dados experimentais.
- Implementar e avaliar o método HDBScanSS* para a detecção e tratamento de ruídos nos dados, utilizando informações parciais de rótulos para otimizar o processo de agrupamento e melhorar a qualidade da entrada dos algoritmos semissupervisionados.

1.3 Organização do Trabalho

Esta pesquisa está organizada em cinco capítulos. O Capítulo 2 apresenta a revisão bibliográfica, abordando os fundamentos teóricos dos algoritmos semissupervisionados e suas aplicações na análise da relação entre estrutura química e atividade biológica. O Capítulo 3 descreve os materiais, métodos, bases de dados, algoritmos e métricas utilizados nos experimentos. O Capítulo 4 expõe e discute os resultados obtidos, comparando criticamente o desempenho das abordagens propostas. Por fim, o Capítulo 5 reúne as conclusões do estudo, destacando as contribuições, limitações e perspectivas para pesquisas futuras.

2 Revisão Bibliográfica

Para o desenvolvimento deste trabalho, será realizada uma revisão sobre o estado da arte no contexto do aprendizado de máquina, com foco nas abordagens mais recentes e relevantes para a análise de relações entre estrutura química e atividade biológica (SAR).

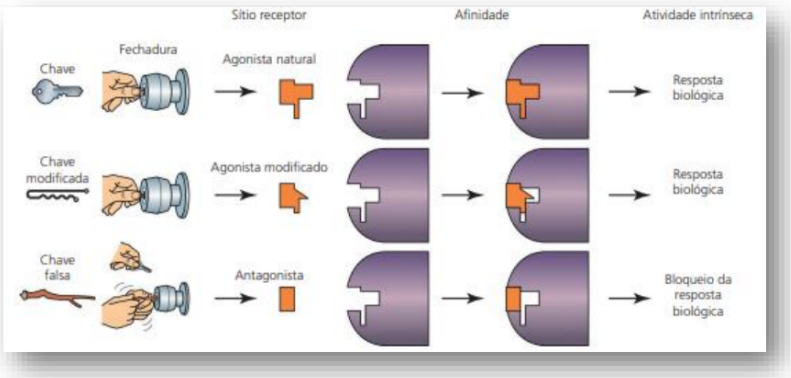
2.1 Relação Estrutura Química e Atividade Biológica

A relação entre estrutura química e atividade biológica estuda como as características moleculares de um composto influenciam suas interações com receptores biológicos. Essas interações são mediadas por forças intermoleculares, incluindo efeitos lipofílicos, polares, eletrostáticos e estéricos (ARROIO; HONÓRIO; SILVA, 2010).

Dois modelos teóricos fundamentais explicam esses mecanismos de interação molecular entre enzimas e substratos ou receptores e ligantes: o modelo chave-fechadura proposto por Fischer (1894) (Figura 2.1) e o modelo do encaixe induzido desenvolvido por Koshland (1995) (Figura 2.2). Estes modelos descrevem padrões de interação molecular entre enzimas e substratos ou receptores e ligantes.

Modelo chave-fechadura: Segundo Fischer, as enzimas possuem um sítio ativo com conformação espacial complementar ao seu substrato específico, formando um complexo enzima-substrato (RINGE; PETSKO, 2008). Nesta analogia, a enzima representa a fechadura e o substrato atua como chave, exigindo complementaridade geométrica exata para a ligação. Assim como apenas a chave correta abre determinada fechadura, somente substratos com conformação molecular específica podem se ligar ao sítio ativo de cada enzima.

Figura 2.1 – Representação do modelo chave-fechadura e do processo de reconhecimento do ligante por seu receptor biológico.



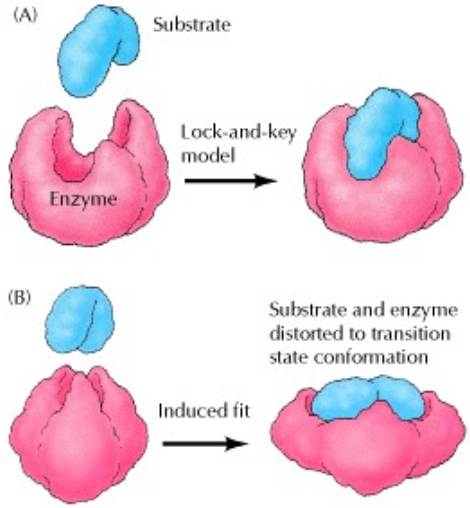
Fonte: Barreiro e Fraga (2014).

O modelo chave-fechadura pressupõe que o sítio ativo da enzima tem uma estrutura rígida, pré-formada e imutável, já adaptada ao substrato correspondente. Essa visão explica bem a alta especificidade de algumas enzimas, como a tripsina e a quimotripsina, que são capazes de distinguir seus substratos de outras moléculas semelhantes com base em características estruturais específicas.

O modelo de Fischer podia explicar as observações de que análogos menores do substrato original da enzima eram catalisados a uma taxa menor, porém não podia explicar como análogos maiores também eram catalisados. Esse foi um argumento central para que Daniel Koshland, em 1958, recebesse a aceitação de seu modelo de ligação enzimática, o "ajuste induzido" (RINGE; PETSKO, 2008). Ele não levava em consideração a flexibilidade intrínseca das moléculas biológicas e, portanto, não explicava como muitas enzimas conseguiam acomodar substratos que inicialmente não parecem se ajustar perfeitamente à sua estrutura.

Modelo encaixe induzido: Trouxe uma nova perspectiva sobre as interações entre enzimas e seus substratos, diferenciando-se do modelo chave-fechadura proposto por Fischer (1894). De acordo com Koshland (1995), uma enzima muda de forma ligeiramente quando se liga ao substrato, resultando em um encaixe ainda mais apertado. Essa flexibilidade (HAMMES; BENKOVIC; HAMMES-SCHIFFER, 2008) permite que o sítio ativo da enzima se ajuste à forma do substrato, moldando-se de maneira a facilitar a ligação e o funcionamento do complexo enzima-substrato (BERG; TYMOCZKO; STRYER, 2002).

Figura 2.2 – O modelo do encaixe induzido, introduzido por Daniel Koshland em 1958.



Fonte: The Cell: A Molecular Approach. 2nd edition.

O modelo de encaixe induzido reconhece a natureza intrinsecamente flexível das enzimas, o que o torna mais adequado para explicar a diversidade de interações moleculares observadas nos sistemas biológicos. Ele descreve como, em muitos casos, a conformação inicial da enzima é apenas parcialmente adequada ao substrato, e somente após a interação inicial ocorre um ajuste conformacional que garante a ligação mais eficiente. Esse ajuste pode influenciar não apenas a

especificidade, mas também a eficiência catalítica da enzima.

Um exemplo clássico que ilustra o modelo do encaixe induzido é o da hexoquinase. Essa enzima, ao se ligar a seu substrato, passa por mudanças conformacionais significativas, evidenciando a capacidade adaptativa proposta por Koshland. O modelo do encaixe induzido não só aprimorou a compreensão sobre os mecanismos enzimáticos, mas também forneceu uma base para o desenvolvimento de teorias ainda mais avançadas sobre as interações biomoleculares. Ele reflete melhor a complexidade dos sistemas vivos, onde a dinâmica molecular desempenha um papel crucial nas funções biológicas.

2.2 Aprendizado de Máquina

Problemas computacionais são tradicionalmente resolvidos por meio de algoritmos com passos bem definidos. No entanto, tarefas complexas como a predição de propriedades biológicas de compostos químicos apresentam desafios particulares, pois envolvem inúmeras variáveis (interações moleculares, efeitos biológicos, ruídos experimentais etc.) que dificultam a modelagem por regras explícitas (MITCHELL, 1997). Para tais casos, as técnicas de Aprendizado de Máquina (AM) têm se mostrado particularmente eficazes, como evidenciado na Figura 2.3, que ilustra as principais abordagens de AM.

O crescimento exponencial de dados em diversos setores (economia, saúde, educação etc.) e a complexidade dos problemas modernos demandam técnicas capazes de aprender padrões a partir de experiências passadas, induzir hipóteses generalizáveis e melhorar o desempenho com a exposição a novos dados (MITCHELL, 1997). Esse processo de indução de conhecimento, formalmente definido como Aprendizado de Máquina (MITCHELL, 1997; FACELI et al., 2011), pode ser categorizado conforme mostrado na Figura 2.3 em três paradigmas principais: Aprendizado supervisionado, Aprendizado não supervisionado e Aprendizado por reforço. As aplicações bem-sucedidas de AM são vastas, incluindo reconhecimento facial e de fala, detecção de fraudes em transações financeiras sistemas especialistas para diagnóstico médico, agentes inteligentes para jogos estratégicos, entre outros (FACELI et al., 2011).

Conforme será detalhado nas seções seguintes, além dos três paradigmas principais mostrados na Figura 2.3, existe ainda o aprendizado semissupervisionado, que combina características dos outros métodos.

Figura 2.3 – O aprendizado de máquina e suas principais subcategorias.

Fonte: Machine learning.

O aprendizado supervisionado é uma subcategoria do aprendizado de máquina, caracterizada pelo uso de conjuntos de dados rotulados para treinar algoritmos. Nesse tipo de aprendizado, os dados de entrada são fornecidos ao modelo junto com as respectivas saídas esperadas, permitindo que o modelo aprenda a mapear corretamente as entradas para as saídas desejadas (IBM, 2024). Cada amostra no conjunto de treinamento, E_i, é composta por um vetor de atributos $\mathbf{x}_i$ e seu rótulo correspondente y_i, representando a classe ou o valor desejado. O objetivo do aprendizado supervisionado é induzir uma função $f(\mathbf{x})$ que, dada uma entrada $\mathbf{x}$, possa prever o valor de y para novos dados não vistos.

Durante o treinamento, o modelo ajusta seus parâmetros internos para minimizar a diferença entre suas previsões e as saídas reais, utilizando uma função de perda (GOODFELLOW; BENGIO; COURVILLE, 2016). O processo de otimização continua até que o erro de previsão seja suficientemente pequeno, o que é monitorado por meio de validação cruzada e dados de teste que não foram utilizados durante o treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O objetivo final é que o modelo seja capaz de generalizar bem para exemplos desconhecidos, realizando previsões precisas em situações reais.

O aprendizado supervisionado pode ser aplicado em tarefas de classificação e regressão. Na classificação, o objetivo é atribuir dados a categorias específicas, como no caso do reconhecimento de imagens, detecção de fraudes ou diagnóstico médico (KOUROU et al., 2015). Já na regressão, o objetivo é prever valores numéricos contínuos, como a previsão de preços de imóveis ou análise de vendas de produtos (ANTIPOV; POKRYSHEVSKAYA, 2018). Essas abordagens são fundamentadas em métodos estatísticos e de otimização (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Algoritmos de aprendizado supervisionado, como árvores de decisão, k-vizinhos mais

próximos (KNN) e regressão logística, são amplamente usados para tratar esses tipos de problemas(HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A escolha do algoritmo adequado depende do tipo de dado e do problema específico. Por exemplo, a regressão linear é usada para modelar a relação entre variáveis contínuas, enquanto a regressão logística é aplicada em problemas de classificação binária (BISHOP; NASRABADI, 2006). Nas próximas subseções, apresentamos dois algoritmos supervisionados utilizados na condução dos experimentos do trabalho: a floresta aleatória (*Random forest*) e a máquina de vetores de suporte (SVM[1]).

