UNIVERSIDADE EVANGÉLICA DE GOIÁS - UniEVANGÉLICA
PROGRAMA DE PÓS-GRADUAÇÃO EM SOCIEDADE, TECNOLOGIA E MEIO AMBIENTE

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE TOXICIDADE EM *Artemia salina* LEACH: SUBSÍDIOS À CONSERVAÇÃO DA BIODIVERSIDADE

JOSIEL ARAUJO LEMES

ANÁPOLIS-GO

2022

JOSIEL ARAUJO LEMES

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE TOXICIDADE EM *Artemia salina* LEACH**: SUBSÍDIOS À CONSERVAÇÃO DA BIODIVERSIDADE**

Dissertação apresentada ao curso de Mestrado do Programa de Pós-Graduação em Sociedade, Tecnologia e Meio Ambiente (PPSTMA) da Universidade Evangélica de Goiás - UniEVANGÉLICA como requisito parcial para a obtenção do título de Mestre em Ciências Ambientais.

Orientadora: Profª. Drª. Josana de Castro Peixoto

Co-orientador: Prof. Dr. Bruno Júnior Neves

ANÁPOLIS-GO

2022

L552

Lemes, Josiel Araujo.

Desenvolvimento de modelos de aprendizado de máquina para predição de toxidade em *Artemia salina* LEACH: subsídios à conservação da biodiversidade / Josiel Araujo Lemes - Anápolis: Universidade Evangélica de Goiás, 2022.

42 p.; il.

Orientadora: Profª. Dra. Josana de Castro Peixoto.
Co-orientador: Prof. Dr. Bruno Junior Neves.
Dissertação (mestrado) – Programa de pós-graduação em Sociedade, Tecnologia e Meio Ambiente – Universidade Evangélica de Goiás, 2022.

1. Toxicidade aquática 2. Modelagem Preditiva 3. *Artemia salina.* I. Peixoto, Josana de Castro II. Neves, Bruno Junior III. Título

Catalogação na Fonte
Elaborado por Hellen Lisboa de Souza CRB1/1570

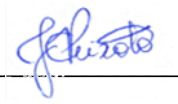
FOLHA DE APROVAÇÃO

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE TOXICIDADE EM *Artemia salina* LEACH**: SUBSÍDIOS À CONSERVAÇÃO DA BIODIVERSIDADE**

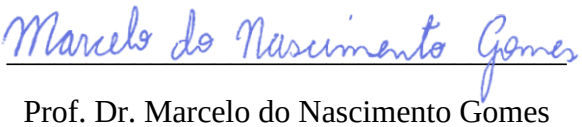
Josiel Araujo Lemes

Dissertação apresentada ao Programa de Pós-graduação em Sociedade, Tecnologia e Meio Ambiente/ PPG STMA da Universidade Evangélica de Goiás/ UniEVANGÉLICA como requisito parcial à obtenção do grau de MESTRE.

Banca examinadora



Profª. Drª. Josana de Castro Peixoto
Centro Universitário de Anápolis, UniEVANGÉLICA
Presidente da Banca Examinadora – Orientadora



Prof. Dr. Marcelo do Nascimento Gomes
Faculdade Metropolitana de Anápolis, FAMA
Membro da Banca Examinadora

Profª. Drª. Lucimar Pinheiro Rosseto
Centro Universitário de Anápolis, UniEVANGÉLICA
Membro da Banca Examinadora

DEDICATÓRIA

Dedico este trabalho à Deus, minha família aos quais sempre me deram total amparo e a todos familiares, amigos e colegas.

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus o Autor da Existência, Aquele que permite que todas as coisas se concretizem.

Em segundo lugar agradeço meus pais pelo apoio em todos os momentos desta trajetória.

Agradeço em especial a professora Dr.ª Josana de Castro Peixoto ao professor Dr. Bruno Junior Neves e ao professor Msc. Roberto Alves Pereira, pois foram protagonistas na minha formação profissional e pessoal, transmitindo além do conhecimento técnico.

Agradeço o corpo docente do Pós-Graduação em Sociedade, Tecnologia e Meio Ambiente (PPSTMA) da UniEVANGÉLICA pelo amparo e suporte na transmissão do conhecimento e saberes.

A toda equipe do *"Laboratory for Molecular Modeling and Drug Design"* (LabMol) na pessoa da professora Prof. Dr. ª Carolina Horta Andrade e Prof. Dr. Bruno Junior Neves pelo suporte na execução dos modelos necessários para realização deste trabalho.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pela bolsa (PROSUP) de mestrado.

“O coração do homem considera o seu caminho,
mas o Senhor lhe dirige os passos. ”

Provérbios 16:9

RESUMO

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE TOXICIDADE EM *Artemia salina* LEACH: SUBSÍDIOS À CONSERVAÇÃO DA BIODIVERSIDADE

Josiel Araujo Lemes[1], Bruno Júnior Neves[2] e Josana de Castro Peixoto[1]

[1]Universidade Evangélica de Goiás, Anápolis, Goiás, Brasil.

[2]Universidade Federal de Goiás, Goiânia, Goiás, Brasil.

Os ecossistemas aquáticos e sua biodiversidade desempenham funções essenciais para a sustentabilidade das comunidades bióticas, na realização de sequestro de carbono e nutrientes, e consequentemente gerando uma "estabilidade" da fração desses compostos no meio ambiente, também sendo cruciais para a produção de pescados, abastecimento de água e recreação. No entanto o uso intensivo de agrotóxicos e produtos químicos industriais, vem sendo considerados responsáveis por uma queda dramática no número de espécies de organismos aquáticos. A realização de ensaios experimentais para avaliação completa da toxicidade dessas substâncias químicas em grande quantidade (em várias doses e concentrações) se torna morosa, onerosa e representa um problema ético. Com avanços em *hardware* e *software*, assim como os grandes avanços no desenvolvimento de algoritmos de aprendizado de máquina, hoje é possível construir, validar e implementar tais modelos computacionais para avaliação de ecotoxicidade durante o registro de novos agrotóxicos, o que torna os modelos de aprendizado de máquina uma alternativa custo-efetiva para analisar o potencial toxicológico e a identificação prévia de potenciais contaminantes. Diante do exposto, o presente projeto tem como objetivo desenvolver modelos de aprendizado de máquina para predição compostos potencialmente tóxicos para organismos aquáticos. Inicialmente, modelos binários baseados em aprendizado de máquina foram desenvolvidos para a espécie *Artemia salina*. Os modelos foram desenvolvidos obedecendo as boas práticas de validação e através da combinação de diferentes tipos de métodos (e.g., *Random Forest - RF*, *Support Vector Machine - SVM* e *Light Gradient Boosting Model - LGBM*) e impressões digitais moleculares (Morgan e FeatMorgan). Dentre os resultados obtidos o melhor foi o com MCC: 0,74; SE: 0,91; SP: 0,83; PPV: 0,84; NPV: 0,91; Kappa: 0,74 e Acurácia: 0,87.

Palavras-chave: Toxicidade aquática; Modelagem preditiva; *Artemia salina*.

ABSTRACT

DEVELOPMENT OF MACHINE LEARNING MODELS FOR TOXICITY PREDICTION IN *Artemia salina* LEACH: SUBSIDIES FOR BIODIVERSITY CONSERVATION

Josiel Araujo Lemes[1], Bruno Júnior Neves[2] e Josana de Castro Peixoto[1]

[1]Universidade Evangélica de Goiás, Anápolis, Goiás, Brasil.

[2]Universidade Federal de Goiás, Goiânia, Goiás, Brasil.

Communities also promote the generation of sources and their source of value to the environment of a generation of carbon, in the sustainability of a source of carbon generation and of energy generation, being their source of carbon generation, stability and supply of fuel. of water and recreation. However, the intensive use of pesticides and industrial chemicals has been thought to lead to a dramatic drop in the number of species of combined organisms. The performance of experiments is for complete evaluation of toxicity, evidence in large quantity (in several doses of tests and trials) With advances in hardware and software, as well as the great advances in the development of possible machines for the evaluation of machines, today it is possible to build , validate and implement ecotoxicity models during the registration of new machine learning models a cost-effective alternative to analyze the toxicological potential and prior identification of potential contaminants. In view of the objective, the project is presented as the preparation of objective models prepared for the machine prepared for the preparation of models. Models designed to learn the machine were designed for an Artemia species. The models were developed through validation practices and the widest variety of method types (eg Random Forest - RF Support Vector Machine - SVM and Gradient Boosting Model - LGBM) and molecular fingerprints (Morgan and FeatMorgan). Among the results obtained, the best was the one with MCC: 0.74; SE: 0.91; SP: 0.83; VPP: 0.84; NPV: 0.91; Kappa: 0.74 and Accuracy: 0.87.

Key words: Aquatic toxicity; Predictive modeling; *Artemia salina*.

LISTA DE FIGURAS

LISTA DE QUADROS

SÍMBOLOS, SIGLAS E ABREVIATURAS

CDB – Convenção sobre Diversidade Biológica

ONU – Organização das Nações Unidas

QSTR – Relação quantitativa estrutura-toxicidade

DL – Dose Letal

LC – Concentração Letal

µM/mL – Microlitro por Mililitro

mg/mL – Miligrama por Mililitro

µg/mL – Microgram por Mililitro

ECFP – Impressões Digitais de Conectividade Estendida

FCFP – Impressões Digitais de Classe Funcional

SE – Sensibilidade

SP – Especificidade

CCR – Taxa de Classificação Correta

PPV – Valor Preditivo Positivo

NPV – Valor Preditivo Negativo

VP – Verdadeiros Positivos

VN – Verdadeiros Negativos

FP – Falsos Positivos

FN – Falsos Negativos

N – Total de Compostos

κ – Kapaa

Pr(a) – Concordância Relativa

Pr(e) – Probabilidade Hipotética

DA – Domínio de Aplicabilidade

SUMÁRIO

1. INTRODUÇÃO

Os ecossistemas aquáticos são divididos em ecossistemas de água doce (dulcícola), incluem rios, córregos, lagoas, lagos, e ecossistemas de água salgada (marinho), correspondem aos mares e oceanos, que se constituem o maior de todos os meios da biosfera e apresentam características próprias, como sua dimensão territorial, ocupando a maior parte da superfície da Terra, cerca de 72% (RAMOS; AZEVEDO, 2010).

O Brasil possui a maior rede hidrográfica do mundo, sendo os ecossistemas aquáticos (fluviais, lacustres permanentes ou temporários) de grande representatividade. A diversidade biológica destes ecossistemas são representativas e alguns dos organismos destes ambientes apresentam valor biológico, na realização de serviços ecológicos essenciais à manutenção da qualidade dos ecossistemas aquáticos, como o sequestro de carbono e nutrientes, e consequentemente gerando uma "estabilidade" da fração desses compostos no meio, também sendo cruciais para a produção de pescados, abastecimento de água e recreação (CERVI et al., 2009; GRIZZETTI et al., 2016; HALPERN et al., 2015; IRFAN; ALATAWI, 2019; ZHANG et al., 2018).

