UFRRJ

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

Uso de algoritmos de aprendizado de máquina para a busca de modelos de previsão da atividade inibitória da enzima N-miristoiltransferase de *Leishmania donovani*

Soraya de Oliveira Bandeira

2024

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA A BUSCA DE MODELOS DE PREVISÃO DA ATIVIDADE INIBITÓRIA DA ENZIMA N-MIRISTOILTRANSFERASE DE *LEISHMANIA DONOVANI*

SORAYA DE OLIVEIRA BANDEIRA

Sob orientação de
Marcelo Dib Cruz

e co-orientação de
Carlos Mauricio Rabello de Sant'Anna

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre** no Programa de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Computacional.

Seropédica, RJ, Brasil
Março de 2024

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

Bandeira, Soraya de Oliveira, 10/06/1984-
B111u Uso de algoritmos de aprendizado de máquina para a busca de modelos de previsão da atividade inibitória da enzima N-miristoiltransferase de Leishmania donovani / Soraya de Oliveira Bandeira. - Niterói, 2024.
112 f.: il.

Orientador: Marcelo Dib Cruz.
Coorientador: Carlos Mauricio Rabello Sant'Anna.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, PPGMMC, 2024.

1. Aprendizado de Máquina. 2. QSAR. 3. Modelagem molecular. 4. Leishmania Donovani. I. Cruz, Marcelo Dib , 05/11/1967-, orient. II. Sant'Anna, Carlos Mauricio Rabello , 13/11/965-, coorient. III Universidade Federal Rural do Rio de Janeiro. PPGMMC. IV. Título.

MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

Seropédica-RJ, 19 de março de 2024.

Soraya de Oliveira Bandeira

Dissertação submetida como requisito parcial para a obtenção de grau de **Mestra**, no Programa de Pós-Graduação em Modelagem Matemática e Computacional PPGMMC, área de Concentração em Modelagem Matemáticae Computacional.

DISSERTAÇÃO APROVADA EM 19/03/2024

Marcelo Dib Cruz Drº UFRRJ (Orientador, Presidente da Banca)

Felipe Leite Coelho da Silva Drº UFRRJ (membro interno)

Marcello Montillo Provenza Drº UFRRJ (Externo à Instituição)

ATA Nº ata/2024 - ICE (12.28.01.23)
(Nº do Documento: 1043)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 04/04/2024 16:39)
FELIPE LEITE COELHO DA SILVA
PROFESSOR DO MAGISTERIO SUPERIOR
DeptM (12.28.01.00.00.00.63)
Matrícula: ###398#2

(Assinado digitalmente em 02/04/2024 16:06)
MARCELO DIB CRUZ
PROFESSOR DO MAGISTERIO SUPERIOR
DCOMP (11.39.97)
Matrícula: ###680#1

(Assinado digitalmente em 02/04/2024 15:40)
MARCELLO MONTILLO PROVENZA
ASSINANTE EXTERNO
CPF: ###.###.147-##

Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: ***1043****, ano:* ***2024****, tipo:* ***ATA****, data de emissão:* ***02/04/2024*** *e o código de verificação:* ***fba7baa2f4***

Agradecimentos

Agradeço ao meu Criador por abrir portas como esta em minha vida e cuidar de mim.

Aos meus pais, parentes, noivo e gatos, por entenderem a minha ausência em muitos momentos e serem o meu conforto em tempos difíceis.

Aos professores D.Sc. Marcelo Dib Cruz e D.Sc. Carlos Maurício Rabello de Sant Anna que como orientador e coorientador me auxiliaram com toda paciência e dedicação.

Ao meu noivo Igor Campos de Almeida Lima, que me incentivou a ingressar e concluir o curso de mestrado. Ele presenciou cada etapa e me auxiliou com os códigos do R, receba toda a minha admiração.

Aos professores D.Sc. Robson Mariano Silva (UFRRJ), M.Sc. Paulo Henrique Couto Simões (UERJ), D.Sc. Marcello Montillo Provenza (UERJ), e D.Sc. Janaína dos Santos Nascimento (IFRJ), que agregaram conhecimentos em trabalhos e experiências.

Ao professor D.Sc. Aderval Severino Luna (Depto. de Química Analítica/IQ) e ao PPgEQ/UERJ que me aceitaram na disciplina de aprendizado de máquina.

Aos meus colegas do curso que participaram juntamente comigo dessa etapa.