2.2.1 Random Forest

O Random Forest, ou Floresta Aleatória, é um dos algoritmos de aprendizado de máquina mais versáteis e robustos atualmente disponíveis. Como mostra a Figura 2.4, ele se baseia na ideia simples mas poderosa de combinar múltiplas árvores de decisão para formar um modelo mais preciso e estável do que qualquer árvore individual poderia ser (RIGATTI, 2017).

A essência do Random Forest está na forma como ele cria diversidade entre suas árvores componentes. Cada árvore é treinada usando uma amostra aleatória dos dados originais, garantindo que diferentes árvores capturem aspectos distintos dos dados. Além disso, durante o processo de construção de cada árvore, apenas um subconjunto aleatório das características é considerado em cada divisão, o que evita que todas as árvores sigam o mesmo padrão e reforça a capacidade de generalização do modelo (BREIMAN, 2001).

Para problemas de classificação, o algoritmo utiliza o conceito de impureza dos nós para tomar decisões sobre como dividir os dados, enquanto em problemas de regressão ele se baseia na redução da variabilidade. O processo continua até que as árvores atinjam um tamanho adequado, balanceando complexidade e capacidade preditiva (BREIMAN et al., 2017).

Uma das grandes vantagens do Random Forest é sua capacidade intrínseca de avaliar a importância de cada variável no modelo final. Isso é feito automaticamente durante o treinamento, analisando o quanto cada característica contribui para melhorar as decisões nas árvores. Essa característica torna o modelo particularmente útil quando a interpretabilidade é importante (BREIMAN, 2001).

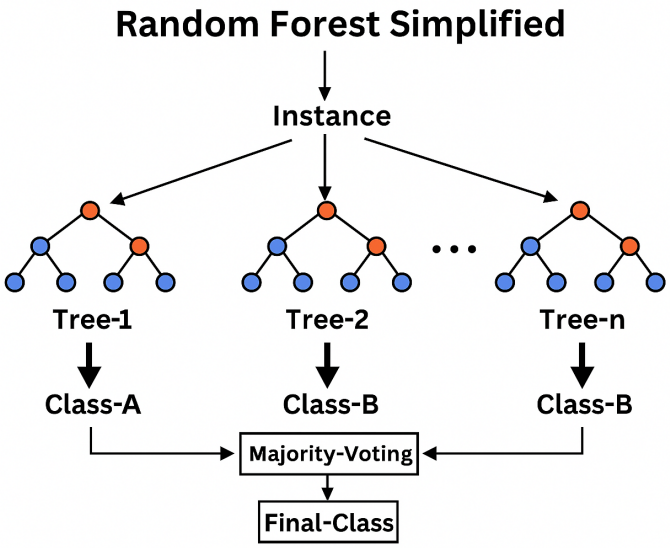
A robustez do Random Forest vem de sua natureza ensemble. Ao combinar muitas árvores, ele naturalmente reduz problemas comuns como overfitting e sensibilidade a ruídos nos dados. Além disso, ele se adapta bem a diferentes tipos de dados e situações, desde conjuntos pequenos até grandes volumes de informação, mantendo boa performance mesmo quando algumas variáveis são altamente correlacionadas (DIETTERICH, 2000).

Essa combinação de características - precisão, robustez e interpretabilidade - faz do Random Forest uma escolha popular em diversas aplicações práticas. Desde sistemas de recomendação até diagnósticos médicos, previsões financeiras e análise de sensores, o algoritmo

[1] Do original, em Inglês, *Support Vector Machines*

tem se mostrado valioso em situações onde se necessita de modelos confiáveis que possam lidar com complexidade sem perder a capacidade de explicação. A Figura 2.4 ilustra como todas essas árvores individuais trabalham em conjunto para formar previsões mais acuradas.

Figura 2.4 – Exemplo random forest.



Fonte: Elaborado pelo autor.

2.2.2 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) é um algoritmo de aprendizado supervisionado amplamente utilizado para problemas de classificação e regressão. Sua principal característica é a capacidade de encontrar um hiperplano ótimo que maximize a margem de separação entre classes distintas, garantindo assim uma generalização robusta para novos dados. (ZHU; GOLDBERG, 2009)

O princípio fundamental do SVM baseia-se na representação dos dados como pontos em um espaço vetorial de dimensão d, onde cada ponto $\mathbf{x}_i$ está associado a um rótulo y_i que indica sua classe (tipicamente $y_i \in \{-1, +1\}$ para problemas binários). O algoritmo identifica o hiperplano ótimo definido pela equação $\mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 0$, onde $\mathbf{w}$ representa o vetor normal ao hiperplano e $\mathbf{b}$ o termo de viés. A otimização busca maximizar a margem $\frac{2}{\|\mathbf{x}\|}$, que corresponde à distância entre o hiperplano e os pontos mais próximos de cada classe, denominados vetores de suporte(ZHU; GOLDBERG, 2009).

Matematicamente, este problema é formulado como uma otimização quadrática com restrições lineares:

$$\min_{\mathbf{w},\mathbf{b}} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{sujeito a} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) \geq 1, \quad \forall i$$

onde:

- $\mathbf{w}$: vetor normal ao hiperplano de separação;
- $\mathbf{b}$: termo de viés que desloca o hiperplano;
- y_i: rótulo da instância i, assumindo valores -1 ou $+1$;
- $\mathbf{x}_i$: vetor de atributos da instância i;
- $\|\mathbf{w}\|$: norma euclidiana de $\mathbf{w}$, que mede a margem entre as classes;
- Restrição $y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) \geq 1$: garante que todas as instâncias sejam classificadas corretamente, respeitando uma margem mínima de separação.

Para conjuntos de dados não linearmente separáveis, o SVM emprega funções de kernel que mapeiam os dados para um espaço de características de maior dimensão. O kernel Gaussiano (RBF - *Radial Basis Function*) é particularmente popular:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \quad \text{com} \quad \gamma = \frac{1}{2\sigma^2}$$

onde:

- $\mathbf{x}_i, \mathbf{x}_j$: vetores de atributos das instâncias i e j;
- $\|\mathbf{x}_i - \mathbf{x}_j\|^2$: distância euclidiana quadrática entre os vetores;
- γ: parâmetro do kernel que controla a sensibilidade do modelo às distâncias entre pontos;
- σ: parâmetro de escala, relacionado à largura da função Gaussiana.

onde o parâmetro γ controla a sensibilidade do kernel às distâncias entre os pontos. A classificação de novas instâncias $\mathbf{x}$ é determinada pela função de decisão:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + \mathbf{b}\right)$$

onde:

- α_i: multiplicadores de Lagrange obtidos durante a otimização do SVM;
- y_i: rótulo da instância i, assumindo valores -1 ou $+1$;
- $K(\mathbf{x}_i, \mathbf{x})$: função de kernel que mede a similaridade entre a instância de treino $\mathbf{x}_i$ e a nova instância $\mathbf{x}$;
- $\mathbf{x}_i$: vetor de atributos das instâncias de treino que atuam como vetores de suporte;

- $\mathbf{b}$: termo de viés do hiperplano;
- $\text{sign}(\cdot)$: função sinal, responsável por atribuir a classe final à instância com base no valor calculado.

onde α_i são os multiplicadores de Lagrange obtidos durante a otimização. Os pontos com $\alpha_i > 0$ correspondem aos vetores de suporte que efetivamente definem a fronteira de decisão.

Do ponto de vista geométrico, o SVM constrói uma superfície de separação que maximiza a margem entre classes, sendo particularmente eficaz em espaços de alta dimensão. A utilização de kernels permite a modelagem de fronteiras de decisão complexas, enquanto a dependência apenas dos vetores de suporte confere ao método propriedades de esparsidade e robustez computacional. Estas características tornam o SVM especialmente adequado para problemas com conjuntos de dados de média dimensão onde se busca um equilíbrio entre capacidade de generalização e complexidade do modelo. (ZHU; GOLDBERG, 2009)

2.3 Aprendizado de Máquina Semissupervisionado

O aprendizado semissupervisionado opera sobre conjuntos de dados parcialmente rotulados, onde a quantidade de exemplos não rotulados $\mathbf{X}_u$ tipicamente supera significativamente os rotulados $\mathbf{X}_l$. Entre as diversas abordagens desenvolvidas para esse contexto, os métodos transdutivos baseados em grafos se destacam por sua eficácia e fundamentação teórica.

Antes de nos aprofundarmos no modelo transdutivo baseado em grafos, cabe mencionar brevemente outras abordagens semissupervisionadas. O *self-training* é um método iterativo onde um classificador inicialmente treinado com dados rotulados é usado para rotular exemplos não rotulados com alta confiança, que são então incorporados ao conjunto de treinamento. Uma variação importante é o *co-training*, que emprega dois classificadores distintos que cooperam mutuamente no processo de rotulação (ZHU, 2005). Outras extensões incluem o *tri-training* (ZHOU; LI, 2005) (com três classificadores) e o STREET (ZHU; LAFFERTY; GHAHRAMANI, 2003), que incorpora técnicas de edição de dados.

Abordagens baseadas em agrupamento, como o SEEDED K-means, utilizam dados rotulados como sementes iniciais para os agrupamentos (BASU; BANERJEE; MOONEY, 2002). Contudo, apesar de sua utilidade em contextos específicos, esses métodos frequentemente se mostram limitados em cenários complexos que envolvem estruturas de dados não lineares. Em contraste, a abordagem baseada em grafos destaca-se por ser particularmente adequada para domínios como bioinformática, processamento de imagens e análise de redes sociais, onde a compreensão das relações entre os pontos de dados é crucial para a análise eficaz

O método de *Label Propagation* proposto por Zhu, Ghahramani e Lafferty (2003) representa a abordagem transdutiva baseada em grafos mais consolidada na literatura. Neste paradigma ocorrem algumas etapas como: (i) **construção do grafo**, onde os dados são modelados como

um grafo onde os nós representam tanto instâncias rotuladas quanto não rotuladas. As arestas conectam vizinhos e seus pesos refletem a similaridade entre as instâncias, tipicamente usando métricas como distância euclidiana ou similaridade de cosseno. (ii) **propagação de rótulos**: os rótulos se propagam iterativamente através da estrutura do grafo, onde cada nó não rotulado recebe influência dos rótulos de seus vizinhos. Esse processo pode ser formalizado como um problema de otimização que minimiza uma função de custo baseada no Laplaciano do grafo. (iii) **convergência**: O algoritmo itera até alcançar um estado de equilíbrio onde a distribuição de rótulos se estabiliza, garantindo suavidade (*smoothness*) na estrutura do grafo - pontos conectados por arestas de alto peso tendem a compartilhar o mesmo rótulo (ZHU, 2005).