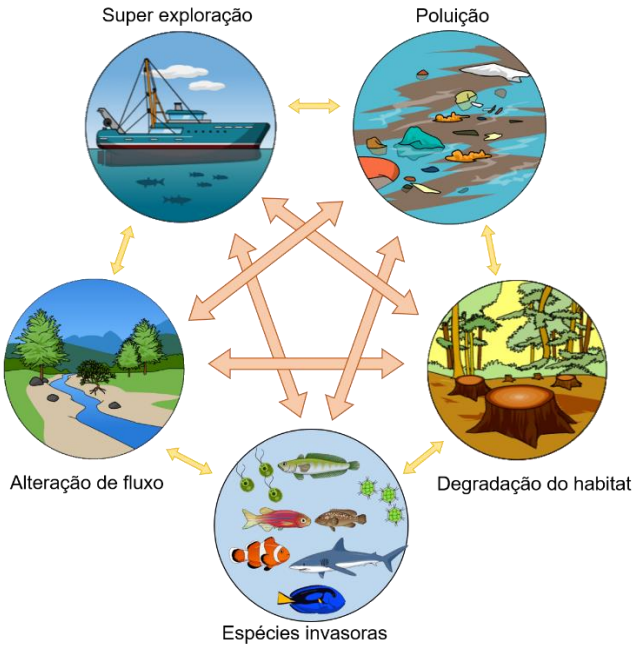
Os ambientes aquáticos, marinhos e continentais abrigam grande diversidade de seres, incluindo algas, bactérias, macrófitas, artrópodes (crustáceos e insetos) e vertebrados. Da fauna que habita os ambientes aquáticos, os peixes representam um pouco mais que a metade das espécies de vertebrados conhecidos no mundo, com 24.618 espécies, sendo que 9.966 espécies ocupam águas doces permanentemente. (NELSON, 1994).

As ameaças à biodiversidade global dos ecossistemas aquáticos podem ser agrupadas em cinco categorias interativas (Figura 01) (ALLAN; FLECKER, 1993; MALMQVIST; RUNDLE, 2002; NAIMAN; TURNER, 2000; RAHEL, 2002; REVENGA et al., 2005). A desestabilização global das estruturas e funções dos ecossistemas, resultam na perda de biodiversidade, afetando diretamente os serviços ecológicos prestados pela fauna e flora aquática (KAFUMBATA; JAMU; CHIOTHA, 2014; SMOL et al., 2005).

Os recursos hídricos e sua biodiversidade encontram-se ameaçados por múltiplas fontes de contaminação provenientes do uso prolongado de produtos químicos em todo o mundo (ARIAS et al., 2007; CHEN et al., 2019). Oriundas de diversas fontes de emissão, como lixos tóxicos provenientes de efluentes industriais e drenagem agrícola (DÍAZ-CRUZ; BARCELÓ, 2008).

.

Figura 1: Categorias interativas de ameaças à biodiversidade dos ecossistemas aquáticos.



Fonte: Autoral, 2021.

2. FUNDAMENTAÇÃO TEÓRICA

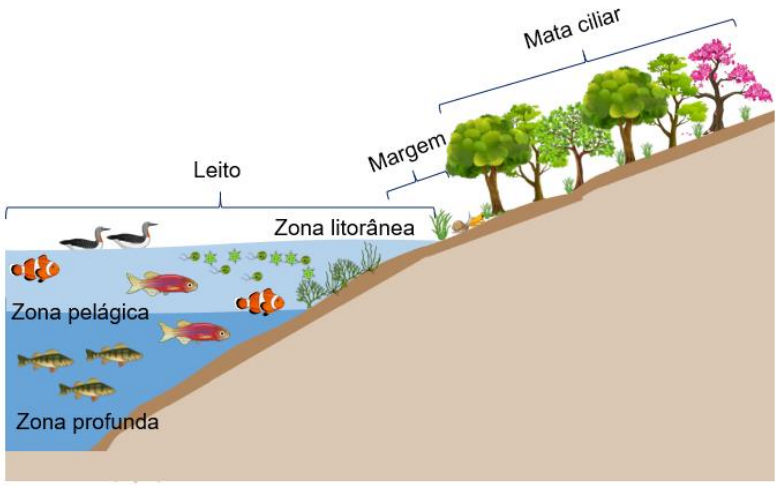
2.1. ECOSSISTEMAS AQUÁTICOS

2.1.1. ECOSSISTEMAS DULCÍCOLAS

As águas doces (Figura 2) fornecem habitats para uma variedade de organismos incluindo bactérias, protozoários, fungos, esponjas, celenterados, vermes, rotíferos, briozoários, moluscos, crustáceos, aracnídeos e vários grupos de insetos (ROCHA, 2006). Importante ressaltar que vários invertebrados de água doce passam parte de seu ciclo de vida no ambiente aquático e parte no ambiente terrestre, como no caso dos Coleoptera, Odonata, Diptera e muitos outros.

Os ecossistemas de água doce são divididos em lênticos e lóticos. Os ecossistemas lênticos incluídos os lagos e lagoas, que são corpos de água cercados por terra, sendo os lagos geralmente são mais extensos e profundos que as lagoas. Já os ecossistemas lóticos hidrologicamente, o rio é um sistema aberto, com fluxo contínuo da nascente à foz, cujo vetor é determinante das características de cada unidade fluvial.

Figura 2: Esquema representativo do ecossistema dulcícola.



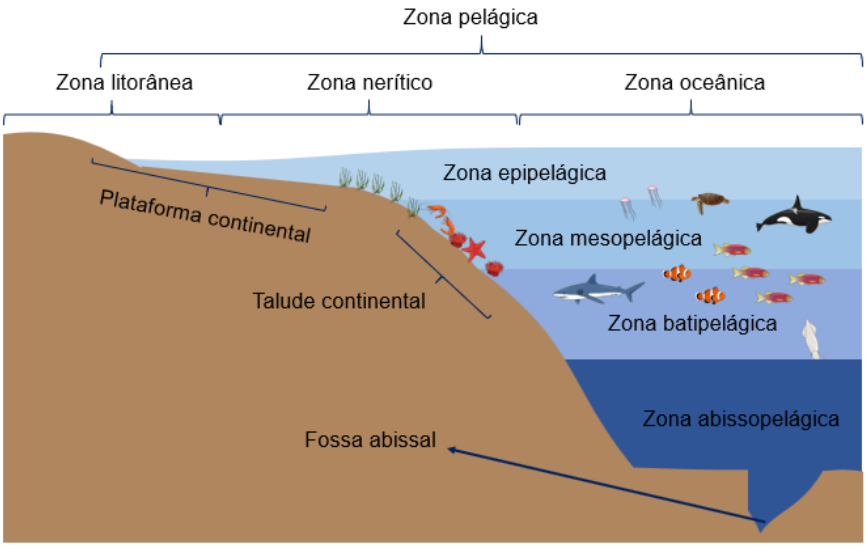
Fonte: Autoral, 2021.

Em termos de biodiversidade, as águas continentais brasileiras apresentam enorme significado global para Algae (25% das espécies do mundo), Porifera (Demospongiae, 33%), Rotifera (25%), Cladocera (Branchiopoda, 20%) e peixes (21%) (AGOSTINHO; THOMAZ; GOMES, 2005). Sendo que a comunidade de macroinvertebrados aquáticos é frequentemente utilizada na avaliação da qualidade de água (MARQUES; FERREIRA; BARBOSA, 1999).

2.1.2. ECOSSISTEMAS MARINHOS

Além de acolher uma ampla variedade de seres vivos, os ecossistemas marinhos (Figura 3) proporcionam serviços essenciais à sobrevivência humana, como alimentos, manutenção do clima, purificação da água, controle de inundações e proteção costeira, além da possibilidade de uso recreativo e espiritual. Associado diretamente à cadeia alimentar e ao ciclo dos nutrientes, o equilíbrio dinâmico dos ecossistemas marinhos está representado pela da fauna e da flora, principalmente as espécies que ocorrem na região entremarés.

Figura 3: Esquema representativo do ecossistema marinho.



Fonte: Autoral, 2021.

Segundo o último Panorama Global da Biodiversidade, editado pela Convenção sobre Diversidade Biológica (CDB) da ONU, os ecossistemas marinhos continuam tendo sua extensão reduzida, o que ameaça serviços ecossistêmicos altamente valiosos e imprescindíveis, como, por exemplo, a absorção de dióxido de carbono da atmosfera, que cumpre papel relevantíssimo na mitigação das mudanças climáticas globais (GROSS; JOHNSTON; BARBER, 2005).

2.2. ECOTOXICOLOGIA AQUÁTICA

A ecotoxicologia aquática é definida como uma ciência, a qual estuda as propriedades e o comportamento dos poluentes nos ecossistemas aquáticos, bem como o impacto dos poluentes nos organismos, populações e comunidades (IRFAN; ALATAWI, 2019; TRUHAUI, 1977). O principal objetivo da ecotoxicologia aquática é avaliar o efeito de substâncias químicas tóxicas sobre organismos representativos do ecossistema aquático, sejam substâncias essas naturais ou sintetizados (BAER, 1996).

Múltiplas fontes de contaminação, provenientes do uso prolongado de produtos químicos, ameaçam a integridade dos recursos hídricos, a nível mundial (DÍAZ-CRUZ; BARCELÓ, 2008).

Análises ecotoxicológicas vêm sendo aplicadas no monitoramento de águas e efluentes industriais no intuito de investigar os efeitos de substâncias químicas manufaturadas e de outros

materiais, antropogênicos ou naturais, em organismos aquáticos (BOUDOU; RIBEYRE, 1997). O primeiro tipo de teste toxicológico a que são submetidos os compostos é o agudo letal, que consiste de uma análise após curta exposição (24h – 48h) do composto com o organismo bioindicador.

Em toxicologia, dose e/ou concentração que induz a morte de 50% de uma população de determinado organismo em teste, se trata do LC_{50} ou LD_{50}. No entanto, a avaliação completa da toxicidade para uma grande quantidade de substâncias químicas (em várias doses e concentrações) por meio destes ensaios experimentais é demorosa, dispendiosa e representa um problema ético.

2.3. TOXICOLOGIA COMPUTACIONAL

O desafio de avaliar a toxicidade de agrotóxicos em organismos aquáticos ressoa fortemente com o tema da química verde que vem ganhando ampla atenção no campo da toxicologia computacional. Nos últimos dez anos, as abordagens de aprendizado de máquina, que são um dos componentes mais importantes da inteligência artificial, têm sido amplamente utilizadas para encontrar relações quantitativas estre estrutura química e toxicidade a partir de conjuntos de dados de compostos com toxicidade experimental pré-determinada (DOBCHEV; PILLAI; KARELSON, 2014; MITCHELL B.O., 2014).

2.3.1. Aprendizagem de máquina

Os métodos computacionais tradicionais são baseadas no conceito de que compostos que compartilham similaridade estrutural (*read-across*) ou certas subestruturas (*structural alerts*) podem compartilhar as mesmas propriedades toxicológicas. No entanto, essas metodologias baseadas em regras têm sido abandonadas por conta de sua limitada taxa de acerto durante as extrapolações (ALVES et al., 2016). Nesse sentido, os esforços no campo da toxicologia preditiva têm sido direcionados para a construção e validação de modelos de classificação usando aprendizado de máquina (ML, do inglês *Machine Learning*), por se tratar de uma metodologia com maior poder de predição (ALVES et al., 2016; GOH; HODAS; VISHNU, 2017).

O ML representa um subcampo da inteligência artificial que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional. Conceitualmente, o algoritmo de ML possui a capacidade de aprender com dados sem ser explicitamente programados para isso (FOURCHES, 2014; TETKO; ENGKVIST; CHEN, 2016). Dois tipos de informação são necessários para a geração de um modelo supervisionado: um conjunto de compostos (variável

independente) e seus respectivos valores de toxicidade experimental (variável dependente). A variável dependente pode ser uma medida quantitativa contínua (e.g., DL50) ou categórica, utilizando um limiar para separar os compostos em classes (e.g., tóxico e não tóxico) quando o intuito é gerar modelos de classificação (KUBINYI, 1993). O processo de desenvolvimento e validação de um modelo pode ser representado em quatro etapas.