A UFRRJ e a diretoria do PPGMMC que proporcionaram ferramentas para chegar até este momento, mesmo em época de Pandemia.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

"Comece fazendo o que é necessário,
depois o que é possível, e de repente
você estará fazendo o impossível!"
(São Francisco de Assis)

RESUMO

DE OLIVEIRA BANDEIRA, Soraya. **Uso de algoritmos de aprendizado de máquina para a busca de modelos de previsão da atividade inibitória da enzima N-miristoiltransferase de *Leishmania donovani*** . 2024. 112f. Dissertação (Mestrado em Modelagem Matemática e Computacional). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

A leishmaniose é uma doença negligenciada (DN) e representa um desafio global. As empresas farmacêuticas investem pouco na fabricação de medicamentos para DNs, porque o retorno é pequeno e essas doenças afetam principalmente as regiões de baixo poder aquisitivo . Os medicamentos usados no tratamento contra leishmaniose são administrados com base na inibição de outras doenças e tem efeitos colaterais. O campo da modelagem molecular tem se mostrado eficiente para o desenvolvimento de novos medicamentos, quanto a recurso e tempo. Este trabalho tem como objetivo aplicar o método QSAR (Quantitative Structure Activity) para a modelagem molecular de estruturas sintetizadas e promissoras no combate a *Leishmania donovani*, juntamente com uma análise de métodos lineares e não lineares de aprendizado de máquina em caráter inibitório do protozoário, o melhor modelo foi aquele com o menor RMSE, nas etapas de calibração e validação cruzada leave-one-out, sendo este o método de regressão linear múltipla (MLR).

Palavras-chave: Leishmaniose, *Leishmania donovani*, QSAR.

ABSTRACT

DE OLIVEIRA BANDEIRA, Soraya. **Using machine learning algorithms to search for models prediction of enzyme inhibitory activity of the N-myristoyltransferase of *Leishmania donovani* .** 2024. 112p. Dissertation (Master in Mathematical and Computational Modeling). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

Leishmaniasis is a neglected disease (ND) and represents a global challenge. The companies pharmaceutical companies invest little in the manufacture of medicines for EDs, as the return is small and these diseases mainly affect regions with low purchasing power. The medicines used to treat leishmaniasis are administered based on the inhibition of other diseases and has side effects. The field of molecular modeling has proven to be effective for the development of new medicines, in terms of resources and time. This work has as objective apply the QSAR (Quantitative Structure Activity) method for molecular modeling of synthesized and promising structures in the fight against *Leishmania donovani*, together with an analysis of linear and non-linear machine learning methods in inhibitory character of the protozoan, the best model was the one with the lowest RMSE, in the calibration stages and leave-one-out cross validation, this being the multiple linear regression (MLR) method.

Keywords: Leishmaniasis, *Leishmania donovani* , QSAR.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações e Siglas

CGIs	Interface comum de ligação
DN	Doenças negligenciadas
LC	Leishmaniose cutânea
LT	Leishmaniose tegumentar
LV	Leishmaniose visceral
NMT	N-miristoiltransferase
QSAR	Relação quantitativa de estrutura-atividade
pBA	Porcentagem de atividade biológica
AM	Aprendizado de máquina
Ld	*Leishmania donovani*
ANVISA	Agência nacional de vigilância sanitária
ASP	Potencial estatístico Astex
CDC	Centro de dados cristalográficos de Cambridge
DE.tors	Peso do potencial de torção do ligante
Esolv	Energia de solvatação
GOLD	Otimização genética na docagem do ligante
HOMO	Orbital molecular ocupado de maior energia
IC_{50}	Concentração do inibidor necessária para inibir in vitro a atividade enzimática em 50%
KI	Dissociação do complexo enzima-inibidor
LUMO	Orbital molecular não ocupado de menor energia
MMFF	Campo de Força Molecular Merck
OMS	Organização mundial de saúde
PDB	Banco de dados de proteínas

PLS	Regressão dos mínimos quadrados parciais
PM6	Método paramétrico 6
Q^2	Coeficiente de correlação da validação cruzada
R^2	Coeficiente de correlação ao quadrado
RLM	Regressão linear múltipla
RMSD	Desvio quadrático médio da raiz
RMSEP	Raiz quadrada do erro média de previsão
S.Hbond	Termo da ligação de hidrogênio entre a proteína
S.int	Soma dos termos de torção interna e de van der Waals interno
SBDD	Droga baseada na estrutura
SEP	Erro-padrão de predição
Spress	Desvio-padrão da validação cruzada

Lista de Símbolos

Q^2	coeficiente de correlação da validação cruzada
R^2	Coeficiente de correlação de determinação múltipla
R^2_{ext}	Coeficiente de correlação de determinação múltipla externa
Q^2_{ext}	Coeficiente de validação de correlação externa
Q	Coeficiente de validação de correlação de Pearson
W	Média dos valores experimentais
y_e	Valores experimentais de y
y_c	Valores calculados de y, ou seja, valores de calibração
y_p	Valores previstos de y, ou seja, valores do conjunto de Validação externa
y_v	Valores calculados de y a partir de uma validação interna (LOO, LNO ou y-randomização), y_c
s	Desvio padrão amostral
CV%	Coeficiente de variação
Mín	Mínimo
Q_1	1º Quartil
$\bar{x}$	Média aritmética
$\tilde{x}$	Mediana
Q_3	3º Quartil
Máx	Máximo