Variantes avançadas como o *Gaussian Field Harmonic Function* (GFHF) e o *Robust Multi-Class Graph Transduction* (RMGT) estendem essa abordagem básica, incorporando regularização e robustez a ruídos. A principal vantagem desses métodos reside em sua capacidade de capturar estruturas complexas e não-lineares nos dados, superando muitas limitações das abordagens baseadas em classificadores individuais.

Comparado aos métodos mencionados anteriormente, a abordagem baseada em grafos oferece vantagens práticas significativas, particularmente em problemas onde a estrutura de vizinhança dos dados contém informação relevante para a tarefa de classificação. Esta característica torna os métodos baseados em grafos especialmente adequados para domínios como bioinformática, processamento de imagens e análise de redes sociais, onde as relações entre os pontos de dados são fundamentais para a análise (ZHU, 2005).

2.3.1 *Gaussian Field Harmonic Function* (GFHF)

O *Gaussian Field Harmonic Function* (GFHF) é um método fundamental no aprendizado semissupervisionado que utiliza propagação de rótulos em estruturas de grafos(ZHU; LAFFERTY; GHAHRAMANI, 2003). Este algoritmo resolve o problema de classificação encontrando uma função harmônica que interpola os rótulos conhecidos através da estrutura de similaridade dos dados (SOUSA, 2015).

O processo do GFHF desenvolve-se em quatro estágios. Na construção do grafo, os dados são representados como vértices conectados por arestas ponderadas seguindo uma métrica de similaridade. A função gaussiana é particularmente adequada para este fim:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right),$$

onde:

- $\mathbf{x}_i, \mathbf{x}_j$: vetores de atributos das instâncias i e j;
- $\|\mathbf{x}_i - \mathbf{x}_j\|^2$: distância euclidiana quadrática entre as instâncias;

- σ: parâmetro de escala que controla a influência da distância no cálculo da similaridade;
- w_{ij}: peso da aresta que conecta os nós i e j no grafo, representando o grau de similaridade entre eles.

onde o parâmetro σ determina a escala de similaridade considerada relevante. Na fase de inicialização, os vértices rotulados mantêm seus valores conhecidos, tipicamente ± 1 para problemas binários, enquanto os não rotulados recebem valores iniciais arbitrários. A propagação iterativa constitui o núcleo do algoritmo, onde cada vértice não rotulado atualiza seu valor como combinação convexa dos valores vizinhos:

$$f(\mathbf{x}_i) \leftarrow \frac{\sum_{j=1}^{l+u} w_{ij} f(\mathbf{x}_j)}{\sum_{j=1}^{l+u} w_{ij}}$$

onde:

- $\mathbf{x}_i$: instância não rotulada cujo valor será atualizado;
- l: número de instâncias rotuladas no grafo;
- u: número de instâncias não rotuladas no grafo;
- w_{ij}: peso da aresta que conecta os nós i e j, indicando o grau de similaridade entre eles;
- $f(\mathbf{x}_j)$: valor atual associado à instância j (rotulada ou propagada);
- $f(\mathbf{x}_i)$: valor atualizado da instância i, obtido como combinação convexa dos vizinhos.

Este processo converge para uma solução onde o valor em cada vértice não rotulado representa precisamente a média harmônica dos valores vizinhos. Finalmente, na etapa de decisão, os valores contínuos são discretizados para produzir as classificações finais. Do ponto de vista analítico, o GFHF minimiza o funcional de energia $\mathcal{E}(f) = \frac{1}{2}\sum_{i,j} w_{ij}(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$ sujeito às restrições dos rótulos conhecidos. Alternativamente, pode ser interpretado como a probabilidade de um passeio aleatório atingir primeiro um vértice positivamente rotulado(SOUSA, 2015).

2.3.2 *k-Nearest Neighbor Label Distribution Propagation* (KNN-LDP)

O *k-Nearest Neighbor Label Distribution Propagation* (KNN-LDP) proposto por Gøttcke, Zimek e Campello (2021) representa uma evolução conceitual importante ao propagar distribuições completas de probabilidade ao invés de valores escalares. Este método integra o *framework* Bayesiano com estimativas de densidade baseadas em vizinhanças locais.

O corpo principal do algoritmo reside na estimativa das probabilidades condicionais, dada por:

$$Pr(c|\mathbf{x}) = \frac{\sum_{\mathbf{x}' \in kNN(\mathbf{x})} Pr(c|\mathbf{x}')}{|kNN(\mathbf{x})|},$$

onde:

- c: classe considerada;
- $\mathbf{x}$: instância cuja probabilidade de pertencer à classe c está sendo estimada;
- $kNN(\mathbf{x})$: conjunto dos k vizinhos mais próximos da instância $\mathbf{x}$;
- $|\cdot|$: cardinalidade do conjunto, ou seja, o número de vizinhos considerados (k);
- $Pr(c|\mathbf{x}')$: probabilidade estimada de que o vizinho $\mathbf{x}'$ pertença à classe c;
- $Pr(c|\mathbf{x})$: probabilidade resultante de que a instância $\mathbf{x}$ pertença à classe c, obtida como a média das probabilidades de seus vizinhos.

que emerge naturalmente da combinação do teorema de Bayes com estimativas de densidade por k-NN:

$$\hat{f}(\mathbf{x}|c_j) = \frac{\sum_{\mathbf{x}' \in kNN(\mathbf{x})} Pr(c_j|\mathbf{x}')}{\sum_{\mathbf{x}' \in \mathbf{L}} Pr(c_j|\mathbf{x}') \cdot Vol_{kNN}(\mathbf{x})}.$$

onde:

- $\mathbf{x}$: instância de interesse;
- c_j: classe considerada;
- $kNN(\mathbf{x})$: conjunto dos k vizinhos mais próximos da instância $\mathbf{x}$;
- $\mathbf{L}$: conjunto de instâncias rotuladas disponíveis;
- $Pr(c_j|\mathbf{x}')$: probabilidade de a instância vizinha $\mathbf{x}'$ pertencer à classe c_j;
- $Vol_{kNN}(\mathbf{x})$: volume da região delimitada pelos k vizinhos mais próximos de $\mathbf{x}$ (normalização da densidade);
- $\hat{f}(\mathbf{x}|c_j)$: estimativa da densidade condicional da instância $\mathbf{x}$ dada a classe c_j.

A implementação eficiente do KNN-LDP baseia-se no processamento prioritário das instâncias segundo sua medida de informação:

$$w(\mathbf{x}) = \sum_{c \in \mathbf{C} \setminus \{\text{“unknown”}\}} Pr(c|\mathbf{x})$$

onde:

- $\mathbf{x}$: instância considerada;
- $\mathbf{C}$: conjunto de todas as classes possíveis no problema;
- $\mathbf{C} \setminus \{\text{“unknown”}\}$: subconjunto de classes excluindo o rótulo “desconhecido”;
- $Pr(c|\mathbf{x})$: probabilidade da instância $\mathbf{x}$ pertencer à classe c;
- $w(\mathbf{x})$: peso atribuído à instância $\mathbf{x}$, refletindo o grau de confiança na sua classificação.

Esta estratégia garante que instâncias com maior certeza sejam processadas primeiro, criando um fluxo de informação unidirecional que dispensa iterações. O método apresenta vantagens notáveis em robustez, pois a propagação de distribuições completas mitiga a propagação de erros, e em transparência, mantendo informações probabilísticas em todas as etapas. A limitação principal ocorre para instâncias isoladas na estrutura de vizinhança, onde o algoritmo prudentemente se abstém de classificar, atribuindo o rótulo “desconhecido” quando a informação disponível é insuficiente(GØTTCKE; ZIMEK; CAMPELLO, 2021).

A implementação com k-NN probabilístico mostra desempenho competitivo, particularmente em espaços de alta dimensionalidade onde relações de vizinhança capturam adequadamente a estrutura subjacente dos dados. A ausência de necessidade de iterações e a preservação de informações probabilísticas tornam o KNN-LDP particularmente adequado para aplicações onde tanto a acurácia quanto a interpretabilidade são importantes(GØTTCKE; ZIMEK; CAMPELLO, 2021).

2.3.3 *Hierarchical Density-Based Spatial Clustering of Applications with Noise – Semi-Supervised* (HDBScanSS*)

O Hierarchical DBSCAN* (HDBSCAN*), proposto por Campello, Moulavi e Sander (2013), é reconhecido como um algoritmo de agrupamento hierárquico baseado em densidade de última geração. Em sua versão estendida (CAMPELLO et al., 2015), o HDBSCAN* funciona como um *framework* capaz de realizar: agrupamento não supervisionado, agrupamento semissupervisionado por meio de restrições (do tipo *should-link* e *should-not-link*), detecção de *outliers* e visualização de dados. No contexto do aprendizado semissupervisionado, especialmente no cenário de agrupamento, para também realizar o agrupamento semissupervisionado diretamente a partir de uma coleção de objetos pré-rotulados*, em vez de depender de restrições de pares de instâncias (como era suportado anteriormente pelo algoritmo). Essa abordagem direta do uso de rótulos demonstrou ser mais simples e eficaz.

Para a extração de soluções de agrupamento "planas"(onde cada objeto é atribuído a um único *cluster*) a partir de sua hierarquia, o HDBSCAN* utiliza um método opcional de pós-processamento chamado *Framework for Optimal Extraction of Clusters* (FOSC). O FOSC é singular por permitir cortes não horizontais na hierarquia, o que significa que *clusters* podem

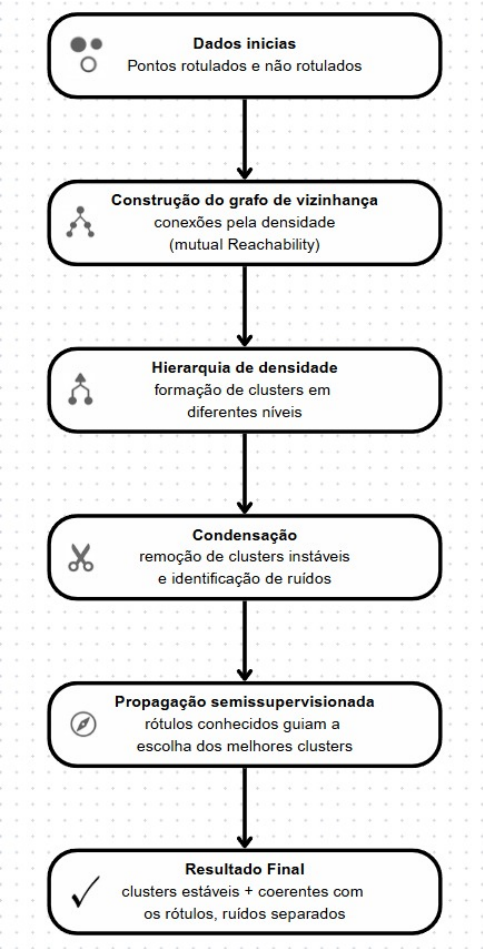
ser extraídos de diferentes níveis hierárquicos, permitindo soluções compostas por *clusters* em vários níveis de densidade.