Na primeira etapa, o conjunto de dados é dividido em dois subconjuntos: conjunto modelagem e conjunto externo. O conjunto de treinamento, geralmente constituído por 80% dos compostos do conjunto total, é utilizado na construção do modelo. Já os compostos do conjunto externo (20%) são utilizados para avaliar a capacidade preditiva do modelo (CHERKASOV et al., 2014); Na segunda etapa, todas as estruturas químicas são convertidas em descritores moleculares (transformação a informação química em um número útil) (TODESCHINI; CONSONNI, 2008); Na terceira etapa, métodos de ML (e.g., *Random Forest* (RF) (BREIMAN, 2001), *Deep Neural Networks* (DNN) (SCHMIDHUBER, 2015), *Support Vector Machine* (SVM) (VAPNIK, 2000) são utilizados para estabelecer relações quantitativas entre os descritores moleculares (conjunto modelagem) e propriedade toxicológica em estudo (QSTR, do inglês, *Quantitative Structure-Activity/Toxicity Relationship*). Esta etapa consiste em gerar uma hipótese capaz estabelecer e otimizar uma relação entre descritores (x) e propriedade toxicológica (Y) utilizando parâmetros de ajuste (a e b) (LAVECCHIA, 2015; MITCHELL B.O., 2014; WELLING, 2011), conforme descrito na Equação 1:

$$Y(x)=a+bx \qquad (1)$$

Na quarta etapa, a capacidade preditiva do modelo gerado é determinada utilizando métricas apropriadas, as quais irão avaliar a habilidade do modelo em predizer corretamente a toxicidade de compostos do conjunto externo. Uma vez validado, o modelo gerado representa uma ferramenta inestimável para a predição de toxicidade de agrotóxicos não testados experimentalmente (CHERKASOV et al., 2014; TROPSHA, 2010).

3. JUSTIFICATIVA

Considerando (i) o papel destrutivo de substâncias químicas nos efluentes industriais; (ii) o declínio drástico de espécies aquáticas; (iii) o papel *Artemia salina* como indicadores toxicológicos de efluentes e ecossistemas marinhos; e (iv) o alto custo e baixa vasão dos métodos experimentais de avaliação toxicológico, o presente projeto justifica-se pelas seguintes razões:

- Os modelos de classificação baseados em ML apresentam performance preditiva superior aos métodos baseados em regras (*read-across* e *structural alerts*) e métodos de regressão clássicos usados para construção de QSAR;

- Os modelos de classificação disponíveis na literatura científica para avaliação de toxicidade aguda em abelhas não obedecem às boas práticas de modelagem preditiva36 e princípios da (OECD) (OECD, 2004);

- Modelos de ML podem reduzir o tempo e custo associados ao registro e reavaliação de agrotóxicos e favorecer a comercialização de produtos químicos menos tóxicos;

- Uma vez validados, os modelos de classificação podem ser utilizados para triagem virtual de todo o espaço químico de agrotóxicos.

4. OBJETIVOS

4.1. OBJETIVO GERAL

Construir modelos de aprendizado de máquina para predição de compostos potencialmente tóxicos para *Artemia salina*.

4.2. OBJETIVOS ESPECÍFICOS

- Compilar, integrar e padronizar dados de compostos de toxicidade para *Artemia salina*;
- Construir e validar modelos de QSTR binários através da combinação de vários descritores moleculares com o método de aprendizado de máquina *Random Forest*;
- Analisar alertas estruturais e interpretação mecanística dos modelos de QSTR construídos;

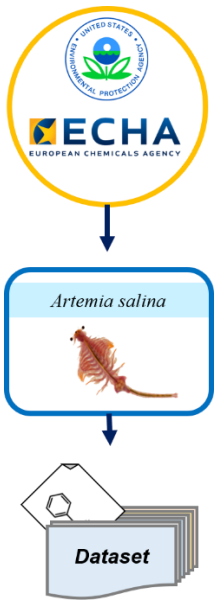
5. MATERIAIS E MÉTODOS

Todas as etapas *in silico* desenvolvidas neste estudo foram executadas em Python v.3.6 (https://www.python.org). O *frame* utilizado é completo, como o módulo para preparação de dados, balanceamento de conjuntos de dados, construção e validação dos modelos QSTR.

5.1. INTEGRAÇÃO, PREPARO E PADRONIZAÇÃO DO CONJUNTO DE DADOS

Todos os compostos com dados de DL_{50} e CL_{50} para *A. salina* foram extraídos dos bancos de dados ECOTOX (https://cfpub.epa.gov/ecotox/) e ChEMBL (https://www.ebi.ac.uk/chembl/) e testados na *in vivo* em *Artemia salina* (Figura 4). Em seguida, compostos com dados de toxicidade expressos na escala de peso molecular (mg/mL, µg/mL, etc.) foram convertidos para escala molar (µM/mL). Ao final deste processo, compostos com LC_{50} ≤10 µM/mL foram classificados como tóxicos, ao passo que compostos com LC_{50} >10 µM/mL foram classificados como não tóxicos ("Chemical Hazard Classification and Labeling : comparison of opp requirements and the ghs", 2004).

Figura 4: Diagrama da extração das *datasets* da base de dados da ECOTOX para *Artemia salina*.

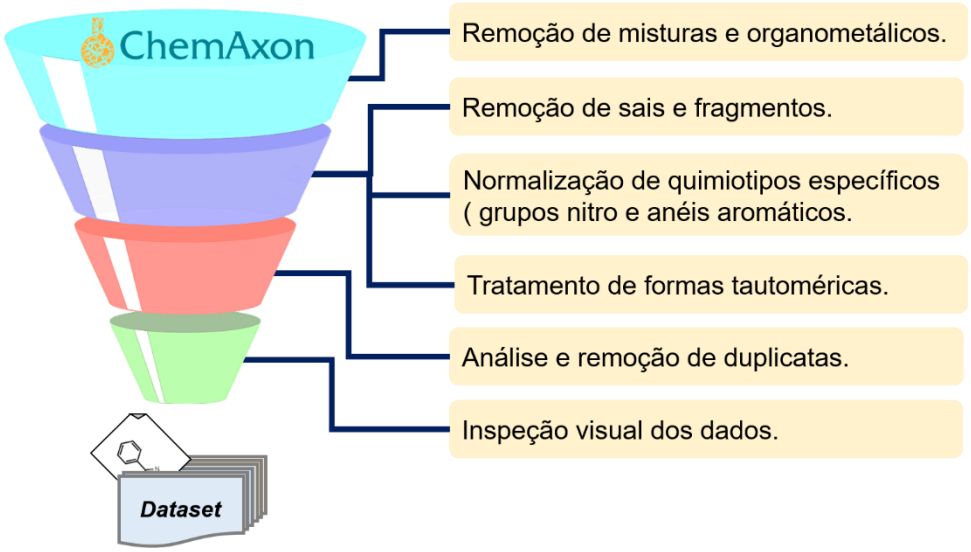


Fonte: Autoral, 2021.

Conjuntos de dados públicos contêm uma fração de registros compilados erroneamente, cuja presença é causada pela falta padronização incompleta dos dados, variações de medição e verificação de qualidade insuficiente. Em termos gerais, esses dados "ruins" englobam entradas que contêm anotações de estereoisômeros pouco claras, estruturas químicas com problemas de valência, valores de toxicidade reportados erroneamente e registros duplicados.

Desta forma todos os compostos foram cuidadosamente padronizados de acordo com o protocolo estabelecido por Fourches e colaboradores (FOURCHES; MURATOV; TROPSHA, 2010, 2015, 2016). Hidrogênios explícitos foram adicionados enquanto que polímeros, sais, metais, compostos organometálicos e misturas foram removidos. Em paralelo, quimiotipos específicos como anéis aromáticos e grupos nitro foram normalizados (Figura 05). Todas estas etapas foram realizadas utilizando o programa Standardizer 20 (v.16.9.5.0, ChemAxon, Budapest, Hungary, http://www.chemaxon.com).

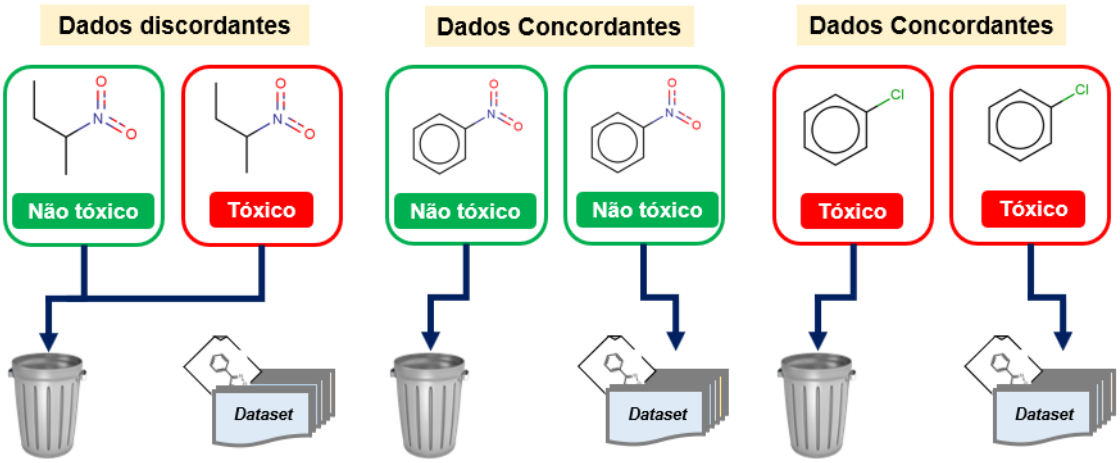
Figura 5: Curagem dos dados.



Fonte: Adaptado de Fourches 2010.

Em seguida, todos os conjuntos de dados foram submetidos a um processo de análise e remoção de duplicatas seguindo os seguintes critérios: (i) duplicatas com propriedades toxicológicas discordantes foram excluídas; e (ii) duplicatas com propriedades toxicológicas concordantes tiveram uma das entradas mantida no conjunto de dados e as demais excluídas (Figura 06).

Figura 6: Análise e remoção de duplicatas.

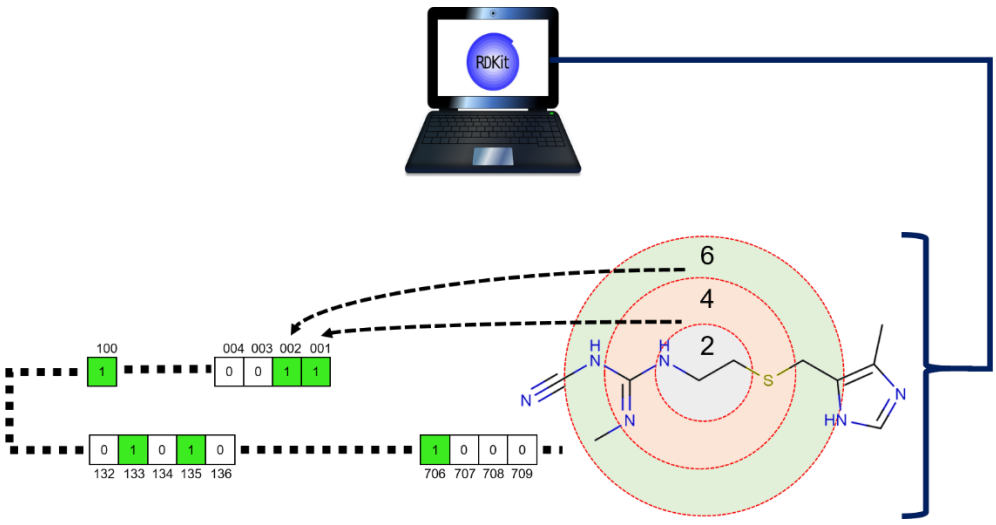


Fonte: Autoral, 2021.