1 Introdução

A leishmaniose é uma patologia que ocorre principalmente em países tropicais, causada por parasitas protozoários do gênero *Leishmania*. A leishmaniose é um desafio mundial, pois esta incluída no grupo de doenças negligenciadas (DNs), que afetam 15 % das pessoas no mundo [1]. De acordo com Santos, 2020, há uma dificuldade na confecção de novos fármacos para os tratamentos, muitas empresas não tem interesse de investir em pesquisas por conta do baixo retorno de receita e o alto custo na descoberta e desenvolvimento de novos medicamentos quimioterápicos para essas DNs, sendo improvável de se desenvolver [1]. As características dos sintomas da leishmaniose podem ser semelhantes a outras doenças comuns, como febre tifoide e tuberculose, por isso os métodos laboratoriais confiáveis são necessários para um diagnóstico diferencial. O exame parasitológico (histopatologia, microscopia e cultura parasitária), sorologia e diagnóstico são testes aplicados para a identificação, mas a presença de características semelhantes das demais doenças são desvantagens no reconhecimento da doença [2].

O campo da modelagem molecular tem se mostrado atrativo para o desenvolvimento de novos fármacos. A descoberta de moléculas bioativas que possam combater a leishmaniose é um processo caro e demorado, novas estratégias são continuamente buscadas para otimizar esse processo. A Triagem virtual (VS) é uma das estratégias recentes que têm sido exploradas para a identificação de moléculas bioativas candidatas. Através de bibliotecas virtuais é possível selecionar em bancos de dados com estruturas químicas candidatas a fármacos, que ajudam a amenizar o crescimento da doença e juntamente com a evolução de equipamentos computacionais a pesquisa in sílico mostrou-se uma alternativa viável no entendimento na pesquisas de novos fármacos oferecendo redução do tempo e de recursos financeiros.

Entre as opções de simulação, a modelagem molecular tem facilitado o desenvolvimento de novos fármacos. Considera-se que existe uma relação entre as propriedades de uma molécula, sua estrutura química e sua atividade biológica, então, buscam-se estabelecer relações matemáticas simples para descrever e prever a atividade de um conjunto de inibidores semelhantes [3] e [4].

O desenho de novas moléculas bioativas pode ser realizado por vários métodos diferentes, um exemplo é a relação de atividade de estrutura quantitativa (QSAR), que relaciona os dados de atividade biológica ou farmacológica com à estrutura química. A geometria dos modelos requer a representação de um conjunto ou matriz de dados incluindo a medida quantitativa da atividade biológica, os parâmetros físico-químicos e estruturais que descrevem as propriedades dos compostos químicos. Ou seja, o conjunto de dados contém os valores da atividade biológica e variáveis descritivas referentes a compostos gerados [5]

O aprendizado de Máquina, em inglês *"Machine Learning"* é um ramo da inteligência

artificial que estuda o desenvolvimento de algoritmos capazes de classificar objetos a partir de um dado conjunto de treinamento, através de estratégias de métodos matemáticos, estatísticos e computacionais de forma otimizada, semelhante a quimiometria onde os objetos normalmente são estruturas químicas e as classes propriedades inerentes a estas.

2 Fundamentação Teórica

2.1 leishmaniose

A leishmaniose é considerada uma doença negligenciada e é mais encontrada em países tropicais. Sua causa se deve a protozoários pertencentes a mais de 20 espécies de gênero *leishmania*, que são responsáveis pelas leishmanioses, ou seja, essa doença é um amplo espectro de doenças, que afeta à saúde pública. De acordo com a OMS - Organização Mundial da Saúde em Janeiro de 2023 estimou-se que 700.000 a 1 milhão de novos casos ocorrem anualmente [6].

Os três principais tipos são leishmaniose tegumentar (LT), leishmaniose visceral (LV) ou calazar e leishmaniose mucocutânea, que são transmitidas pelos insetos hematófagos conhecidos como flebótomos ou flebotomíneos [2].

2.1.1 Leishmaniose Visceral

A leishmaniose visceral que é abordada neste trabalho, tem sua transmissão causada por um dos protozoários chamado *Leishmania donovani*, a mesma pode apresentar sintomas iniciais de febre irregular prolongada; anemia; indisposição; falta de apetite; perda de peso e inchaço do abdômen devido ao aumento do fígado e do baço [7].

As características da leishmaniose podem ser semelhantes a outras doenças comuns, como febre tifoide, malária e tuberculose e onde métodos laboratoriais confiáveis são necessários para um diagnóstico diferencial. O exame parasitológico (histopatologia, microscopia e cultura parasitária), sorologia e diagnóstico são testes aplicados para a identificação da leishmaniose, mas a presença de características semelhantes das demais doenças são desvantagens na identificação [2].

De acordo com [8] e [26], anualmente surgem 500 mil novos casos de infecção que são relatados principalmente em comunidades carentes, sendo o reposicionamento de novos medicamentos uma estratégia ideal para combater esses parasitas.