Gertrudes (2019) introduz uma nova medida de qualidade semissupervisionada para o FOSC, que opera diretamente com rótulos, em oposição às restrições. Esta medida é baseada nos critérios B3 *Precision* e B3 *Recall*, originalmente propostos por Bagga e Baldwin (1998). Para combinar a *Precision* e o *Recall*, pode-se usar uma medida conservadora, o B3 F-Measure, que é a média harmônica de ambos. Esta medida de B3 F-Measure satisfaz as propriedades de aditividade e localidade exigidas pelo FOSC, permitindo uma decomposição eficiente para cálculo da qualidade de clusters.

2.3.3.1 Etapas do Algoritmo

O funcionamento pode ser descrito em quatro fases principais:

Figura 2.5 – Fluxo do algoritmo HDBScasSS*.



Fonte: Elaborado pelo autor.

1. **Construção do grafo de vizinhança:** A partir da distância de *mutual reachability*, define-se a conectividade entre pontos:

$$d_{mreach}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\}$$

onde:

- $\text{core}_k(x)$: distância do ponto x ao seu k-ésimo vizinho mais próximo;
- $d(a, b)$: distância euclidiana entre os pontos a e b.

2. **Formação da árvore hierárquica de densidade:** Os pontos vão sendo agrupados em diferentes níveis de densidade, formando uma hierarquia.

3. **Condensação da hierarquia:** Subárvores pouco estáveis são eliminadas, marcando pontos como ruído.

4. **Propagação semissupervisionada:** No *HDBScanSS**, os rótulos disponíveis são incorporados para guiar a escolha da partição mais consistente, reduzindo ambiguidades do HDBSCAN tradicional.

2.3.3.2 Critério de Seleção de *Clusters*

O *HDBScanSS** adapta a função de custo para considerar rótulos já conhecidos. Seja $\mathcal{L}$ o conjunto de pontos rotulados e $\mathcal{U}$ o conjunto de pontos não rotulados:

$$\text{Custo}(C) = \text{Persistência}(C) - \lambda \cdot \text{Inconsistência}(C, \mathcal{L})$$

onde:

- Persistência(C): mede a estabilidade do cluster ao longo da hierarquia;
- Inconsistência$(C, \mathcal{L})$: penaliza clusters que agrupam instâncias com rótulos conflitantes;
- λ: parâmetro de balanceamento.

Assim, os *clusters* escolhidos são estáveis e consistentes com os rótulos fornecidos.

Além disso, a nova medida baseada em rótulos pode ser combinada com a medida não supervisionada de estabilidade do **cluster** (discutida na seção 2.4 da tese), tornando a extração de *clusters* eficaz, independentemente de apenas uma parte, todas ou nenhuma das categorias estarem representadas por observações rotuladas. O HDBSCAN* e o FOSC adotam uma abordagem suave para o agrupamento semissupervisionado, tratando rótulos ou restrições como expectativas prévias do usuário, em vez de restrições rígidas que devem ser impostas. Isso permite que o algoritmo priorize suas suposições de modelo implícitas (como conectividade baseada em densidade) na construção da hierarquia, usando a informação externa como preferência para extrair a solução final.

2.4 Trabalhos Relacionados

A evolução das abordagens de aprendizado semissupervisionado (SSL) para modelagem QSAR tem sido marcada por contribuições significativas.O trabalho de Zhu (2005) estabeleceu os princípios fundamentais que ilustram o potencial das estratégias colaborativas para superar os desafios impostos pela escassez de dados rotulados em problemas complexos de classificação.Levatić et al. (2013) estabeleceram os fundamentos para aplicação de SSL na relação SAR, abordando o desafio da escassez de dados rotulados e buscando melhorar a precisão preditiva.

O estudo conduzido por Gertrudes (2019) oferece contribuições relevantes para a análise SAR em contextos semissupervisionados, servindo como importante referencial teórico-metodológico para o presente projeto de pesquisa. A investigação proposta pelo autor explorou sistematicamente três configurações experimentais distintas, combinando técnicas de propagação de rótulos, seleção de atributos e classificação via SVM. Na primeira abordagem experimental, o autor implementou um fluxo sequencial iniciando com propagação de rótulos, cujos resultados foram subsequentemente utilizados para seleção de atributos. Contudo, os resultados demonstraram que esta estratégia não proporcionou melhorias significativas no desempenho do classificador SVM quando comparado aos métodos convencionais. A segunda configuração adotou uma abordagem inversa, realizando inicialmente a seleção de atributos antes do processo de propagação de rótulos. Esta inversão na ordem das operações revelou-se mais promissora, registrando modestas melhorias de performance em relação aos métodos puramente supervisionados. A configuração mais sofisticada, apresentada como terceira abordagem por Gertrudes (2019) , combinou de forma iterativa ambas as técnicas. O processo iniciava com uma etapa de seleção de atributos, seguida pela propagação de rótulos utilizando o algoritmo SSDBSCAN++ – desenvolvido pelo autor. A etapa final consistia em recombinar os dados pré-rotulados com os rótulos propagados para uma nova fase de seleção de atributos, desta vez considerando todo o conjunto de características disponíveis. Esta abordagem híbrida demonstrou resultados particularmente satisfatórios, proporcionando melhorias substanciais no processo de aprendizado realizado pela SVM. Estes achados, particularmente os relativos à terceira configuração experimental, oferecem *insights* valiosos para o desenvolvimento de estratégias semissupervisionadas eficazes, destacando a importância da ordem e da integração adequada das diferentes etapas do processo de aprendizado.

Paralelamente, Watson, Cortes-Ciriano e Watson (2020) conduziram uma análise crítica sobre a validação de algoritmos de aprendizado de máquina na descoberta de fármacos, revelando um aspecto crucial: enquanto os modelos apresentam boa calibração próximo aos dados de treinamento, sua acurácia preditiva diminui progressivamente com o aumento da distância química, como demonstrado pelos fatores de enriquecimento calculados para 24 alvos proteicos.

Respondendo a esses desafios, Levatić et al. (2018) e Levatić et al. (2020) desenvolveram as árvores de regressão semissupervisionadas, uma abordagem específica para QSAR que visa superar as limitações impostas pela disponibilidade limitada de dados rotulados. Essa linha de pesquisa foi posteriormente expandida por Levatić et al. (2024) , que propôs um algoritmo

inovador para classificação multi-rótulo e hierárquica multi-rótulo em cenários semissupervisionados. Baseado em árvores de clusterização preditivas (PCTs). Este método utiliza tanto o espaço alvo quanto o descritivo para avaliar divisões candidatas, demonstrando em estudos empíricos um desempenho preditivo superior às PCTs supervisionadas em diversos conjuntos de dados estruturados.

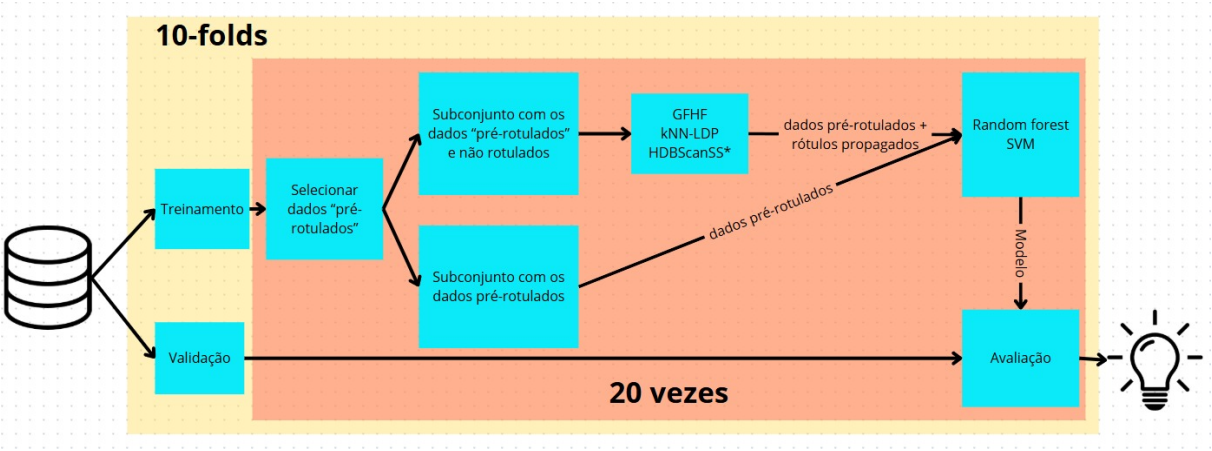
Na intersecção entre modelagem computacional e química medicinal, Nunes (2023) investigou sistematicamente derivados de isatina-tiazol, empregando a plataforma SwissADME para predição de propriedades farmacocinéticas segundo os parâmetros de Lipinski e Veber. Embora alguns compostos tenham apresentado perfis de seletividade promissores, indicando potencial segurança terapêutica, outros revelaram limitações significativas que restringiram o avanço para estudos pré-clínicos. Os desafios sintéticos associados a esses compostos, marcados por etapas complexas e reagentes especializados, ressaltam a importância de estratégias computacionais eficientes na triagem inicial de candidatos.

Complementando essas abordagens, Fontes (2023) desenvolveu um método semissupervisionado híbrido que combina as vantagens de SVM e Random Forest em um esquema de co-training. Esta arquitetura demonstrou particular eficácia no cenário comum de dados predominantemente não rotulados, alcançando ganhos de desempenho de 5% em relação aos classificadores individuais e conseguindo rotular a maioria das instâncias não rotuladas.

3 Metodologia

Neste capítulo será explicada a metodologia do estudo proposto. Serão apresentados a base de dados, os modelos utilizados para treinamento e como os mesmos funcionam e serão avaliados. Todo o processo metodológico é apresentado na Figura 3.1 e será discutido em detalhes nas próximas seções.

Figura 3.1 – Configuração experimental inicial realizada no presente projeto de pesquisa.



Fonte: Elaborado pelo autor.

3.1 Bases de Dados

Os conjuntos de dados químicos utilizados nos estudos SAR estão resumidos na Tabela 3.1, as mesmas bases utilizadas nos experimentos de Gertrudes (2019). Os conjuntos de dados "PPARD121" e "TGFB" foram obtidos mediante solicitação aos autores. Esses conjuntos já possuíam seus atributos calculados pelos próprios autores, que utilizaram o software eDragon 1.0[1]. Os demais conjuntos estão publicamente disponíveis para download. Nestes casos, as moléculas foram convertidas em atributos numéricos usando a plataforma química online Ochem.[2].

De acordo Gertrudes (2019), as moléculas de cada conjunto foram processadas na plataforma Ochem aplicando as opções padrão do software, incluindo padronização, neutralização, remoção de sais, limpeza de estruturas e otimização. Além disso, também foram calculados índices de átomo E-state, Alogps e o conjunto de descritores fornecido pela ferramenta *PyDescriptor* (MASAND; RASTIJA, 2017), também disponível na plataforma Ochem.