5.2. DESCRITORES MOLECULARES

As impressões digitais moleculares circulares do tipo Morgan (ECFP) e FeatMorgan (FCFP) foram calculadas no programa de código aberto RDKit (http://www.rdkit.org), (RINIKER; LANDRUM, 2013). Ambas as impressões digitais foram geradas com tamanho de raio variando entre 2−6 e com comprimento de 2.048 bits.

Figura 7: Impressões digitais moleculares circulares.



Fonte: Autoral, 2021.

5.3. MODELOS DE APREDIZAGEM BINÁRIOS

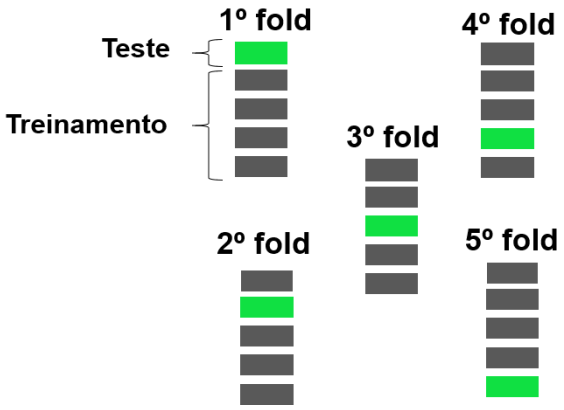
Os modelos binários de aprendizado de máquina foram construídos combinando impressões digitais moleculares com o algoritmo *Random Forest - RF*, *Support Vector Machine - SVM* e *Light Gradient Boosting Model - LGBM* (BREIMAN, 2001). Os modelos foram gerados utilizando o pacote de código aberto Scikit-learn (http://scikit-learn.org/stable, (PEDREGOSA et al., 2011).

Os parâmetros de busca da grade foram definidos utilizando diversas combinações de árvores (100, 150, 200, 250, 300) e o coeficiente kappa de Cohen's como função de pontuação do estimador.

5.4. VALIDAÇÃO CRUZADA EXTERNA DE 5-FOLDS

Os conjuntos de dados foi dividido aleatoriamente em cinco subconjuntos de igual tamanho; então um desses subconjuntos (20% de todos os compostos) foi considerado como conjunto teste e as demais partes juntas formaram o conjunto de treinamento (80% do conjunto completo). Este procedimento foi repetido cinco vezes, permitindo que cada um dos cinco subconjuntos fosse usado como conjunto de validação teste. Os modelos foram construídos utilizando apenas usando o conjunto treinamento. É importante ressaltar que os compostos no conjunto teste momentâneo não foram utilizados para construir os modelos (ALVES et al., 2015a, 2015b).

Figura 8: Validação cruzada externa de 5-folds.



Fonte: Autoral, 2021.

5.5. AVALIAÇÃO DE PREDITIVIDADE DOS MODELOS

O desempenho preditivo dos modelos binários gerados foi avaliado utilizando a sensibilidade (SE), especificidade (SP), taxa de classificação correta (CCR), valor preditivo

positivo (PPV) e valor preditivo negativo (NPV). Essas métricas foram calculadas da seguinte forma:

$$\mathrm{SE} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \quad (1)$$

$$\mathrm{SP} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}} \quad (2)$$

$$\mathrm{CCR} = \frac{\mathrm{SE} + \mathrm{SP}}{2} \quad (3)$$

$$\mathrm{PPV} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \quad (4)$$

$$\mathrm{NPV} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FN}} \quad (5)$$

Aqui, VP representa o número de verdadeiros positivos, VN o número de verdadeiros negativos, FP o número de falsos positivos, FN é o número de falsos negativos e N é o número total de compostos. A SE representa uma medida de correção na predição de produtos químicos tóxicos. A SP caracteriza uma medida de correção na predição de produtos químicos não tóxicos. Já o CCR é a média aritmética dos valores de SE e SP. Por fim, PPV e NPV foram calculados para estimar a probabilidade de acerto durante a predição um novo composto tóxico ou não tóxico. Valores de CCR, SE, SP, PPV e NPV maiores que 0,65 indicam que os modelos são preditivos.

Além das métricas acima, o κ foi utilizado para comparar o quão bem as predições concordam com os valores experimentais (COHEN, 1960; VIERA; GARRETT, 2005). Ele será calculado utilizando as seguintes equações:

$$\Pr(a) = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{N}} \quad (6)$$

$$\Pr(\mathrm{e}) = \frac{(\mathrm{VP} + \mathrm{FP})\ \mathrm{x}\ (\mathrm{VP} + \mathrm{FN}) + (\mathrm{VN} + \mathrm{FN})\ \mathrm{x}\ (\mathrm{VN} + \mathrm{FP})}{\mathrm{N}} \quad (7)$$

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (8)$$

Nas equações acima, Pr(a) representa a concordância relativa observada entre a predição realizada pelo modelo e os valores conhecidos e Pr(e) representa a probabilidade hipotética de

concordância. Valores de κ maiores que 0.6 indicam boa concordância entre os valores preditos e experimentais.

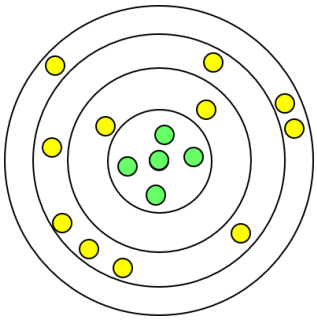
5.6. DOMÍNIO DE APLICABILIDADE

O DA foi estimado com base na distância Euclidiana entre os compostos do conjunto de modelagem (5FECV) e cada compostos no subconjunto de dados de validação externa. No presente estudo, o DA foi definido como um limiar de distância L_D entre um composto submetido a uma predição e seu vizinho mais próximo no conjunto modelagem. O DA foi calculado utilizando a seguinte equação:

$$L_D = \bar{y} + Z\sigma \qquad (9)$$

Na qual $\bar{y}$ é a distância Euclidiana média de k vizinhos mais próximos dentro do conjunto de modelagem, σ é o desvio padrão dessas distâncias Euclidianas e Z um parâmetro arbitrário para controlar o nível de significância. Se a distância de um composto exceder o limiar estabelecido, a predição foi considerada como menos confiável (TROPSHA, 2010).

Figura 9: Domínio de aplicabilidade.



Fonte: Autoral, 2021.

* Círculos de cor verde que estão dentro do limiar estabelecido, já os círculos em amarelo exemplificando os compostos que excedem o limiar estabelecido.

5.7. ANÁLISE DO ESPAÇO QUÍMICO

A análise do espeço químico foi realizada utilizando o programa DataWarrior (SANDER et al., 2015), sendo calculado o se descritor interno SkelSpheres, como tambéma análise de *cliffs* e similaridade entre os compostos ativos e inativos. Esta análise, permite identificar *cliffs* de

atividade em que compostos estruturalmente similares podem apresentar atividades extremamente discordantes.

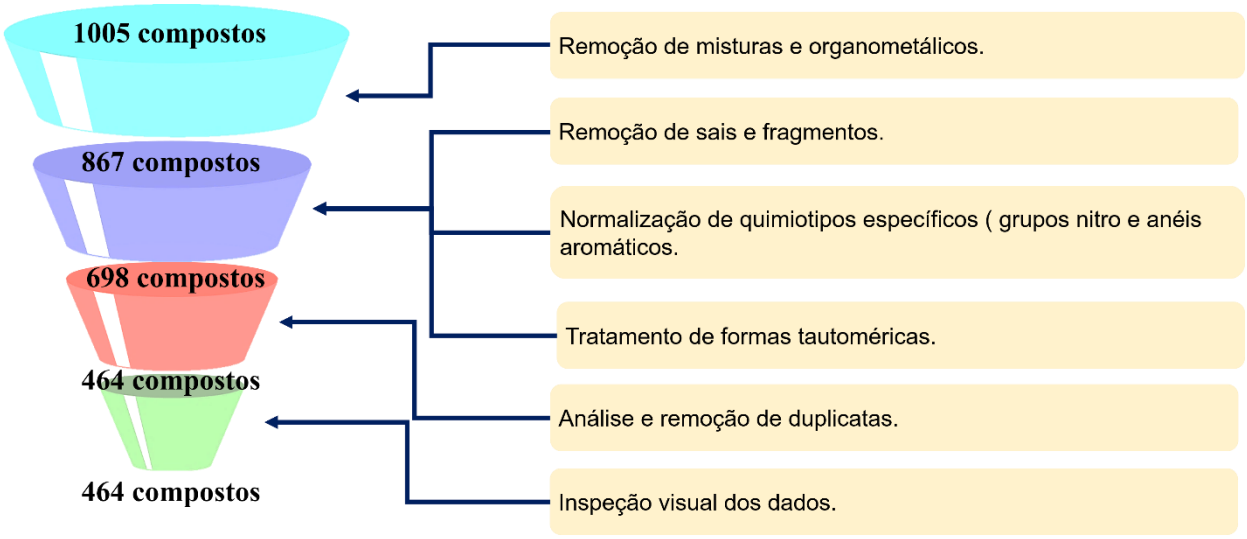
6. RESULTADOS E DISCUSSÕES

6.1. INTEGRAÇÃO, PREPARO E PADRONIZAÇÃO DO CONJUNTO DE DADOS

Todos os compostos com dados de DL_{50} e CL_{50} para *A. salina* foram extraídos dos bancos de dados ECOTOX (https://cfpub.epa.gov/ecotox/) e ChEMBL (https://www.ebi.ac.uk/chembl/) e testados na *in vivo* em *Artemia salina* (Figura 4), como já explicado no item 4.3 e curado seguindo protocolos pré-estabelecidos por Fourches e colaboradores (2010 e 2016).

Com os dados coletados, foram obtidas 725 moléculas com dados de LC_{50} para *A. salina* foram preparadas, padronizadas e visualmente inspecionadas. Tal prática permitia a identificação de 101 duplicatas e 31 compostos com dados toxicológicos inconclusivos.

Figura 10. Exemplo de Curagem dos dados para *Artemia salina*.



6.2. MODELOS DE APRENDIZADO DE MÁQUINA

Os modelos foram gerados e validados usando os conjuntos de dados de DL_{50} e CL_{50} para *A. salina* preparadas e curadas anteriormente, para distinguir compostos ativos e inativos. Ao todo, foram gerados 71 modelos para o conjunto de dados através da combinação de três algoritmos de aprendizado de máquina (Random Forest, Support Vector Machine e Light Gradient Boosting Machine) com 2 descritores moleculares (ECFP e FCFP), diferindo o raio entre 2 a 6.

A escolha dos *fingerprints* circulares (ECFP e FCFP) ao invés de outros descritores (AtomPair, MACCS e etc) é devido principalmente os descritores circulares terem a capacidade de descrever de forma mais precisa a presença e ausência de grupos funcionais em uma estrutura química e assim, representarem melhor a molécula. Além disso, os descritores circulares foram construídos especificamente para serem usados na construção de modelos que relacionam a estrutura química e a sua atividade (CERETO-MASSAGUÉ et al., 2015b).