Em 2020 [10] realizaram um estudo sobre as funções que identificam algumas proteínas citosólicas desconhecidas de *Leishmania donovani* através do classificador *Random Forest* (RF), os resultados obtidos no estudo indicaram que o método tem precisão e significância, sendo um alvo promissor para o desenvolvimento de candidatos a vacinas para descoberta de novos medicamentos terapia de LV fatal e que existem várias proteínas desconhecidas, hipotéticas que foram caracterizadas em detalhes durante os últimos anos e que desempenham um papel dinâmico ao parasita no processo celular e propagação da infecção. Essas proteínas são alvos importantes para o desenvolvimento de terapêuticas com funções específicas como sobre-

vivência intracelular, patogenicidade, resistência a fármacos, e controle da doença. Descobrir a função molecular é um percurso demorado, que consome tempo e recursos, por isso a utilização computacional diminui o tempo a ser gasto através de algoritmo [10]. De acordo com [2], a leishmaniose pode ser prevenida com o controle em contato com o vetor principal que seria flebotomíneos infectados com o protozoário e não existem estudos publicados sobre a eficácia de diagnóstico para prevenção e tratamento de doenças.

Os tratamentos disponíveis para leishmaniose visceral são inadequados e há uma necessidade premente de novas terapêuticas. Os esforços de descoberta de medicamentos para ambas as doenças dependem principalmente da triagem fenotípica. No entanto, a otimização de compostos fenotipicamente ativos é prejudicado pela falta de informação sobre seu(s) alvo(s) molecular(es) [11].

De acordo com dados da Secretaria de Vigilância de Saúde, publicados em 2020, os casos de leishmaniose Visceral tiveram um aumento significativo entre o início da série histórica registrada no ano 2000 e a partir de 2003 a ocorrência manteve-se mais ou menos constante (figura 2.1). Quando os casos são estratificados por regiões, a série histórica (figura 2.2) mostra que a região nordeste apresenta o maior índice de novos casos registrados. Por conta deste cenário escolheu-se a LV em especial para este e trabalho, pois ela se mostrou mais emergente frente a situação do crescimento da doença.

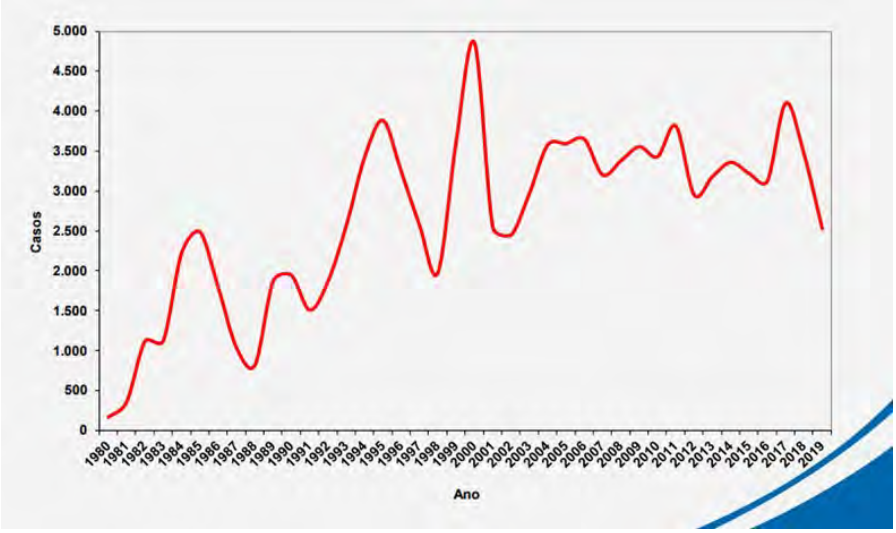


Figura 2.1 – Casos de leishmaniose Visceral no Brasil, 1980 a 2019
Fonte: Secretaria de Vigilância de Saúde - 2020.