Os conjuntos "ACE", "BZR", "COX2", "DHFR", "PPARD121" e "TGFB" possuem atividade biológica representada por valores contínuos. Para categorizar a atividade biológica

[1] Software público disponível em http://www.vcclab.org/
[2] Software público disponível em https://ochem.eu/

no cenário de classificação semissupervisionada, foi utilizada a estratégia aplicada em Rivera-Borroto et al. (2011), onde a classe de uma molécula é definida pela mediana da atividade biológica. Moléculas com atividade biológica maior ou igual à mediana recebem a classe ativa, enquanto aquelas com atividade inferior à mediana recebem a classe inativa. Essa abordagem resulta em distribuição balanceada das classes, evitando problemas durante a avaliação.

Tabela 3.1 – Lista de conjunto de dados obtidos para execução dos experimentos no cenário semissupervisionado.

	Conjunto	**#inst**	**#atrib.**	**#ativ. biol.**
ACE	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	114	16.278	[2,10, 9,90]
ACHE	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	111	16.278	*
AT1	(SILVA, 2013)	59	1.722	*
BZR	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	163	16.288	[5,50, 8,90]
BBB	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	3.096	16.305	*
COX2	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	322	16.286	[4,00, 9,00]
DHFR	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	397	16.284	[3,30, 9,80]
ERRBA	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	925	16.298	*
EP2	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	1.344	16.291	*
FONTAINE	(FONTAINE et al., 2005)	435	16.290	*
GPB	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	66	16.280	*
GTPase	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	1.288	16.292	*
M1	(GAULTON et al., 2017)	769	16.304	*
MIC	(GAULTON et al., 2017)	219	16.271	*
PPARD121	(MALTAROLLO, 2013)	121	654	[5,95, 9,27]
TGFB	(ARAÚJO et al., 2015)	59	1.719	[4,96, 7,92]
THERM	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	76	16.274	*
THR	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	88	16.273	*
TTR	(SUTHERLAND; O'BRIEN; WEAVER, 2004)	203	16.278	*

Fonte: Adaptado de Gertrudes (2019).

3.2 Partição dos dados e informação parcial

Para melhor avaliação dos algoritmos de classificação semissupervisionada, cada conjunto foi particionado usando validação cruzada com 10 *folds*, conforme ilustrado na Figura 3.1. Nos experimentos, foram utilizados 9 folds (conjunto de treinamento) para realizar a seleção aleatória dos objetos rotulados (informação parcial), propagação dos rótulos e treinamento do classificador. O *fold* restante (conjunto de teste) foi usado para avaliar a qualidade do classificador.

Para a semissupervisão, objetos rotulados foram selecionados do conjunto de treinamento, garantindo pelo menos dois rótulos de cada classe, repetindo a seleção aleatória 20 vezes - resultando em 20 variações de cada conjunto de treinamento para cada porcentagem de objetos rotulados. Para estudar a influência da quantidade de dados rotulados, definimos os valores de 5%, 10% e 15% dos dados de treinamento.

3.3 Implementação e definição de hiperparâmetros

Para execução dos experimentos, implementações de algoritmos e configurações dos seus hiperparâmetros foram realizadas. Foram consideradas três categorias principais de métodos: (i) algoritmos semissupervisionados (kNN-LDP , GFHF e HDBScanSS*), (ii) métodos supervisionados tradicionais (*Random Forest* e SVM), (iii) abordagens híbridas que combinam propagação semissupervisionada com classificação supervisionada. A Tabela 3.2 sumariza sistematicamente estas configurações.

Tabela 3.2 – Algoritmos e hiperparâmetros utilizados nos experimentos.

Algoritmo	**Hiperparâmetros**	**Fonte**
kNN-LDP	k=3 vizinhos, heap máxima, `cKDTree`	Gøttcke, Zimek e Campello (2025)
GFHF	Kernel RBF ($\gamma = 0.1$), Laplaciano normalizado	Implementado pelo autor
SVM	Kernel RBF, $C = 1.0$, $\gamma =' scale'$	`scikit-learn`
Random Forest	100 árvores, critério='gini'	`scikit-learn`
HDBScanSS*	Classes = 5	Gertrudes (2019)

Fonte: Elaborado pelo autor.

O kNN-LDP foi implementado com uma estrutura de heap máxima para eficiência computacional na propagação de rótulos, utilizando a vizinhança definida por 3 vizinhos mais próximos calculados via `cKDTree` para acelerar as consultas espaciais. O algoritmo GFHF empregou um kernel de base radial (RBF) com parâmetro $\gamma = 0.1$ e operador Laplaciano normalizado para garantir convergência estável. Os métodos supervisionados de referência incluíram SVM com kernel RBF e parâmetros padrão, além de *Random Forest* com 100 árvores de decisão.

link para os codigos :github.com/CaioSilas

3.3.1 Identificação de Ruído com HDBScanSS*

O método HDBScanSS é empregado para lidar com instâncias de ruído, que são automaticamente rotuladas como -1 e, em seguida, desconsideradas durante a avaliação de acurácia. Trata-se de uma variação semissupervisionada do algoritmo HDBSCAN, um método de agrupamento hierárquico baseado em densidade, capaz de identificar *clusters* de forma adaptativa, sem a necessidade de definir previamente o número de grupos, além de reconhecer pontos que não pertencem a nenhum cluster, classificando-os como ruído.

Na sua versão semissupervisionada, o HDBScanSS* utiliza uma fração de rótulos previamente conhecidos para orientar o processo de *clustering*. Essa integração de informação supervisionada aprimora a separação das classes, favorece a detecção de padrões relevantes e reduz a influência de instâncias anômalas ou inconsistentes.

Ao ignorar os pontos identificados como ruído no cálculo da acurácia, o método assegura uma avaliação mais fidedigna do desempenho, resultando em modelos mais robustos e generalizáveis. Assim, sua aplicação neste trabalho tem como objetivo não apenas reduzir os efeitos do ruído, mas também potencializar a eficácia dos algoritmos subsequentes na etapa de aprendizado semissupervisionado.

Conforme detalhado na subseção 2.3.3, o HDBScanSS* adota uma abordagem semissupervisionada suave, em que os rótulos conhecidos atuam como expectativas prévias que guiam a formação de *clusters* sem impor restrições rígidas. Essa característica o torna especialmente adequado para cenários com escassez de rótulos e maior suscetibilidade a instâncias anômalas.

No fluxo metodológico representado na Figura 3.1, o HDBScanSS* se integra como uma etapa preliminar, anterior ao treinamento dos classificadores supervisionados e semissupervisionados. Sua função é identificar e descartar pontos classificados como ruído (-1), assegurando que apenas instâncias relevantes sejam consideradas nas etapas posteriores. Com isso, reduz-se a influência de exemplos atípicos e aumenta-se a consistência dos dados de entrada, reforçando a robustez do processo experimental.

3.4 Avaliação

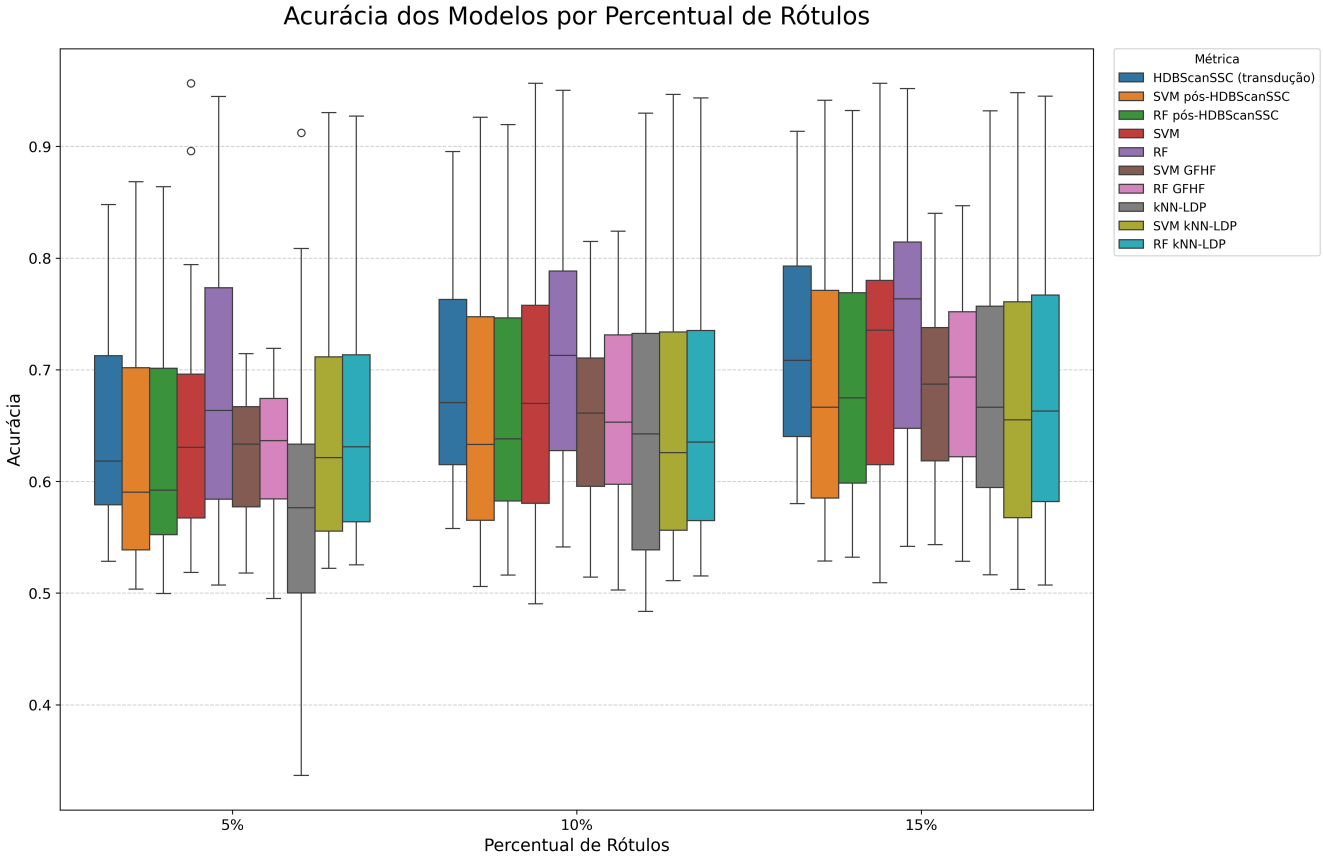
Para avaliação do desempenho dos modelos, a acurácia foi adotada como métrica. Cabe destacar que a investigação contemplou tanto abordagens supervisionadas quanto estratégias híbridas combinando métodos de propagação de rótulos com classificadores supervisionados.

Especificamente, avaliaram-se duas configurações distintas: (i) **abordagens supervisionadas**: *Random Forest* (RF) e *Support Vector Machine* (SVM), treinadas exclusivamente com a fração rotulada original do conjunto de treinamento. (ii) **abordagens híbridas**: Combinações entre os algoritmos de propagação de rótulos (GFHF e kNN-LDP) e os classificadores supervisionados, representadas pelas configurações RF-GFHF, SVM-GFHF, RF-kNN-LDP e SVM-kNN-LDP, nas quais os classificadores são treinados tanto com os rótulos originais quanto com aqueles inferidos pelos algoritmos de propagação. Esta estrutura experimental permite avaliar comparativamente o possível ganho preditivo proporcionado pela incorporação estratégica de informações semissupervisionadas.