Em seguida, os modelos gerados foram calibrados alterando os limites de probabilidade (valor padrão = 50%) usados para a rotulação das classes (ZAKHAROV et al., 2014). A estimativa de probabilidade representa um importante parâmetro para avaliação de confiança de predições de atividade biológica. Usualmente, compostos com estimativas de probabilidade >50% são rotulados como ativos, enquanto compostos com probabilidade <50% são rotulados como inativos. Entretanto, os modelos de classificação desenvolvidos com conjuntos de dados desbalanceados, geralmente fornecem estimativas de probabilidade pobres para a classe minoritária. Curiosamente, mesmo quando o desempenho geral é satisfatório, o modelo tem dificuldade em distinguir entre as classes e a confiança dessas predições é baixa (NICULESCU-MIZIL; CARUANA, 2005; WALLACE; DAHABREH, 2012). Consequentemente, o limiar de 50% pode não representar um "corte" adequado para rotulação de compostos como ativos e inativos.

Por outro lado, o processo de calibração não resultou em grandes melhorias nos desempenhos estatísticos interno e externo dos modelos gerados para a predição de toxicidade para *A. salina*.

O melhor melhores métricas estatísticas foram obtidas utilizando o descritor *fingerprint* FCFP, de raio 2, com o algoritmo *Light*GBM, calibrado com o limiar de 0,25 com acurácia balanceada de 0.78, sensibilidade (SE) de 0,8, especificidade (SP) de 0,77, valor preditivo positivo (PPV) e negativo (NPV) de 0,78 e 0,79, respectivamente e coeficiente de kappa Cohen's (κ) de 0,57 para a validação interna e na validação externa de acurácia balanceada de 0.87, SE de 0,91, SP de 0,83, PPV e NPV de 0,84 e 0,91, respectivamente e κ de 0,74. Encontram-se no Quadro 1 e 2.

Por outro lado, o melhor modelo escolhido foi utilizando o descritor *fingerprint* FCFP, de raio 2, com o algoritmo *Ligth*GBM, descalibrado, apresentando acurácia balanceada de 0.78, sensibilidade (SE) de 0,75, especificidade (SP) de 0,81, valor preditivo positivo (PPV) e negativo

(NPV) de 0,8 e 0,76, respectivamente e coeficiente de kappa Cohen's (κ) de 0,56 para a validação interna e na validação externa de acurácia balanceada de 0.86, SE de 0,85, SP de 0,87, PPV e NPV de 0,87 e 0,85, respectivamente e κ de 0,72. Pois o mesmo não possui calibração e tem métricas semelhantes com modelo com *Light*GBM, calibrado, seguindo a regra da parcimônia em modelos mais simples podem ser modelos melhores. Encontram-se no Quadro 1 e 2.

Entre os cinco melhores modelos de aprendizado de máquina obtidos na validação externa, temos a predominância do descritor FCFP, possuindo apenas um modelo em quarto lugar em que foi utilizado o descritor ECFP. Essa preferência ao descritor FCFP, supõem-se devido a sua única diferença com o ECFP, em que são ainda mais abstraídas e, em vez de indexar um átomo específico no ambiente, eles indexam o papel desse átomo. Assim, átomos ou grupos diferentes com função igual ou semelhante não são distinguidos pela impressão digital (CERETO-MASSAGUÉ et al., 2015a).

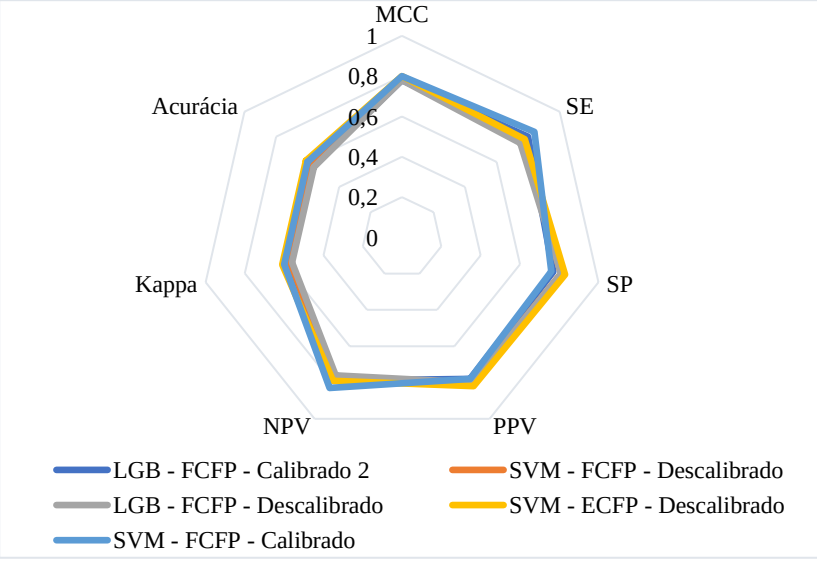
Além disso, o algoritmo *Light*GBM, pode ser utilizado eficientemente em datasets desbalanceadas, pois o seu método de *boosting* constrói uma árvore de cada vez sendo uma dependente da outra, sendo assim o aprendizado, o combinado de decisões no decorrer do tempo. Diferente dos métodos de RF e SVM, em que cada árvore é treinada independentemente e o aprendizado é a média de todas as árvores e que determina um hiperplano em um espaço n-dimensional que separa os ativos dos inativos, respectivamente.

Todos os modelos de ML construídos e o gráfico estatísticos dos melhores modelos para cada algoritmo, *fingerprints* e os seus raios, encontram-se no anexo.

Quadro 1: Resultados estatísticos dos melhores modelos de aprendizado de máquina na validação interna.

Validação interna								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**κ**	**ACC**
LGB - FCFP - Calibrado	2	0.78	0.80	0.77	0.78	0.79	0.57	0.57
SVM - FCFP - Descalibrado		0.79	0.76	0.81	0.80	0.77	0.57	0.57
LGB - FCFP - Descalibrado		0.78	0.75	0.81	0.80	0.76	0.56	0.56
SVM - ECFP - Descalibrado		0.8	0.78	0.83	0.82	0.79	0.61	0.61
SVM - FCFP - Calibrado		0.8	0.84	0.76	0.78	0.83	0.60	0.6

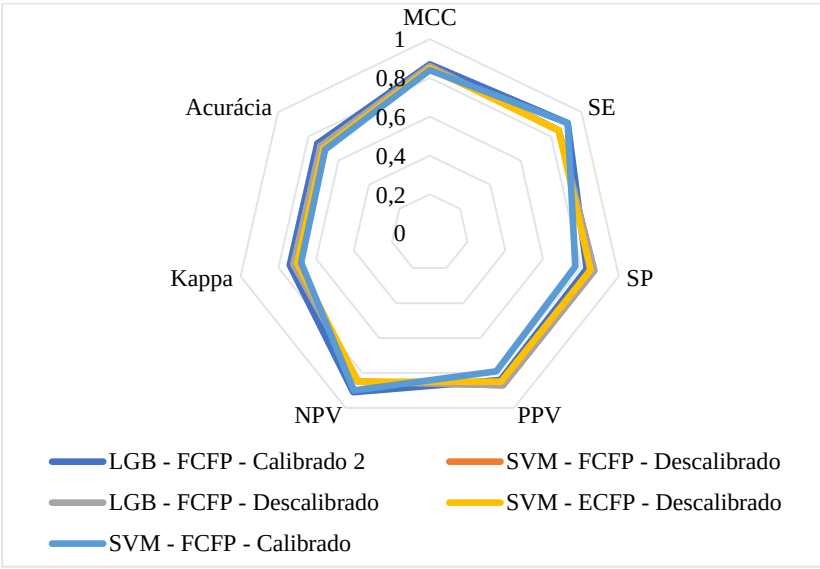
Figura 101: Gráfico dos resultados estatísticos dos melhores modelos de aprendizado de máquina na validação interna.



Quadro 2: Resultados estatísticos dos melhores modelos de aprendizado de máquina na validação externa.

Validação externa								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**κ**	**ACC**
LGB - FCFP - Calibrado	2	0.87	0.91	0.83	0.84	0.91	0.74	0.74
SVM - FCFP - Descalibrado		0.86	0.85	0.87	0.87	0.85	0.72	0.72
LGB - FCFP - Descalibrado		0.86	0.85	0.87	0.87	0.85	0.72	0.72
SVM - ECFP - Descalibrado		0.85	0.85	0.85	0.85	0.85	0.70	0.7
SVM - FCFP - Calibrado		0.84	0.91	0.77	0.79	0.90	0.68	0.69

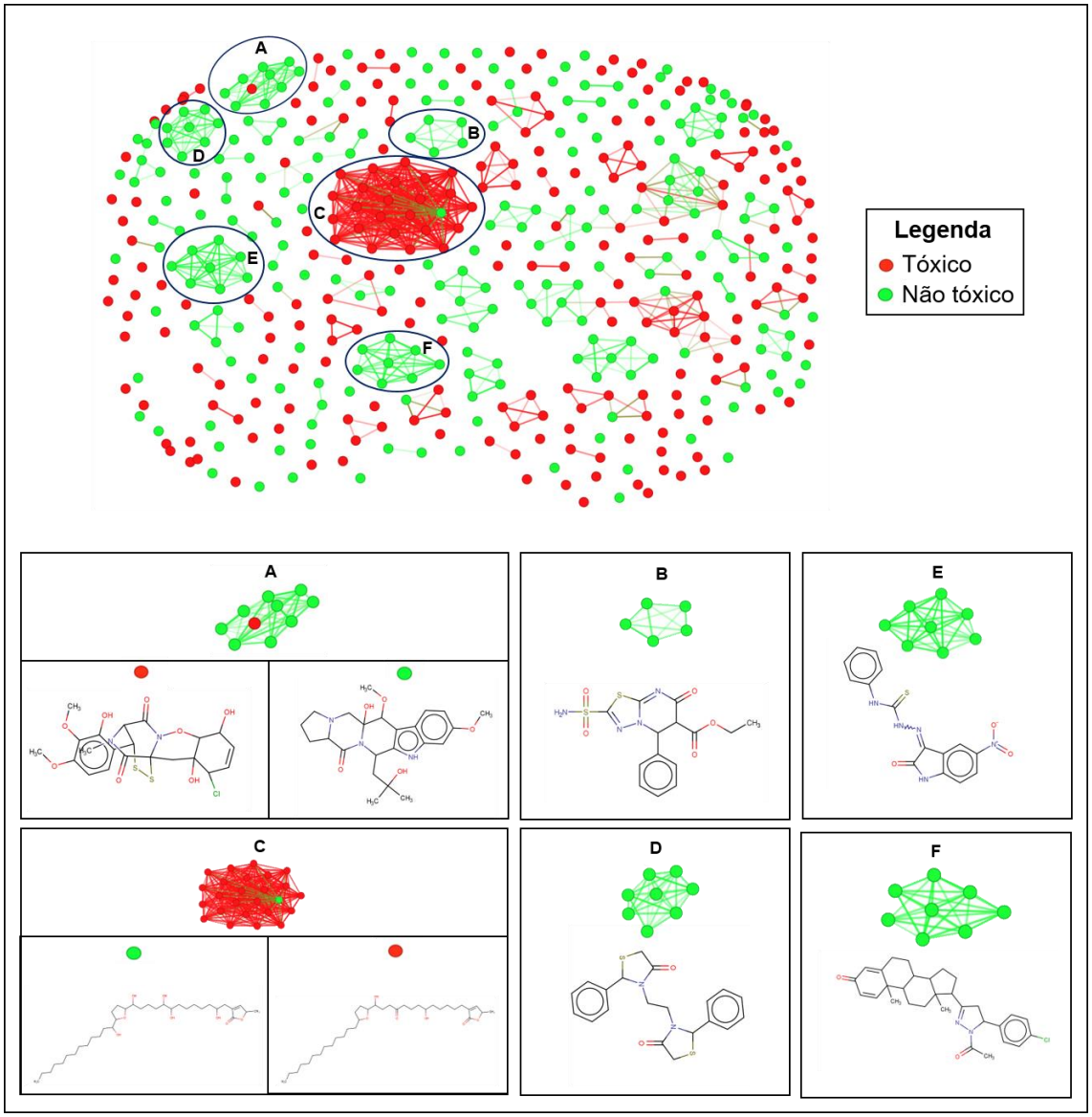
Figura 11: Gráfico dos resultados estatísticos dos melhores modelos de aprendizado de máquina na validação externa.