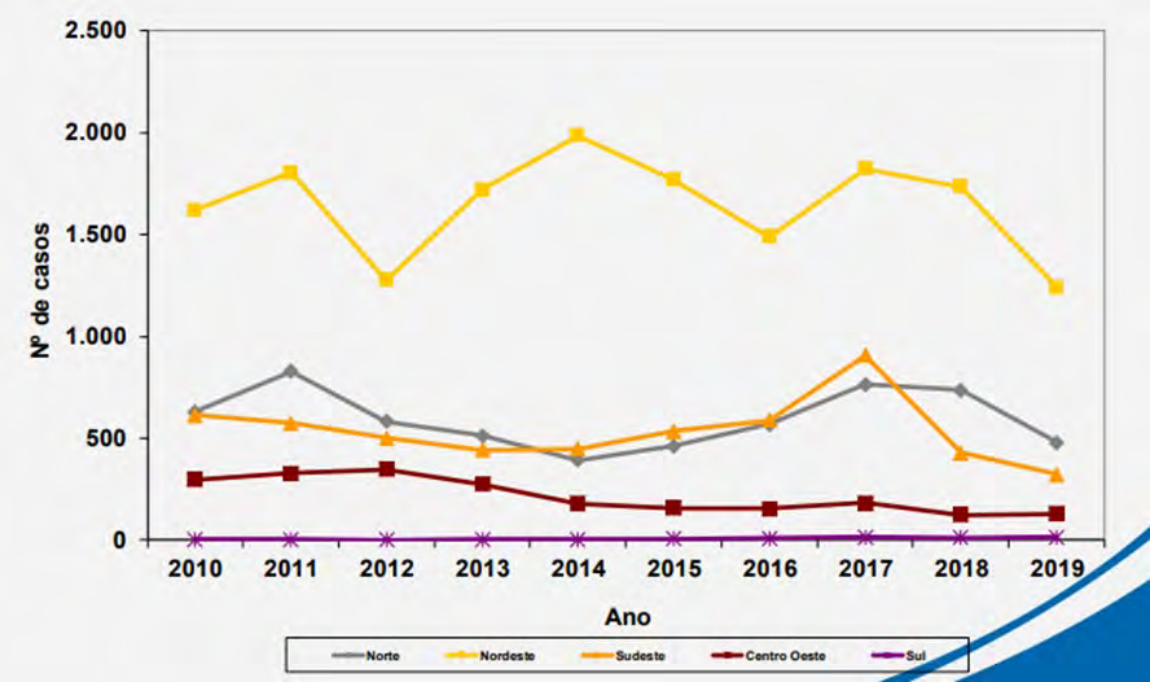


Figura 2.2 – Casos de leishmaniose Visceral por região no Brasil, 2010 a 2019
Fonte: Secretaria de Vigilância de Saúde - 2020.

Em 2020, de acordo com a OMS, o Brasil representou mais de 97% dos casos de LV na região das Américas.[6] Na resolução WHA60.13 da Organização Mundial de Saúde (*World Health Organization*, o Brasil destaca-se negativamente tanto para LT quanto LV, figurando entre as regiões do mundo com maior número de novos casos de leishmaniose [12]. Segundo o Ministério da Saúde [122], em 2019 o Brasil contabilizou 2.529 casos de leishmaniose, e no seu site retirado em fevereiro de 2024 informam que a média de casos é de 3500 de leishmaniose visceral no Brasil. A Fundação Oswaldo Cruz (Fiocruz) em 2021 divulgou no seu portal que de março a outubro de 2021 foram registrados 2062 casos suspeitos de leishmaniose Visceral Canina.[13]

> "O cenário é preocupante, por isso organizações e instituições ligadas à Saúde promovem o 'Agosto Verde' – mês dedicado ao combate e prevenção desta enfermidade que acomete tanto seres humanos quanto cães e gatos."Fonte [14]

2.1.1.1 *Leishmania donovani*

A ocorrência anual de leishmaniose visceral é de 500 mil casos de infecção no mundo, sendo uma doença fatal se não for tratada. O reposicionamento de novos medicamentos é uma estratégia ideal para combater esses parasitas [26]. Entre esses protozoários, a *Leishmania donovani*, que é um dos agentes causador da leishmaniose visceral, ela mais resistente ao óxido nítrico (NO) e ao peróxido de hidrogênio do que *Leishmania major* [9].

2.2 Transmissão

O ciclo de transmissão ocorre quando o mosquito fêmea realiza o repasto no ser hospedeiro infectado, causando formas amastigotas do protozoário que seguem para o intestino

anterior do inseto. Nesse processo ocorrem reações bioquímicas e morfológicas para sobrevivência dentro do hospedeiro e posteriormente se dá a infecção [15].

No trato digestivo do vetor, o parasita assume uma forma promastigota pro-cíclica, onde se evita a interrupção no intestino do mosquito e posteriormente a forma promastigota metacíclica infectante, chegando nas partes bocais do mosquito, que de acordo com [16], o mesmo sobrevive por meio de açúcares retidos no sangue, que favorecem a sobrevivência da doença dentro do inseto.

O flebotomíneo fêmea inocula através de sua saliva as formas promastigotas no hospedeiro vertebrado, em seguida, essas formas penetram nas células fagocítico mononuclear local, transformando-se em amastigotas e se multiplicam através divisão longitudinal binária. Dentro do macrófagos do ser vertebrado são gerados e se reinicia um novo ciclo morfológico para sobrevivência dentro do hospedeiro e posteriormente se dá a infecção [15, 17].