4 Resultados

Os resultados comparativos dos algoritmos de aprendizado semissupervisionado, sumarizados na Figura 4.1, revelam padrões distintos de comportamento entre os métodos testados considerando diferentes configurações do parâmetro de vizinhança no kNN-LDP e variados percentuais de dados rotulados.

Figura 4.1 – Resultados.



Conforme evidenciado na Figura 4.1, o algoritmo SVM apresenta sensibilidade ao percentual de dados rotulados, com melhoria progressiva em estabilidade e desempenho conforme aumenta a quantidade de exemplos rotulados. A análise visual dos gráficos indica que sua performance mostra-se relativamente insensível às variações no parâmetro de vizinhança do kNN-LDP.

Os resultados apresentados na Figura 4.1 demonstram que o método *Random Forest* se destaca por sua robustez e consistência em todos os cenários avaliados. Sua abordagem baseada em *ensemble* mostra-se particularmente eficaz, com desempenho consistentemente superior ao SVM nas diversas configurações testadas.

A Figura 4.1 permite observar que as combinações que incorporaram o método GFHF

apresentam resultados competitivos, especialmente no que diz respeito à estabilidade preditiva. Nota-se que a combinação com *Random Forest* mantém bom desempenho em todos os níveis de rotulação.

Em contrapartida, como visível na Figura 4.1, as variações que utilizaram kNN-LDP mostram limitações significativas em desempenho e estabilidade. Os gráficos revelam particular sensibilidade ao parâmetro de vizinhança e dificuldade em generalizar com poucos exemplos rotulados.

A análise global dos resultados apresentados na Figura 4.1 permite concluir que abordagens baseadas em *ensemble*, particularmente o *Random Forest* em suas variações, apresentam as características mais promissoras para o problema estudado. A visualização dos dados confirma que métodos de propagação global como o GFHF podem melhorar a robustez dos modelos, enquanto estratégias baseadas em vizinhança local demonstram limitações relevantes.

Outro ponto relevante diz respeito à presença de valores *nan* nas tabelas 4.1 e 4.2, especificamente associados ao algoritmo kNN-LDP. Esses valores decorrem do fato de que, em determinadas instâncias, o método se abstém de classificar, atribuindo o rótulo *nan* quando a informação disponível na vizinhança local é insuficiente para propagar distribuições de probabilidade consistentes. Essa característica, embora confere maior prudência ao algoritmo ao evitar classificações arbitrárias, compromete sua capacidade de cobertura em cenários onde há instâncias isoladas ou mal conectadas na estrutura de vizinhança.

Tabela 4.1 – Resultados de experimentos - Parte 1

Dataset	% Rot.	HDBScanSS*	SVM pós-HDB	RF pós-HDB	SVM	RF	SVM GFHF	RF GFHF	kNN-LDP	SVM kNN-LDP	RF kNN-LDP
BBB.data	5%	0.7169	0.7145	0.7162	0.7417	**0.7665**	0.7096	0.7070	0.7288	0.7317	0.7328
	10%	0.7557	0.7493	0.7484	0.7624	**0.7861**	0.7387	0.7392	0.7530	0.7525	0.7564
	15%	0.7800	0.7658	0.7657	0.7739	**0.8003**	0.7589	0.7627	0.7615	0.7601	0.7665
EP2.data	5%	0.7767	0.7949	0.7860	**0.8958**	0.8783	0.6562	0.6487	0.8085	0.8441	0.8327
	10%	0.8039	0.8419	0.8268	0.8957	**0.8834**	0.6954	0.6832	0.8196	0.8602	0.8474
	15%	0.8180	0.8656	0.8451	**0.8960**	0.8860	0.7226	0.7062	0.8196	0.8682	0.8549
ACE.data	5%	0.6354	0.6064	0.6170	0.6388	**0.7804**	0.6615	0.6744	0.5765	0.6506	0.6547
	10%	0.7193	0.7004	0.7135	0.7297	**0.7904**	0.7094	0.7390	0.7034	0.7149	0.7135
	15%	0.7607	0.7422	0.7558	0.7863	**0.8285**	0.7523	0.7839	0.7522	0.7616	0.7674
ACHE.data	5%	0.5404	0.5259	0.5386	0.5305	**0.5735**	0.5469	0.5334	0.5124	0.5222	0.5253
	10%	0.5809	0.5512	0.5618	0.5585	**0.6108**	0.5643	0.5607	0.5396	0.5347	0.5266
	15%	0.6194	0.5680	0.5753	0.5677	**0.6211**	0.6033	0.6075	0.5962	0.5756	0.5771
AT1.data	5%	0.5823	0.5338	0.5488	0.5753	**0.5830**	0.5720	0.5728	0.4733	nan	nan
	10%	**0.6131**	0.5495	0.5590	0.5122	0.5700	0.5827	0.5803	0.4965	0.5281	0.5285
	15%	**0.6365**	0.5928	0.6020	0.5565	0.5968	0.5888	0.5970	0.5267	0.5319	0.5440
BZR.data	5%	0.5691	0.5385	0.5557	0.5344	0.5852	0.5822	**0.5959**	0.5426	0.5438	0.5466
	10%	0.6170	0.5843	0.5974	0.6022	**0.6413**	0.6087	0.6147	0.5975	0.5879	0.5895
	15%	0.6438	0.5766	0.5947	0.6504	**0.6609**	0.6333	0.6364	0.6293	0.5997	0.6188
COX2.data	5%	0.5905	0.5743	0.5783	0.6504	**0.6609**	0.6333	0.6364	0.6293	0.5997	0.6188
	10%	0.6421	0.6108	0.6246	0.6379	**0.6675**	0.6177	0.6302	0.6425	0.6257	0.6353
	15%	**0.6794**	0.6381	0.6457	0.6563	0.6890	0.6539	0.6593	0.6665	0.6550	0.6629

Tabela 4.2 – Resultados de experimentos - Parte 2

Dataset	% Rot.	HDBScanSS*	SVM pós-HDB	RF pós-HDB	SVM	RF	SVM GFHF	RF GFHF	kNN-LDP	SVM kNN-LDP	RF kNN-LDP
DHFR.data	5%	0.6135	0.5813	0.5909	0.6306	0.6635	0.6593	**0.6628**	0.5957	0.5860	0.5896
	10%	0.6705	0.6240	0.6366	0.7077	**0.7421**	0.6992	0.7025	0.6549	0.6463	0.6533
	15%	0.7085	0.6663	0.6749	0.7361	**0.7740**	0.7234	0.7290	0.6962	0.6906	0.7016
errba.data	5%	0.6282	0.6132	0.6186	0.6490	**0.7258**	0.6216	0.6090	0.6317	0.6208	0.6307
	10%	0.6827	0.6651	0.6698	0.7099	**0.7707**	0.6651	0.6610	0.6615	0.6537	0.6666
	15%	0.7149	0.6963	0.7040	0.7354	**0.7939**	0.6949	0.6934	0.6897	0.6925	0.6994
FONTAINE.data	5%	0.7588	0.7503	0.7549	0.7564	**0.8001**	0.7145	0.7192	0.6351	0.6220	0.6312
	10%	**0.8530**	0.8435	0.8481	0.8094	0.8411	0.8148	0.8240	0.7646	0.7573	0.7601
	15%	**0.8781**	0.8660	0.8672	0.8513	0.8717	0.8401	0.8469	0.8211	0.8221	0.8250
GPB.data	5%	**0.5759**	0.5293	0.5195	0.5374	0.5545	0.5336	0.5381	0.5000	nan	nan
	10%	**0.5976**	0.5593	0.5786	0.4905	0.5746	0.5344	0.5323	0.5064	0.5471	0.5495
	15%	**0.6200**	0.5630	0.5856	0.5093	0.5815	0.5704	0.5523	0.5165	0.5033	0.5072
GTPase.data	5%	0.8478	0.8684	0.8640	**0.9564**	0.9447	0.7118	0.7114	0.9122	0.9303	0.9271
	10%	0.8954	0.9262	0.9195	**0.9565**	0.9503	0.7515	0.7479	0.9297	0.9466	0.9433
	15%	0.9134	0.9412	0.9320	**0.9565**	0.9519	0.7750	0.7774	0.9318	0.9481	0.9449
M1.data	5%	0.7861	0.7857	0.7865	0.7942	**0.8246**	0.6854	0.6892	0.7802	0.7918	0.7877
	10%	0.7963	0.7906	0.7932	0.8052	**0.8465**	0.7113	0.7233	0.7999	0.8100	0.8066
	15%	0.8131	0.8038	0.8150	0.8085	**0.8629**	0.7219	0.7412	0.8137	0.8241	0.8183
MIC.data	5%	0.7080	0.6893	0.6866	0.6502	**0.7083**	0.6722	0.6740	0.6238	0.6324	0.6361
	10%	**0.7699**	0.7458	0.7445	0.7529	0.7622	0.7627	0.7630	0.7122	0.7023	0.7138
	15%	**0.8057**	0.7765	0.7724	0.7663	0.7702	0.7849	0.7870	0.7427	0.7400	0.7418
PPARD121.data	5%	0.5955	0.5837	0.5847	0.5594	0.5741	0.5955	**0.5990**	0.4535	0.5392	0.5326
	10%	**0.6582**	0.6331	0.6380	0.6295	0.6537	0.6465	0.6513	0.5790	0.5825	0.5819
	15%	**0.6826**	0.6551	0.6496	0.6547	0.6549	0.6550	0.6571	0.6125	0.6184	0.6207
TGFB.data	5%	0.6190	0.5903	0.5923	0.5990	0.5943	0.6397	**0.6400**	0.3367	nan	nan
	10%	0.6194	0.5712	0.5865	0.6072	0.6552	**0.6612**	0.6530	0.4835	0.5644	0.5723
	15%	0.6609	0.6217	0.6270	0.6330	0.6785	**0.6872**	0.6848	0.5430	0.5435	0.5471
THERM.data	5%	0.5662	0.5388	0.5450	0.5848	**0.6061**	0.5179	0.5333	0.5000	nan	nan
	10%	0.5822	0.5346	0.5483	0.5563	**0.6140**	0.5513	0.5655	0.5379	0.5479	0.5573
	15%	0.6249	0.5772	0.5784	0.5971	**0.6399**	0.5718	0.5871	0.5928	0.5596	0.5869
THR.data	5%	**0.5284**	0.5035	0.4997	0.5186	0.5072	0.5201	0.4952	0.4986	nan	nan
	10%	**0.5578**	0.5059	0.5161	0.5113	0.5413	0.5142	0.5028	0.5206	0.5112	0.5153
	15%	**0.5802**	0.5288	0.5321	0.5369	0.5418	0.5434	0.5285	0.5525	0.5271	0.5237
ttr.data	5%	0.6181	0.6008	0.6090	0.6008	**0.6749**	0.6090	0.6234	0.5389	0.5454	0.5554
	10%	0.7041	0.6687	0.6807	0.6698	**0.7129**	0.6359	0.6473	0.5951	0.5876	0.5934
	15%	0.7483	0.7209	0.7284	0.7364	**0.7635**	0.6780	0.6964	0.6648	0.6550	0.6624

A limitação observada do kNN-LDP está intimamente relacionada às propriedades dos dados químicos utilizados. Os descritores moleculares empregados apresentam alta dimensionalidade e esparsidade, o que reduz a densidade efetiva de vizinhos relevantes em torno de certos compostos. Como consequência, o kNN-LDP encontra dificuldades em propagar rótulos de forma estável, sobretudo em conjuntos pequenos e com poucos exemplos rotulados. Esse comportamento contrasta com o GFHF, cuja propagação global suaviza a influência de instâncias isoladas, e com o Random Forest, que não depende de vizinhanças explícitas.