6.3. ANÁLISE DO ESPAÇO QUÍMICO

Através da análise do espaço demostra a presença de grupos tóxicos e não tóxicos bem definidos (grupos B, D, E e F) com poucos *cliffs* de atividade (grupos A e C). Essa análise permite justificar a realização das boas métricas de validação interna e externa dos modelos construídos e apresentados no item anterior. Pois na Figura 12 A e C, podemos visualizar a semelhança estrutural dos compostos apresentadas nestes grupos e em todo o espaço químico, o que não interferiu no bom aprendizado dos modelos. Pois mesmo assim, o modelo conseguiu aprender muito bem, mesmo sem a continuidade entre modificações estruturais e valores de atividade biológica.

Figura 12: Análise do espaço químico e comparação dos dados.



7. CONCLUSÕES

Foi compilado, integrado e padronizados o maior conjunto de dados de compostos com dados de DL_{50} e CL_{50} para *A. salina* para estudos de ecotoxicidade em *A. salina* envolvendo a integração de métodos *in silico*.

Foi construído e validado estatisticamente modelos de QSTR binários através da combinção dos descritores FCFP, raio 2 e algoritmo de aprendizado de máquina *Ligth*GBM, com métrica estatísticas de acurácia balanceada de 0.78, sensibilidade (SE) de 0,75, especificidade (SP) de 0,81, valor preditivo positivo (PPV) e negativo (NPV) de 0,8 e 0,76, respectivamente e coeficiente de kappa Cohen's (κ) de 0,56 para a validação interna e na validação externa de acurácia balanceada de 0.86, SE de 0,85, SP de 0,87, PPV e NPV de 0,87 e 0,85, respectivamente e κ de 0,72.

A interpretação mecanística do melhor modelo gerado (combinação de descritores Morgan com método de aprendizado de máquina *Ligth*GBM) através dos mapas de probabilidade predita, é possível entender e ilustrar a relação entre grupos e/ou fragmentos estruturais que contribuem para toxicidade e não toxicidade em *A. salina*.

O desenvolvimento de servidores para a avaliação preditiva de ecotoxicidade em *A. salina* podem ser desenvolvidos utilizando ferramentas de código aberto como Flask, RDKit, e Python (MILLMAN; AIVAZIS, 2011). O pacote Flask é uma *microframework* popular e extensível que permite a construção de servidores e aplicativos utilizando Python. Já o RDKit é uma interface de programação de aplicativos de quimioinformática que calcula descritores moleculares como impressões digitais de Morgan (ROGERS; HAHN, 2010) (representação numérica da estrutura química) para o desenvolvimento de modelos de aprendizado de máquina.

O desenvolvimento e implementação de infraestrutura tecnológica como a realizada neste trabalho proporciona maior acesso à dados e informações científicas relacionada a poluição por compostos químicos, o que possibilita subsidiar o compartilhamento de dados e assim a execução de ações de conservação e de uso sustentável bem como a promoção do conhecimento da nossa biodiversidade.

Desta forma este estudo colabora para melhor planejamento de pesquisadores e profissionais da área química para o desenvolvimento de compostos industriais que após utilizados afetem de forma direta ou indireta os recursos hídricos e consequentemente a população de *A. salina* e/ou comunidade aquática.

8. REFERÊNCIAS BIBLIOGRÁFICAS

AGOSTINHO, A. A.; THOMAZ, S. M.; GOMES, L. C. Conservação da biodiversidade em águas. **Megadiversidade**, v. 1, n. 1, p. 70–78, 2005.

ALLAN, J. D.; FLECKER, A. S. Biodiversity Conservation in Running Waters. **BioScience**, v. 43, n. 1, p. 32–43, jan. 1993.

ALVES, V. M. et al. Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. **Toxicology and Applied Pharmacology**, v. 284, n. 2, p. 273–280, 2015a.

ALVES, V. M. et al. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. **Toxicology and Applied Pharmacology**, v. 284, n. 2, p. 262–272, 2015b.

ALVES, V. M. et al. QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. **Green Chemistry**, v. 18, n. 24, p. 6501–6515, 2016.

ARIAS, A. R. L. et al. Utilização de bioindicadores na avaliação de impacto e no monitoramento da contaminação de rios e córregos por agrotóxicos. **Ciência & Saúde Coletiva**, v. 12, n. 1, p. 61–72, mar. 2007.

BAER, K. N. Fundamentals of Aquatic Toxicology: Effects, Environmental Fate, and Risk Assessment. **Journal of the American College of Toxicology**, v. 15, n. 5, p. 453–454, out. 1996.

BOUDOU, A.; RIBEYRE, F. Aquatic Ecotoxicology: From the Ecosystem to the Cellular and Molecular Levels. **Environmental Health Perspectives**, v. 105, n. SUPPL. 1, p. 21, fev. 1997.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CERETO-MASSAGUÉ, A. et al. Molecular fingerprint similarity search in virtual screening. **Methods**, v. 71, n. C, p. 58–63, 2015a.

CERETO-MASSAGUÉ, A. et al. Molecular fingerprint similarity search in virtual screening. **Methods**, v. 71, p. 58–63, jan. 2015b.

CERVI, A. C. et al. Macrófitas aquáticas do Município de General Carneiro, Paraná, Brasil. **Biota Neotropica**, v. 9, n. 3, p. 215–222, set. 2009.

Chemical Hazard Classification and Labeling : comparison of opp requirements and the ghs.

CHEN, J. et al. Aquatic ecosystem health assessment of a typical sub-basin of the Liao River based on entropy weights and a fuzzy comprehensive evaluation method. **Scientific Reports**, v. 9, n. 1, p. 1–13, 2019.

CHERKASOV, A. et al. **QSAR modeling: Where have you been? Where are you going to?Journal of Medicinal Chemistry**, 2014.

COHEN, J. A coefficient of agreement of nominal scales. **Educational and Psychological Measurement**, v. 20, p. 37–46, 1960.

DÍAZ-CRUZ, M. S.; BARCELÓ, D. Trace organic chemicals contamination in ground water recharge. **Chemosphere**, v. 72, n. 3, p. 333–342, jun. 2008.

DOBCHEV, D.; PILLAI, G.; KARELSON, M. In Silico Machine Learning Methods in Drug Development. **Current Topics in Medicinal Chemistry**, v. 14, n. 16, p. 1913–1922, out. 2014.

FOURCHES, D. et al. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. **Journal of Chemical Information and Modeling**, v. 50, n. 7, p. 1189–204, jul. 2010.

FOURCHES, D. Cheminformatics: At the Crossroad of Eras. In: GORB, L.; KUZ'MIN, V. E.; MURATOV, E. N. (Eds.). . **Application of Computational Techniques in Pharmacy and Medicine**. [s.l.] Springer Netherlands, 2014. p. 539–546.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. **Journal of Chemical Information and Modeling**, v. 50, n. 7, p. 1189–1204, jul. 2010.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. **Curation of chemogenomics dataNature Chemical Biology**Nature Publishing Group, , 21 jul. 2015.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. **Journal of Chemical Information and Modeling**, v. 56, n. 7, p. 1243–1252, 25 jul. 2016.

GOH, G. B.; HODAS, N. O.; VISHNU, A. Deep learning for computational chemistry. **Journal of Computational Chemistry**, v. 38, n. 16, p. 1291–1307, jun. 2017.

GRIZZETTI, B. et al. Assessing water ecosystem services for water resource management. **Environmental Science & Policy**, v. 61, p. 194–203, jul. 2016.

GROSS, T.; JOHNSTON, S.; BARBER, C. V. A Convenção sobre Diversidade Biológica : Entendendo e Influenciando o Processo Um Guia para Entender e Participar Efetivamente da. **Instituto de Estudos Avançados da Universidade das Nações Unidas**, p. 72, 2005.

HALPERN, B. S. et al. Spatial and temporal changes in cumulative human impacts on the world's ocean. **Nature Communications**, v. 6, n. 1, p. 7615, nov. 2015.

IRFAN, S.; ALATAWI, A. M. M. Aquatic Ecosystem and Biodiversity: A Review. **Open Journal of Ecology**, v. 09, n. 01, p. 1–13, 2019.

KAFUMBATA, D.; JAMU, D.; CHIOTHA, S. Riparian ecosystem resilience and livelihood strategies under test: Lessons from Lake Chilwa in Malawi and other lakes in Africa. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 369, n. 1639, p. 7–9, 2014.

KUBINYI, H. **QSAR: Hansch Analysis and Related Approaches**. Weinheim, Germany: Wiley-VCH Verlag GmbH, 1993.

LAVECCHIA, A. Machine-learning approaches in drug discovery: Methods and applications. **Drug Discovery Today**, v. 20, n. 3, p. 318–331, 2015.

MALMQVIST, B.; RUNDLE, S. Threats to the running water ecosystems of the world. v. 29, n. 2, p. 134–153, 2002.

MARQUES, M. G. S. M.; FERREIRA, R. L.; BARBOSA, F. A. R. A comunidade de macroinvertebrados aquáticos e características limnológicas das lagoas Carioca e da Barra, Parque Estadual do Rio Doce, MG. **Revista Brasileira de Biologia**, v. 59, n. 2, p. 203–210, maio 1999.

MILLMAN, K. J.; AIVAZIS, M. Python for Scientists and Engineers. **Computing in Science & Engineering**, v. 13, n. 2, p. 9–12, mar. 2011.

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

NAIMAN, R. J.; TURNER, M. G. A Future Perspective on North America's Freshwater Ecosystems. **Ecological Applications**, v. 10, n. 4, p. 958, ago. 2000.

NICULESCU-MIZIL, A.; CARUANA, R. **Predicting good probabilities with supervised learning**. Proceedings of the 22nd international conference on Machine learning - ICML '05. **Anais**...New York, New York, USA: ACM Press, 2005.

OECD. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models. **Biotechnology**, n. November, p. 1–2, 2004.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, 2011.

RAHEL, F. J. Homogenization of freshwater faunas. **Annual Review of Ecology and Systematics**, v. 33, p. 291–315, 2002.

RAMOS; AZEVEDO. **Ecossistemas aquáticos**. [s.l: s.n.].

REVENGA, C. et al. Prospects for monitoring freshwater ecosystems towards the 2010 targets. **Philosophical Transactions of the Royal Society**, v. 360, n. 1454, p. 397–413, fev. 2005.

RINIKER, S.; LANDRUM, G. A. Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. **Journal of Cheminformatics**, v. 5, n. 9, p. 1–7, 2013.

ROCHA, O. Organismos de água doce. **Avaliação do Estado do Conhecimento da Biodiversidade Brasileira**, p. 15–40, 2006.