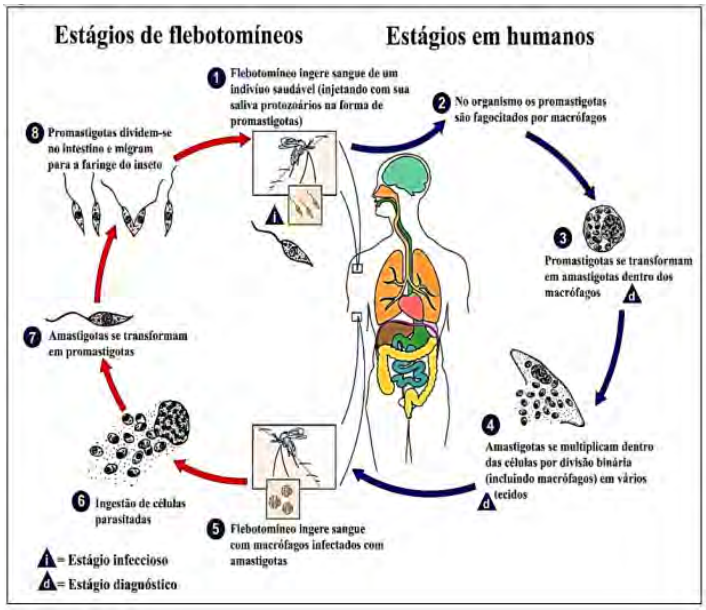


Figura 2.3 – Ciclo de vida de *Leishmania*em flebotomíneo e em humano
Fonte:Adaptado de CDC (https://www.cdc.gov/dpdx/leishmaniasis/index.html).

2.3 Tratamento

Há uma complexidade no consumo dos fármacos que combatem as doença DN, o exemplo disso são os efeitos colaterais por conta da toxidade ao organismo e a resistência aos medicamentos e o alto custo[18].

O tratamento de medicamentos para as doenças negligenciadas pela OMS não tem recurso de quimioterapia direta, por isso a importância de criação de fármacos viáveis é algo promissor para humanos que vivem em populações pobres e negligenciadas. Os esforços de descoberta de medicamentos para as doenças negligenciadas dependem principalmente de uma triagem fenotípica. No entanto, a otimização de compostos fenotipicamente ativos é prejudicado pela falta de informação sobre seu(s) alvo(s) molecular(es)[11].

2.3.1 Tratamento para leishmaniose

Não existe vacina contra a leishmaniose humana, somente para canina. A medida mais utilizada é a prevenção e o combate da doença, que se baseia no controle de vetores e dos reservatórios, diagnóstico precoce, proteção individual, tratamento dos doentes, manejo ambiental e educação em saúde[19].

Diferente de outras leishmanioses, a LV costuma ser letal se não tratado. O SQ109, é um composto em fase IIb/III de ensaios clínicos para tratar *Mycobacterium tuberculosis* e tem um potente efeito inibitório sobre o crescimento de *Leishmania donovani*, mas a atividade de dele na porcentagem de infecção de amastigotas de *Leishmania donovani* em macrófagos murinos mostraram que o composto afeta a porcentagem de macrófagos infectados de maneira dose-dependente aproximadamente 100 vezes mais do que seus efeitos sobre promastigotas, [**??**]. Entre outros, a população de macrófagos hospedeiros por *Leishmania* também difere entre espécies cutânea e visceral . Parasitas viscerotrópicos infectam células de Kupffer, baçomacrófagos e macrófagos da medula óssea, enquanto os parasitas cutâneos infectam macrófagos e células dendríticas [9]. O tratamento é bem complexo, pois tem diferentes manifestações de espécies de *Leishmania*[20]. Uma proposta para a dificuldade no tratamento da leishmaniose é apresentada em [2]:

> " O tratamento disponível para leishmaniose está sobrecarregado com resistência a alguns dos medicamentos atualmente disponíveis. O mecanismo de resistência aos fármacos está frequentemente relacionado à menor absorção de fármacos, aumento do fluxo, taxa rápida de metabolismo do fármaco, modificações do fármaco alvos e sobre-expressão de transportadores de fármacos."

Entre os fármacos disponíveis para tratamento, os principais são antimoniato pentavalente, anfotericina B e miltefosina.

O antimoniato pentavalente é um fármaco com antimoniais, que pode causar efeitos adversos como arritmia cardíaca, pancreatite, hepatotoxicidade e nefrotoxicidade [21], fora que não são recomendados para mulheres grávidas devido as limitações como administração parenteral [22].

A anfotericina B é um fármaco antifúngico sendo usado como tratamento de segunda linha [24]. A miltefosina é o primeira fármaco oral para o tratamento da LV, mas ela tem alguns efeitos colaterais como nefrotoxicidade, hepatotoxicidade e teratogênico [25].

Por conta desses efeitos colaterais e a resistência de medicamentos, é que se faz necessário novos estudos que possam aumentar a atividade da fármaco, ação sinérgica, diminuir a duração e dosagem, minimizar custo no tratamento, reduzir os efeitos colaterais [26].

Existem várias proteínas desconhecidas e hipotéticas que foram caracterizadas em detalhes durante os últimos anos onde desempenham um papel dinâmico ao parasita no processo celular e na propagação da infecção.