4.1 Impacto do HDBScanSS*

No que se refere ao impacto do HDBScanSS*, a análise sugere que sua aplicação foi particularmente benéfica para os métodos semissupervisionados. Ao identificar e remover instâncias ruidosas ou inconsistentes antes da etapa de propagação de rótulos, o algoritmo atenuou as limitações do kNN-LDP, reduzindo o número de casos em que o resultado seria “desconhecido”. Além disso, o GFHF também se beneficiou do pré-processamento, uma vez que sua propagação depende da consistência estrutural dos grafos. A remoção de ruídos pelo HDBScanSS* contribuiu para a formação de grafos mais homogêneos, favorecendo a estabilidade das distribuições harmônicas.

Tabela 4.3 – Melhores resultados por dataset e proporção de rótulos (Parte 1).

Dataset	Proporção	Algoritmo	Acurácia
ACE	5%	RF	0.7804
ACE	10%	RF	0.7904
ACE	15%	RF	0.8285
ACHE	5%	RF	0.5735
ACHE	10%	RF	0.6108
ACHE	15%	RF	0.6211
AT1	5%	RF	0.5830
AT1	10%	HDBScanSS* (transdução)	0.6131
AT1	15%	HDBScanSS* (transdução)	0.6365
BBB	5%	RF	0.7665
BBB	10%	RF	0.7861
BBB	15%	RF	0.8003
BZR	5%	RF GFHF	0.5959
BZR	10%	RF	0.6413
BZR	15%	RF	0.6609
COX2	5%	RF	0.6609
COX2	10%	RF	0.6675
COX2	15%	RF	0.6890
DHFR	5%	RF	0.6635
DHFR	10%	RF	0.7421
DHFR	15%	RF	0.7740

Tabela 4.4 – Melhores resultados por dataset e proporção de rótulos (Parte 2).

Dataset	Proporção	Algoritmo	Acurácia
EP2	5%	SVM	0.8958
EP2	10%	SVM	0.8957
EP2	15%	SVM	0.8960
FONTAINE	5%	RF	0.8001
FONTAINE	10%	HDBScanSS* (transdução)	0.8530
FONTAINE	15%	HDBScanSS* (transdução)	0.8781
GPB	5%	RF	0.5545
GPB	10%	HDBScanSS* (transdução)	0.5976
GPB	15%	HDBScanSS* (transdução)	0.6200
GTPase	5%	SVM	0.9564
GTPase	10%	SVM	0.9565
GTPase	15%	SVM	0.9565
M1	5%	RF	0.8246
M1	10%	RF	0.8465
M1	15%	RF	0.8629
MIC	5%	RF	0.7083
MIC	10%	HDBScanSS* (transdução)	0.7699
MIC	15%	HDBScanSS* (transdução)	0.8057
PPARD121	5%	SVM GFHF	0.5955
PPARD121	10%	HDBScanSS* (transdução)	0.6582
PPARD121	15%	HDBScanSS* (transdução)	0.6826

Tabela 4.5 – Melhores resultados por dataset e proporção de rótulos (Parte 3).

Dataset	Proporção	Algoritmo	Acurácia
TGFB	5%	SVM GFHF	0.6397
TGFB	10%	RF	0.6552
TGFB	15%	SVM GFHF	0.6872
THERM	5%	RF	0.6061
THERM	10%	RF	0.6140
THERM	15%	HDBScanSS* (transdução)	0.6249
THR	5%	SVM GFHF	0.5201
THR	10%	HDBScanSS* (transdução)	0.5578
THR	15%	HDBScanSS* (transdução)	0.5802
TTR	5%	RF	0.6749
TTR	10%	HDBScanSS* (transdução)	0.7041
TTR	15%	RF	0.7635

Observou-se que, em diversos conjuntos (como *BBB*, *EP2* e *GTPase*), classificadores supervisionados tradicionais, em especial o SVM e o *Random Forest*, apresentaram acurácia absoluta superior. No entanto, em bases mais desafiadoras (*AT1*, *GPB*, *THERM* e *TTR*), a aplicação do HDBScanSS* contribuiu para estabilizar os resultados, reduzindo variações indesejadas e atenuando os efeitos do ruído. Esse efeito foi mais pronunciado em cenários com apenas 5% de instâncias rotuladas, nos quais a robustez adquirida pela filtragem superou a busca por ganhos brutos de acurácia.

Esses achados indicam que a integração do HDBScanSS* ao *pipeline* experimental atua como um mecanismo de pré-processamento inteligente, beneficiando classificadores mais sensíveis a ruídos e fortalecendo a confiabilidade das análises comparativas realizadas neste estudo.

5 Considerações Finais

Este trabalho teve como objetivo principal analisar as relações entre estrutura química e atividade biológica (SAR) utilizando algoritmos semissupervisionados. Para isso, foram aplicadas técnicas de aprendizado de máquina a bases de dados contendo informações sobre moléculas e suas interações com alvos biológicos. Os resultados demonstraram a eficácia dos algoritmos semissupervisionados na melhoria da precisão preditiva, especialmente em cenários com escassez de dados rotulados.

Os experimentos realizados permitiram comparar diferentes abordagens, como *Support Vector Machines* (SVM), *Random Forest* (RF) e os algoritmos de propagação de rótulos *Gaussian Field Harmonic Function* (GFHF) e *k-Nearest Neighbor Label Distribution Propagation* (kNN-LDP).

Os resultados indicaram que o Random Forest apresentou desempenho mais consistente e robusto em praticamente todos os cenários testados. Essa superioridade pode ser explicada por alguns fatores: (i) o RF é um método *ensemble*, que combina múltiplas árvores de decisão e, com isso, reduz a variância do modelo e a suscetibilidade a ruídos; (ii) sua estrutura permite capturar relações complexas entre atributos sem sobreajustar excessivamente os dados, o que é particularmente útil em contextos de escassez de rótulos; (iii) a diversidade entre as árvores gera previsões mais estáveis mesmo quando a quantidade de dados rotulados é pequena.

Por outro lado, o kNN-LDP apresentou limitações notáveis em termos de estabilidade e acurácia. Isso ocorre porque sua estratégia depende fortemente da definição de vizinhanças locais. Em conjuntos de dados de alta dimensionalidade, a noção de proximidade pode se tornar menos significativa (*curse of dimensionality*), o que compromete a propagação de rótulos. Além disso, quando existem poucos rótulos disponíveis, o algoritmo tende a propagar incertezas, resultando em classificações inconsistentes. Esse comportamento também foi agravado pela presença de instâncias isoladas, nas quais o kNN-LDP frequentemente se absteve de atribuir rótulos, retornando valores “nan”.

Já o GFHF mostrou desempenho intermediário, com vantagens em termos de suavidade global da propagação de rótulos. Por basear-se em uma formulação harmônica, esse método garante que instâncias conectadas por arestas de alto peso no grafo recebam rótulos consistentes, o que confere robustez frente a ruídos. No entanto, seu desempenho ainda depende da qualidade da construção do grafo e da escolha adequada de parâmetros como σ na função de similaridade, o que pode explicar resultados menos expressivos em alguns cenários.

De forma geral, os experimentos reforçam que métodos baseados em *ensemble* como o RF tendem a apresentar maior robustez, enquanto algoritmos de propagação local, como o kNN-LDP, são mais sensíveis à escassez de rótulos e às características da distribuição dos dados.

A análise dos modelos também evidenciou que a proporção de dados rotulados influencia diretamente a precisão dos algoritmos, reforçando a importância de abordagens semissupervisionadas para lidar com a escassez de dados anotados em pesquisas de Química Medicinal.

Com base nos resultados obtidos, pode-se concluir que os algoritmos semissupervisionados são uma alternativa viável e eficiente para a modelagem de relações SAR. O uso dessas técnicas permitiu um melhor aproveitamento dos dados não rotulados, resultando em modelos preditivos mais precisos e generalizáveis. Ademais, o *Random Forest* se destacou como uma solução promissora devido à sua robustez e estabilidade, especialmente em conjuntos de dados com poucas amostras rotuladas.

A análise dos resultados também permitiu avaliar a contribuição do HDBScanSS* como parte integrante da metodologia. Embora não tenha se mostrado consistentemente superior aos algoritmos supervisionados em termos de acurácia absoluta, seu papel de tratamento de ruído foi decisivo em diferentes cenários. Em bases com maior presença de instâncias anômalas e em situações de rotulação restrita (5%), o HDBScanSS* mitigou a influência de ruídos e promoveu maior estabilidade nos modelos subsequentes.

Dessa forma, o HDBScanSS* consolida-se como uma ferramenta metodológica complementar, cujo valor reside não apenas no desempenho isolado, mas principalmente na sua capacidade de aprimorar a robustez e a generalização dos classificadores supervisionados e semissupervisionados empregados neste trabalho.

5.1 Trabalhos Futuros

Um ponto relevante a ser considerado para trabalhos futuros diz respeito à sensibilidade dos métodos semissupervisionados em relação ao tamanho e à qualidade das bases de dados empregadas. Observou-se que, em conjuntos com número reduzido de instâncias, o desempenho dos algoritmos apresentou maior instabilidade, refletindo a dificuldade de capturar padrões consistentes quando há escassez de exemplos disponíveis. Essa limitação sugere que investigações futuras poderiam explorar estratégias que tornem os modelos mais robustos frente a cenários de baixa amostragem, como técnicas de aumento de dados, aprendizado por transferência ou integração de múltiplas bases relacionadas. Além disso, seria pertinente avaliar de forma mais sistemática a influência do ruído e de rótulos incorretos, dado que bases pequenas tendem a ser mais sensíveis a esse tipo de inconsistência. O aprofundamento nessas direções pode contribuir para ampliar a aplicabilidade prática dos algoritmos semissupervisionados no contexto da descoberta de fármacos, especialmente em situações em que os dados disponíveis são escassos e heterogêneos.

Referências

ANTIPOV, E. A.; POKRYSHEVSKAYA, E. B. Predicting real estate prices using machine learning algorithms. *Journal of Property Research*, v. 35, n. 1, p. 48–70, 2018.

ARAÚJO, S. C.; MALTAROLLO, V. G.; SILVA, D. C.; GERTRUDES, J. C.; HONÓRIO, K. M. ALK-5 Inhibition: A Molecular Interpretation of the Main Physicochemical Properties Related to Bioactive Ligands. *Journal of the Brazilian Chemical Society*, scielo, v. 26, p. 1936 – 1946, 09 2015. ISSN 0103-5053.