ROGERS, D.; HAHN, M. Extended-connectivity fingerprints. **Journal of Chemical Information and Modeling**, v. 50, n. 5, p. 742–754, 24 maio 2010.

SANDER, T. et al. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. **Journal of Chemical Information and Modeling**, v. 55, n. 2, p. 460–473, 23 fev. 2015.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural Networks**, v. 61, p. 85–117, jan. 2015.

SMOL, J. P. et al. Climate-driven regime shifts in the biological communities of arctic lakes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 12, p. 4397–4402, 2005.

TETKO, I. V; ENGKVIST, O.; CHEN, H. Does 'Big Data' exist in medicinal chemistry, and if so, how can it be harnessed? **Future Medicinal Chemistry**, v. 8, n. 15, p. 1801–1806, out. 2016.

TODESCHINI, R.; CONSONNI, V. **Handbook of molecular descriptors**. Weinheim, Germany: Wiley-VCH Verlag GmbH, 2008.

TROPSHA, A. Best Practices for QSAR Model Development, Validation, and Exploitation. **Molecular Informatics**, v. 29, n. 6–7, p. 476–488, 12 jul. 2010.

TRUHAUI, R. Ecotoxicology : Objectives , Principles and Perspectives. p. 151–173, 1977.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. [s.l: s.n.].

VIERA, A. J.; GARRETT, J. M. Understanding interobserver agreement: the kappa statistic. **Family medicine**, v. 37, n. 5, p. 360–363, maio 2005.

WALLACE, B. C.; DAHABREH, I. J. **Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them)**. IEEE 12th International Conference on Data Mining. **Anais**...IEEE, dez. 2012.

WELLING, M. **A first encounter with Machine Learning**. 1. ed. Irvine: University of California, 2011.

ZAKHAROV, A. V. et al. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. **Journal of Chemical Information and Modeling**, v. 54, n. 3, p. 705–712, mar. 2014.

ZHANG, K. et al. Freshwater lake ecosystem shift caused by social-economic transitions in Yangtze River Basin over the past century. **Scientific Reports**, v. 8, n. 1, p. 17146, dez. 2018.

9. ANEXO

Quadro 3: Todos os modelos de aprendizado de máquina gerados.

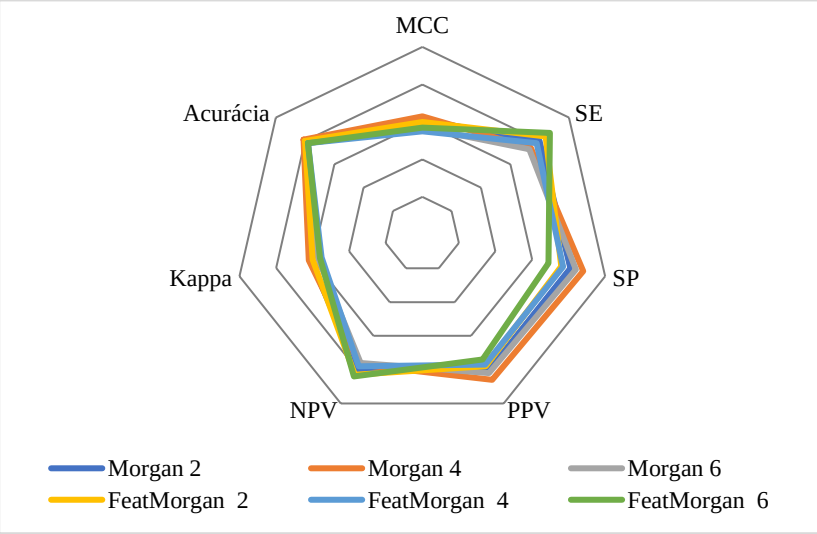
Fingerprint	Raio	Algoritmo	Calibrado	Validação	Limiar	Bal-acc	Se	Sp	PPV	NPV	κ	MCC
featmorgan	2	RF	Não	interna	-	0.7	0.74	0.65	0.68	0.71	0.39	0.39
featmorgan	2	SVM	Não	interna	-	0.79	0.76	0.81	0.8	0.77	0.57	0.57
featmorgan	2	LGB	Não	interna	-	0.78	0.75	0.81	0.8	0.76	0.56	0.56
featmorgan	2	RF	Sim	interna	0.5	0.71	0.73	0.7	0.71	0.72	0.43	0.43
featmorgan	2	SVM	Sim	interna	0.4	0.8	0.84	0.76	0.78	0.83	0.6	0.6
featmorgan	2	LGB	Sim	interna	0.25	0.78	0.8	0.77	0.78	0.79	0.57	0.57
featmorgan	2	RF	Não	externa	-	0.71	0.74	0.68	0.69	0.73	0.42	0.42
featmorgan	2	SVM	Não	externa	-	0.86	0.85	0.87	0.87	0.85	0.72	0.72
featmorgan	2	LGB	Não	externa	-	0.86	0.85	0.87	0.87	0.85	0.72	0.72
featmorgan	2	RF	Sim	externa	0.5	0.7	0.7	0.7	0.7	0.7	0.4	0.4
featmorgan	2	SVM	Sim	externa	0.4	0.84	0.91	0.77	0.79	0.9	0.68	0.69
featmorgan	2	LGB	Sim	externa	0.25	0.87	0.91	0.83	0.84	0.91	0.74	0.74
featmorgan	4	RF	Não	interna	-	0.7	0.72	0.69	0.7	0.71	0.4	0.4
featmorgan	4	SVM	Não	interna	-	0.78	0.78	0.77	0.77	0.78	0.55	0.55
featmorgan	4	LGB	Não	interna	-	0.78	0.79	0.77	0.77	0.78	0.56	0.56
featmorgan	4	SVM	Sim	interna	0.5	0.78	0.78	0.77	0.77	0.78	0.55	0.55
featmorgan	4	LGB	Sim	interna	0.73	0.78	0.78	0.79	0.79	0.78	0.57	0.57
featmorgan	4	RF	Não	externa	-	0.7	0.72	0.68	0.69	0.71	0.4	0.4
featmorgan	4	SVM	Não	externa	-	0.83	0.87	0.79	0.8	0.86	0.66	0.66
featmorgan	4	LGB	Não	externa	-	0.82	0.89	0.74	0.77	0.88	0.63	0.64
featmorgan	4	RF	Sim	externa	0.5	0.69	0.72	0.66	0.67	0.7	0.38	0.38
featmorgan	4	SVM	Sim	externa	0.5	0.83	0.87	0.79	0.8	0.86	0.66	0.66
featmorgan	4	LGB	Sim	externa	0.73	0.8	0.85	0.74	0.76	0.83	0.59	0.6
featmorgan	6	RF	Não	interna	-	0.72	0.71	0.72	0.72	0.71	0.43	0.43
featmorgan	6	SVM	Não	interna	-	0.76	0.77	0.75	0.76	0.77	0.53	0.53
featmorgan	6	LGB	Não	interna	-	0.77	0.82	0.72	0.75	0.8	0.55	0.55
featmorgan	6	RF	Sim	interna	0.5	0.74	0.76	0.71	0.72	0.75	0.47	0.47
featmorgan	6	SVM	Sim	interna	0.39	0.78	0.87	0.69	0.74	0.84	0.56	0.57
featmorgan	6	LGB	Sim	interna	0.28	0.78	0.84	0.72	0.75	0.82	0.56	0.56
featmorgan	6	RF	Não	externa	-	0.71	0.72	0.7	0.7	0.72	0.42	0.42
featmorgan	6	SVM	Não	externa	-	0.82	0.85	0.79	0.8	0.84	0.63	0.64
featmorgan	6	LGB	Não	externa	-	0.81	0.85	0.77	0.78	0.84	0.61	0.62
featmorgan	6	RF	Sim	externa	0.5	0.71	0.72	0.7	0.7	0.72	0.42	0.42
featmorgan	6	SVM	Sim	externa	0.39	0.79	0.93	0.64	0.72	0.91	0.57	0.6
featmorgan	6	LGB	Sim	externa	0.28	0.82	0.89	0.74	0.77	0.88	0.63	0.64
morgan	2	RF	Não	interna	-	0.81	0.82	0.81	0.81	0.82	0.63	0.63
morgan	2	SVM	Não	interna	-	0.8	0.78	0.83	0.82	0.79	0.61	0.61
morgan	2	LGB	Não	interna	-	0.8	0.79	0.82	0.81	0.79	0.61	0.61
morgan	2	RF	Sim	interna	0.49	0.81	0.82	0.81	0.81	0.82	0.63	0.63

Fingerprint	Raio	Algoritmo	Calibrado	Validação	Limiar	Bal-acc	Se	Sp	PPV	NPV	κ	MCC
morgan	2	SVM	Sim	interna	0.49	0.8	0.8	0.81	0.81	0.8	0.61	0.61
morgan	2	LGB	Sim	interna	0.49	0.8	0.8	0.81	0.81	0.8	0.61	0.61
morgan	2	RF	Não	externa	-	0.84	0.93	0.74	0.78	0.92	0.68	0.69
morgan	2	SVM	Não	externa	-	0.85	0.85	0.85	0.85	0.85	0.7	0.7
morgan	2	LGB	Não	externa	-	0.82	0.89	0.74	0.77	0.88	0.63	0.64
morgan	2	RF	Sim	externa	0.49	0.84	0.93	0.74	0.78	0.92	0.68	0.69
morgan	2	SVM	Sim	externa	0.49	0.84	0.85	0.83	0.83	0.85	0.68	0.68
morgan	2	LGB	Sim	externa	0.49	0.82	0.89	0.74	0.77	0.88	0.63	0.64
morgan	4	RF	Não	interna	-	0.82	0.81	0.83	0.83	0.81	0.64	0.64
morgan	4	SVM	Não	interna	-	0.79	0.76	0.82	0.81	0.77	0.57	0.58
morgan	4	LGB	Não	interna	-	0.78	0.85	0.7	0.74	0.83	0.56	0.56
morgan	4	RF	Sim	interna	0.51	0.82	0.8	0.83	0.83	0.81	0.63	0.63
morgan	4	SVM	Sim	interna	0.59	0.81	0.74	0.88	0.86	0.77	0.62	0.63
morgan	4	LGB	Sim	interna	0.96	0.79	0.76	0.81	0.8	0.77	0.57	0.57
morgan	4	RF	Não	externa	-	0.81	0.83	0.79	0.79	0.82	0.61	0.61
morgan	4	SVM	Não	externa	-	0.8	0.78	0.81	0.8	0.79	0.59	0.59
morgan	4	LGB	Não	externa	-	0.84	0.93	0.74	0.78	0.92	0.68	0.69
morgan	2	RF	Sim	externa	0.51	0.81	0.83	0.79	0.79	0.82	0.61	0.61
morgan	4	SVM	Sim	externa	0.59	0.81	0.72	0.89	0.87	0.76	0.61	0.62
morgan	4	LGB	Sim	externa	0.96	0.84	0.85	0.83	0.83	0.85	0.68	0.68
morgan	6	RF	Não	interna	-	0.81	0.81	0.82	0.82	0.81	0.63	0.63
morgan	6	SVM	Não	interna	-	0.78	0.77	0.78	0.78	0.77	0.55	0.55
morgan	6	LGB	Não	interna	-	0.79	0.84	0.74	0.77	0.83	0.58	0.59
morgan	6	RF	Sim	interna	0.5	0.81	0.8	0.82	0.82	0.8	0.62	0.62
morgan	6	SVM	Sim	interna	0.57	0.78	0.73	0.84	0.82	0.76	0.57	0.57
morgan	6	LGB	Sim	interna	0.54	0.79	0.84	0.75	0.77	0.82	0.58	0.59
morgan	6	RF	Não	externa	-	0.81	0.83	0.79	0.79	0.82	0.61	0.61
morgan	6	SVM	Não	externa	-	0.72	0.72	0.72	0.72	0.72	0.44	0.44
morgan	6	LGB	Não	externa	-	0.78	0.91	0.64	0.71	0.88	0.55	0.57
morgan	6	RF	Sim	externa	0.5	0.81	0.83	0.79	0.79	0.82	0.61	0.61
morgan	6	SVM	Sim	externa	0.57	0.73	0.7	0.77	0.74	0.72	0.46	0.46
morgan	6	LGB	Sim	externa	0.54	0.78	0.91	0.64	0.71	0.88	0.55	0.57

Quadro 4: Resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Support Vector Machine* Calibrado Interno.