Essas proteínas são alvos importantes para o desenvolvimento de terapêuticas com funções específicas como sobrevivência intracelular, patogenicidade, resistência a fármacos, e controle da doença. Descobrir uma estrutura química promissora é um percurso demorado, que consome tempo e recursos, por isso a utilização computacional reduz o tempo dedicado ao processo de descoberta [10].

2.4 N-Miristoiltransferase

A miristoilação é a acilação de uma determinada glicina da porção N-terminal de proteínas. Seu processo ocorre em proteínas de eucariotos pela ligação covalente de um miristato, um ácido graxo de 14 carbonos, à glicina da porção N-terminal. Proteínas N-miristoiladas possuem diversos destinos intracelulares e o miristoil possui um papel crítico nas mediações de interações proteína-proteína e proteína-membrana. Em outras proteínas N-miristoiladas, o ancoramento de um ligante produz uma mudança conformacional que expõe ou sequestra a cadeia acil [27].

O ciclo catalítico da N-miristoiltransferase (NMT), que é uma enzima, envolve um mecanismo sequencial de catálise. A NMT de proteínas catalisa a ligação covalente do ácido mirístico no resíduo de glicina amino-terminal após a remoção da metionina de iniciação. A N-miristoilação desempenha um papel importante nas interações proteína-proteína - proteínas na proteína da membrana que, por sua vez, facilitam uma variedade de vias de transdução de sinal. NMT é essencial no ciclo de vida da *Leishmania donovani*, sendo promissora como alvo de fármacos para o tratamento de LV [3]

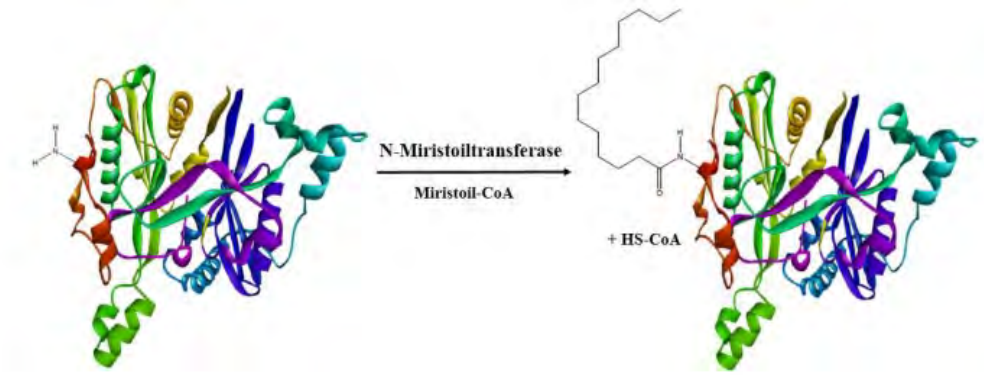


Figura 2.4 – Miristoilação
Fonte:Adaptado de [7].

2.5 Modelagem Molecular

Modelagem molecular, de acordo a IUPAC, é o estudo das estruturas e das propriedades moleculares pela aplicação da química computacional e técnicas de visualização gráfica, visando fornecer uma representação espacial, incluindo um conjunto de circunstâncias [29]. Com a modelagem molecular é possível adiantar a criação de novos fármaco, através da criação virtual 2D ou 3D pode se identificar relação entre as propriedades de uma molécula, sua estrutura química e sua atividade biológica, então, buscam-se estabelecer relações matemáticas simples para descrever e prever um conjunto de estruturas com atividade semelhante [7] e [4]. Trata-se de um conjunto de recursos matemáticos e computacionais que possibilita a análise de sistemas moleculares, SANT'ANNA, 2009. Essa interação de dados da modelagem com a inibição do agente, através novas moléculas bioativas, pode ser realizado por vários métodos diferentes, um dos quais é a relação de atividade de estrutura quantitativa (QSAR), que relaciona os dados de atividade biológica com à estrutura química. A geometria dos modelos requer a representação de um conjunto ou matriz de dados incluindo a medida quantitativa da atividade biológica e os parâmetros físico-químicos e estruturais que descrevem as propriedades dos compostos químicos. Ou seja, o conjunto de dados contém os valores da atividade biológica e

variáveis descritivas referentes a compostos gerados [5], sendo um método muito utilizado em quimiometria, onde os objetos normalmente são moléculas e as classes são os descritores.

2.6 Triagem Virtual

O campo da modelagem molecular tem se mostrado atrativo para o desenvolvimento de novos fármacos. A descoberta de moléculas bioativas que possam combater a leishmaniose é um processo caro e demorado, novas estratégias são necessárias para aperfeiçoar esse processo. A Triagem virtual (VS) é uma das estratégias recentes que têm sido exploradas para a identificação de moléculas bioativas candidatas, [35]. Através de bibliotecas virtuais é possível selecionar em bancos de dados estruturas químicas de remédios que ajudam a amenizar o crescimento da doença e juntamente com a evolução de equipamentos computacionais a pesquisa *in silico* mostrou-se uma alternativa viável no entendimento nas pesquisas de novos fármacos oferecendo redução do tempo e de recursos financeiros [35].