ARROIO, A.; HONÓRIO, K. M.; SILVA, A. B. da. Propriedades químico-quânticas empregadas em estudos das relações estrutura-atividade. *Química Nova*, SciELO Brasil, v. 33, p. 694–699, 2010.

BAGGA, A.; BALDWIN, B. Entity-based cross-document coreferencing using the vector space model. In: *COLING 1998 Volume 1: The 17th international conference on computational linguistics*. [S.l.: s.n.], 1998.

BARREIRO, E. J.; FRAGA, C. A. M. *Química Medicinal-: As bases moleculares da ação dos fármacos*. [S.l.]: Artmed Editora, 2014.

BARRETO, C. A. d. S. Seleção e rotulagem de instâncias para métodos semissupervisionados indutivos. Universidade Federal do Rio Grande do Norte, 2023.

BASU, S.; BANERJEE, A.; MOONEY, R. Semi-supervised clustering by seeding. In: *Proceedings of the 19th International Conference on Machine Learning (ICML)*. [s.n.], 2002. p. 19–26. Disponível em: <https://www.cs.utexas.edu/~mooney/papers/basu-icml02.pdf>.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. *Biochemistry*. 5th. ed. [S.l.]: W. H. Freeman, 2002. Explica detalhadamente os modelos chave-fechadura e ajuste induzido.

BISHOP, C. M.; NASRABADI, N. M. *Pattern Recognition and Machine Learning*. Springer, 2006. v. 1. ISBN 978-0-387-31073-2. Disponível em: <http://www.library.wisc.edu/selectedtocs/bg0137.pdf>.

BISSANTZ, C.; KUHN, B.; STAHL, M. A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry*, ACS Publications, v. 53, n. 14, p. 5061–5084, 2010.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A.; STONE, C. J. *Classification and regression trees*. [S.l.]: Routledge, 2017.

BROOIJMANS, N.; KUNTZ, I. D. Molecular recognition and docking algorithms. *Annual review of biophysics and biomolecular structure*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 32, n. 1, p. 335–373, 2003.

CAMPELLO, R. J.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: SPRINGER. *Pacific-Asia conference on knowledge discovery and data mining*. [S.l.], 2013. p. 160–172.

CAMPELLO, R. J.; MOULAVI, D.; ZIMEK, A.; SANDER, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, ACM New York, NY, USA, v. 10, n. 1, p. 1–51, 2015.

COME, J. A. A. d. S. et al. *Prospecção de novas moléculas naturais e sintéticas na inibição in vitro da arginase recombinante de Leishmania (Leishmania) amazonensis*. Tese (Doutorado) — Universidade de São Paulo, 2019.

DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. *International workshop on multiple classifier systems*. [S.l.], 2000. p. 1–15.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011.

FISCHER, E. Einfluss der configuration auf die wirkung der enzyme. *Berichte der Deutschen Chemischen Gesellschaft*, v. 27, p. 2985–2993, 1894. Artigo original propondo o modelo chave-fechadura.

FLACH, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. [S.l.]: Cambridge University Press, 2012.

FONTAINE, F.; PASTOR, M.; ZAMORA, I.; SANZ, F. Anchor-grind: Filling the gap between standard 3d qsar and the grid-independent descriptors. *Journal of Medicinal Chemistry*, v. 48, n. 7, p. 2687–2694, 2005. PMID: 15801859. Disponível em: <https://doi.org/10.1021/jm049113+>.

FONTES, A. F. d. C. *Algoritmo de aprendizagem semi-supervisionada*. Tese (Doutorado) — Instituto Politecnico do Porto (Portugal), 2023.

GAULTON, A.; HERSEY, A.; NOWOTKA, M.; BENTO, A. P.; CHAMBERS, J.; MENDEZ, D.; MUTOWO-MEULLENET, P.; ATKINSON, F.; BELLIS, L. J.; CIBRIÁN-UHALTE, E.; DAVIES, M.; DEDMAN, N.; KARLSSON, A.; MAGARIÑOS, M. P.; OVERINGTON, J. P.; PAPADATOS, G.; SMIT, I.; LEACH, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.*, v. 45, n. Database-Issue, p. D945–D954, 2017.

GERTRUDES, J. C. *Semi-supervised learning approaches with applications in medicinal chemistry*. Tese (Doutorado) — Universidade de São Paulo, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. MIT Press, 2016. Disponível em: <https://www.deeplearningbook.org/>.

GØTTCKE, J. M. N.; ZIMEK, A.; CAMPELLO, R. J. G. B. Non-parametric semi-supervised learning by bayesian label distribution propagation. In: *SISAP*. [S.l.]: Springer, 2021. (Lecture Notes in Computer Science, v. 13058), p. 118–132.

GØTTCKE, J. M. N.; ZIMEK, A.; CAMPELLO, R. J. G. B. Bayesian label distribution propagation: A semi-supervised probabilistic k nearest neighbor classifier. *Inf. Syst.*, v. 129, p. 102507, 2025. Disponível em: <https://doi.org/10.1016/j.is.2024.102507>.

HAMMES, G. G.; BENKOVIC, S. J.; HAMMES-SCHIFFER, S. Flexibility, diversity, and cooperativity: Pillars of enzyme catalysis. *Biochemistry*, v. 47, n. 9, p. 3317–3321, 2008.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. ed. Springer, 2009. ISBN 978-0-387-84857-0. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-84858-7>.

IBM. *O que é aprendizado supervisionado?* 2024. <https://www.ibm.com/br-pt/topics/supervised-learning>. Accessed: 2024-12-3.

KOSHLAND, D. Angew. chemie int. ed. *English*, v. 33, p. 2375, 1995.

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, v. 13, p. 8–17, 2015.

LEVATIĆ, J.; CECI, M.; KOCEV, D.; DŽEROSKI, S. Semi-supervised predictive clustering trees for (hierarchical) multi-label classification. *International Journal of Intelligent Systems*, Wiley Online Library, v. 2024, n. 1, p. 5610291, 2024.

LEVATIĆ, J.; CECI, M.; STEPIŠNIK, T.; DŽEROSKI, S.; KOCEV, D. Semi-supervised regression trees with application to qsar modelling. *Expert Systems with Applications*, v. 158, p. 113569, 2020. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417420303936>.

LEVATIĆ, J.; DZEROSKI, S.; SUPEK, F.; SMUC, T. Semi-supervised learning for quantitative structure-activity modeling. *Informatica (Slovenia)*, v. 37, n. 2, p. 173–179, 2013. Disponível em: <http://www.informatica.si/index.php/informatica/article/view/447>.

LEVATIĆ, J.; KOCEV, D.; CECI, M.; DŽEROSKI, S. Semi-supervised trees for multi-target regression. *Information Sciences*, v. 450, p. 109–127, 2018. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S002002551830210X>.

MALTAROLLO, V. G. *Aplicação de estratégias in silico para o desenvolvimento de ligantes com afinidade pelo receptor PPARδ*. Tese (Doutorado) — Universidade Federal do ABC, 2013.

MASAND, V. H.; RASTIJA, V. Pydescriptor: A new pymol plugin for calculating thousands of easily understandable molecular descriptors. *Chemometrics and Intelligent Laboratory Systems*, v. 169, p. 12 – 18, 2017. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016974391730312X>.

MENEZES, R. P. d.; SCOTTI, L.; SCOTTI, M. T. Aprendizado de máquina aplicado a qsar. *Química Nova*, SciELO Brasil, v. 47, n. 7, p. e–20240024, 2024.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Education, 1997.

MONTANARI, C. A.; BOLZANI, V. d. S. Planejamento racional de fármacos baseado em produtos naturais. *Química nova*, SciELO Brasil, v. 24, p. 105–111, 2001.

NUNES, J. S. *Planejamento estrutural, síntese e avaliação da atividade biológica de tiazóis derivados da isatina*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2023.

OPAS/OMS. *OMS publica primeiro relatório global sobre inteligência artificial na saúde e seis princípios orientadores para sua concepção e uso*. 2021. Acessado em: 22 jan. 2025. Disponível em: <https://www.paho.org/pt/noticias/28-6-2021-oms-publica-primeiro-relatorio-global-sobre-inteligencia-artificial-na-saude-e>.

RIGATTI, S. J. Random forest. *Journal of Insurance Medicine*, American Academy of Insurance Medicine 1700 Magnavox Way, Fort Wayne, IN 46804, v. 47, n. 1, p. 31–39, 2017.

RINGE, D.; PETSKO, G. A. How enzymes work. *Science*, American Association for the Advancement of Science, v. 320, n. 5882, p. 1428–1429, 2008.

RIVERA-BORROTO, O. M.; PONCE, Y. M.; VEGA, J. M. Garcia-de-la; GRAU-ÁLBALO, R. d. C. Comparison of combinatorial clustering methods on pharmacological data sets represented by machine learning–selected real molecular descriptors. *J. Chem. Inf. Model.*, v. 51, p. 3036–3049, 2011.

SILVA, D. C. *Estudos de modelagem molecular para uma série de ligantes do receptor tipo 1 da antiotensina II com atividade anti-hipertensiva*. Tese (Doutorado) — Universidade Federal do ABC, 2013.

SILVERIO, P. S. d. S. N. 3d-qsarpy: Combinando estratégias de seleção de atributos e técnicas de aprendizado de máquina para construir modelos qsar 3d. Universidade Federal do Rio Grande do Norte, 2021.

SOUSA, C. A. de. An overview on the gaussian fields and harmonic functions method for semi-supervised learning. In: IEEE. *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2015. p. 1–8.

SUTHERLAND, J. J.; O'BRIEN, L. A.; WEAVER, D. F. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med.Chem.*, v. 47, n. 22, p. 5541–5554, 2004.

VERLI, H.; BARREIRO, E. J. Um paradigma da química medicinal: a flexibilidade dos ligantes e receptores. *Química nova*, SciELO Brasil, v. 28, p. 95–102, 2005.

WATSON, O.; CORTES-CIRIANO, I.; WATSON, J. A. A semi-supervised learning framework for quantitative structure–activity regression modelling. *Bioinformatics*, v. 37, n. 3, p. 342–350, 08 2020. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btaa711>.

YOUNG, D. C. *Computational drug design - A Guide for Computational and Medicinal Chemists*. [S.l.]: Wiley, 2009.

ZHOU, Z.-H.; LI, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, IEEE, v. 17, n. 11, p. 1529–1541, 2005.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In: *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*. [s.n.], 2003. Accessed: 2025-06-12. Disponível em: <https://aaai.org/papers/icml03-118-semi-supervised-learning-using-gaussian-fields-and-harmonic-functions/>.

ZHU, X.; GOLDBERG, A. B. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers, 2009. (Synthesis Lectures on Artificial Intelligence and Machine Learning). Disponível em: <http://dx.doi.org/10.2200/S00196ED1V01Y200906AIM006>.

ZHU, X.; LAFFERTY, J.; GHAHRAMANI, Z. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*. [S.l.: s.n.], 2003. v. 3, p. 58–65.

ZHU, X. J. Semi-supervised learning literature survey. *MINDS UW Madison*, University of Wisconsin-Madison Department of Computer Sciences, 2005.