***Support Vector Machine* (SVM) Calibrado Interno**								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**Kappa**	**Acurácia**
Morgan	2	0,61	0,80	0,81	0,81	0,80	0,61	0,80
	4	0,63	0,74	0,88	0,86	0,77	0,62	0,81
	6	0,57	0,73	0,84	0,82	0,76	0,57	0,78
FeatMorgan	2	0,60	0,84	0,76	0,78	0,83	0,60	0,80
	4	0,55	0,78	0,77	0,77	0,78	0,55	0,78
	6	0,57	0,87	0,69	0,74	0,84	0,56	0,78

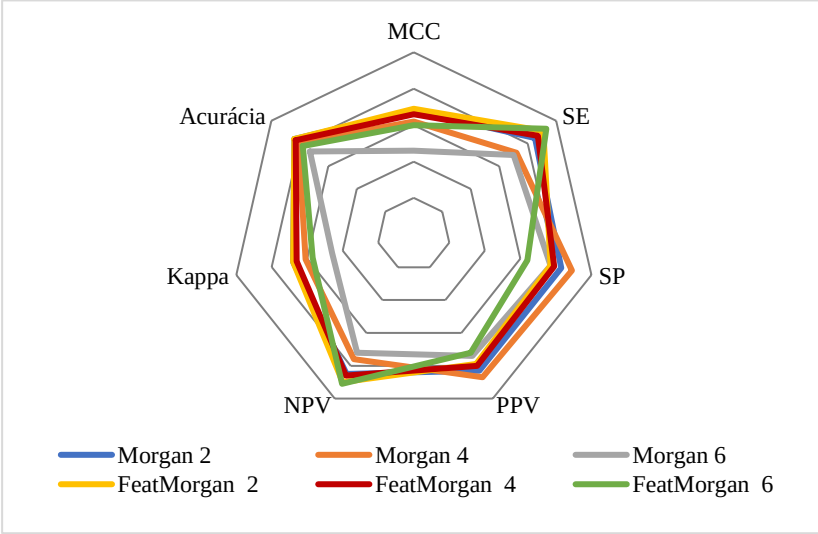
Figura 13: Gráfico dos resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Support Vector Machine* Calibrado Interno.



Quadro 5: Resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Support Vector Machine* Calibrado Externo.

Support Vector Machine (SVM) Calibrado Externo								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**Kappa**	**Acurácia**
Morgan	2	0,68	0,85	0,83	0,83	0,85	0,68	0,84
	4	0,62	0,72	0,89	0,87	0,76	0,61	0,81
	6	0,46	0,70	0,77	0,74	0,72	0,46	0,73
FeatMorgan	2	0,69	0,91	0,77	0,79	0,90	0,68	0,84
	4	0,66	0,87	0,79	0,80	0,86	0,66	0,83
	6	0,60	0,93	0,64	0,72	0,91	0,57	0,78

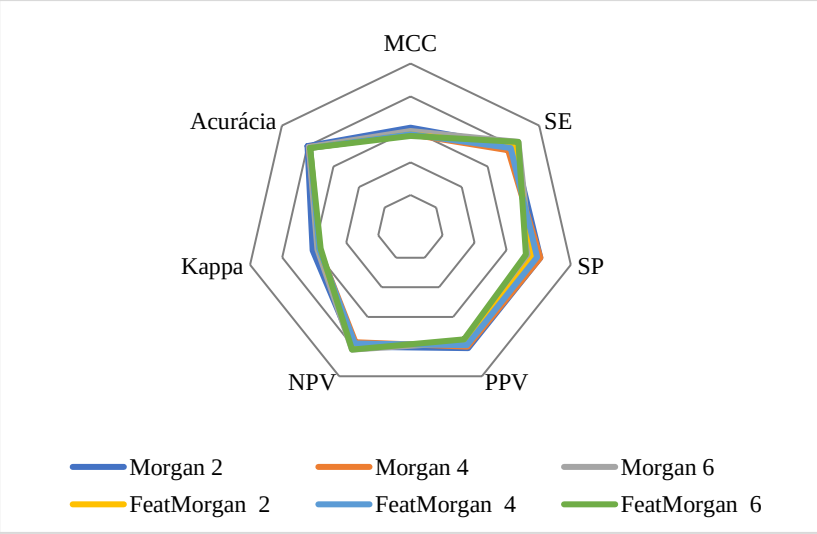
Figura 14: Gráfico dos resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Support Vector Machine* Calibrado Externo.



Quadro 6: Resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Light Gradient Boosting Machine* Calibrado Interno.

***Light Gradient Boosting Machine* (LightGBM) Calibrado Interno**								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**Kappa**	**Acurácia**
Morgan	2	0,61	0,80	0,81	0,81	0,80	0,61	0,80
	4	0,57	0,76	0,81	0,80	0,77	0,57	0,79
	6	0,59	0,84	0,75	0,77	0,82	0,58	0,79
FeatMorgan	2	0,57	0,80	0,77	0,78	0,79	0,57	0,78
	4	0,57	0,78	0,79	0,79	0,78	0,57	0,78
	6	0,56	0,84	0,72	0,75	0,82	0,56	0,78

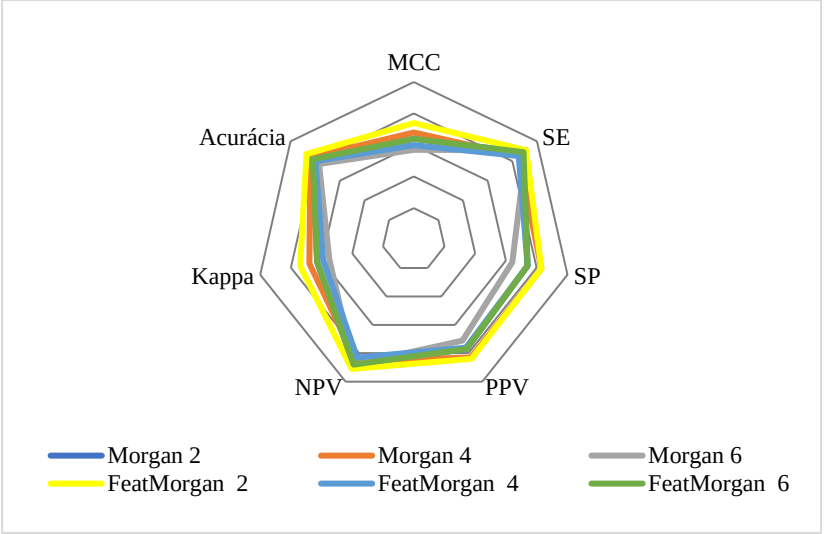
Figura 15: Gráfico dos resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Light Gradient Boosting Machine* Calibrado Interno.



Quadro 7: Resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Light Gradient Boosting Machine* Calibrado Externo.

***Light Gradient Boosting Machine* (LightGBM) Calibrado Externo**								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**Kappa**	**Acurácia**
Morgan	2	0,64	0,89	0,74	0,77	0,88	0,63	0,82
	4	0,68	0,85	0,83	0,83	0,85	0,68	0,84
	6	0,57	0,91	0,64	0,71	0,88	0,55	0,77
FeatMorgan	2	0,74	0,91	0,83	0,84	0,91	0,74	0,87
	4	0,60	0,85	0,74	0,76	0,83	0,59	0,80
	6	0,64	0,89	0,74	0,77	0,88	0,63	0,82

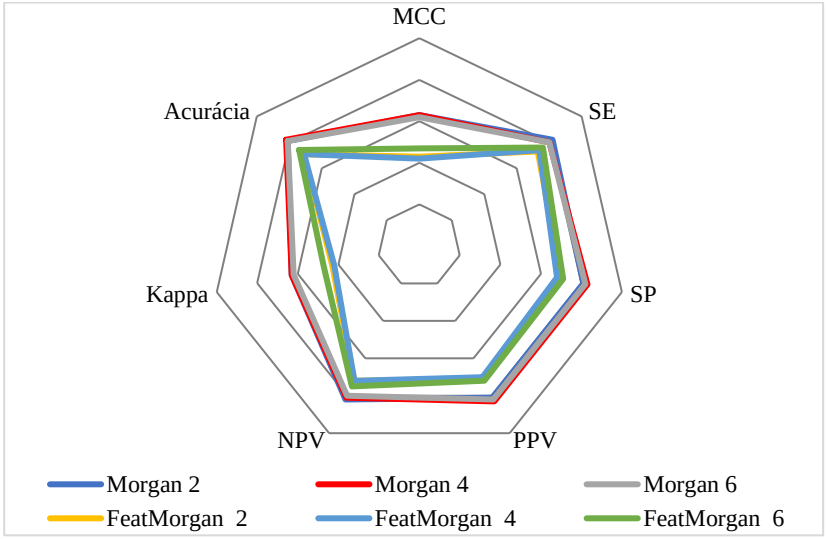
Figura 16: Gráfico dos resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Light Gradient Boosting Machine* Calibrado Externo.



Quadro 8: Resultados estatísticos do modelo de aprendizado de máquina com algoritmo Random Forest Calibrado Interno.

***Random Forest* Calibrado Interno**								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**Kappa**	**Acurácia**
Morgan	2	0,63	0,82	0,81	0,81	0,82	0,63	0,81
	4	0,63	0,80	0,83	0,83	0,81	0,63	0,82
	6	0,62	0,80	0,82	0,82	0,80	0,62	0,81
FeatMorgan	2	0,43	0,73	0,70	0,71	0,72	0,43	0,71
	4	0,42	0,74	0,68	0,70	0,72	0,42	0,71
	6	0,47	0,76	0,71	0,72	0,75	0,47	0,74

Figura 17: Gráfico dos resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Random Forest* Calibrado Interno.



Quadro 9: Resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Random Forest* Calibrado Externo.

***Random Forest* Calibrado Externo**								
Modelo	**Raio**	**MCC**	**SE**	**SP**	**PPV**	**NPV**	**Kappa**	**Acurácia**
Morgan	2	0,69	0,93	0,74	0,78	0,92	0,68	0,84
	4	0,61	0,83	0,79	0,79	0,82	0,61	0,81
	6	0,61	0,83	0,79	0,79	0,82	0,61	0,81
FeatMorgan	2	0,40	0,70	0,70	0,70	0,70	0,40	0,70
	4	0,69	0,72	0,66	0,67	0,70	0,38	0,69
	6	0,42	0,72	0,70	0,70	0,72	0,42	0,71

Figura 18: Gráfico dos resultados estatísticos do modelo de aprendizado de máquina com algoritmo *Random Forest* Calibrado Externo.