O site do Protein Data Bank (PDB), oferece um banco de dados de acesso público, deste banco foi realizada uma triagem virtual para este trabalho, com a seleção do código de estrutura 3D referente ao alvo [36]. A estrutura de combate, cristalográfica, selecionada para este trabalho foi com o código 2wuu, estrutura da N-miristoiltransferase da *L. donovani* [37] que apresenta todo o modelo da enzima . De acordo com [49], o PDB pode apresentar dados de experimentos específicos incertos, sendo necessário uma inspeção minuciosa, o que demanda tempo e análise em laboratório, por este motivo as amostras deste trabalho foram retiradas de [43], onde há detalhes sobre a padronização dos experimentos, visto que são estruturas sintetizadas, sendo diferente do modelo alvo, que para este trabalho só precisa da estrutura 3D na realização da docagem.

2.7 Relações Quantitativas Estrutura - Atividade(QSAR)

O QSAR 3D permite explorar os tipos de liberdade conformacional e diferentes alinhamentos na busca de uma conformação representativa do modo de interação dos farmacóforos.

O QSAR 4D envolve o cálculo de simulação de dinâmica molecular (SDM) para cada molécula e, assim, são gerados um enorme número de confôrmeros [38]. Estes confôrmeros são utilizados para obter um perfil de amostragem conformacional (CEP, do inglês *Conformational Ensemble Profile*). Assim, o diferencial deste método é que ele incorpora a flexibilidade conformacional do ligante. Cada conformação do CEP de cada molécula é então colocada em uma caixa tridimensional virtual, onde é definido o alinhamento [38]. O alinhamento é a maneira pela qual os compostos do conjunto de treinamento são comparados, onde são escolhidos de modo a abranger uma estrutura em comum dos compostos dos grupos treinamento e teste. Alinhamentos usando átomos da direita, esquerda e do meio, ou que usam átomos que abrangem toda a região similar dos ligantes devem ser usados para garantir um bom alinhamento. Na tabela 2.1 tem a exemplificação do tipo estrutura e quais descritores estão inseridos.

Tabela 2.1 – Descritores em relação a modelagem, adaptado de [28]

Representação Molecular	**Descritores**
0D	Peso molecular, número de átomos, número de ligações, soma de propriedades atômicas.
1D	Número de fragmentos.
2D (descritores topológicos)	Índice de Zagreb, Wiener, descritores BCUT, Vetor de autocorrelação.
3D(descritores geométricos)	Descritores 3D - MoRSE, descritores WHIM, vetores de autocorrelação 3D.
3D (Propriedades de superfície)	Análise comparativa dos campos moleculares (CoMFA).
4D	Coordenadas 3D e amostragem de conformações, valores da energia de interação, descritores GRIND.

O desenvolvimento de um modelo QSAR pode ser compreendido em 3 etapas principais, que serão relacionadas nos subitens abaixo:

2.7.1 Geração de Descritores

Um descritor molecular é chamado também como um descritor químico, onde sua representação matemática é alguma característica da molécula [39, 41]. Relação quantitativa estrutura-atividade (QSAR) também correlaciona a estrutura dos ligantes, por meio de descritores moleculares, com suas atividades biológicas.

As informações codificadas pelos descritores geralmente dependem do tipo de representação molecular e o algoritmo definido para o cálculo. Alguns descritores identificam a massa molecular, a topologia (índices de conectividade), a geometria, a parte físico-química (exemplo calor de formação) ou a parte eletrônica (exemplo é o momento dipolo). [39, 41]. Existem muitos programas que estão disponíveis na literatura que podem ser utilizados para fazer modelagem. Alguns conhecidos são: MobyDigs, BuildQSAR, VCCLAB, QSAR+, BILIN, MOLGEN QSPR, CORAL, CODESSA PRO, WOLF[42], SPARTAN e GOLD.

2.7.2 Descritores x atividade biológica

A metade da concentração inibitória máxima IC_{50}, é a medida da potência de uma substância na inibição de uma função biológica ou bioquímica específica [43]. A etapa seguinte consiste em correlacionar os descritores obtidos com os dados da atividade biológica, construindo assim um modelo matemático. Quando essas correlações envolvem valores experimentais de atividade, por exemplo, o IC_{50}, um modelo de regressão pode ser usado. No entanto, quando as correlações envolvem informações categóricas, como, por exemplo, ativo, inativo, tóxico, não tóxico, etc., as abordagens de classificação são usadas [44]. Os modelos quimiométricos gerados são utilizados na identificação de propriedades moleculares relevantes, na predição de parâmetros de atividade e de propriedades farmacocinéticas de novos compostos e na proposição de novas estruturas para síntese e avaliação biológica.

A figura **??** apresenta um exemplo de matriz de amostras e descritores, onde as amostras (no caso, as moléculas obtidas) são alocadas nas linhas enquanto as variáveis (os descritores) são alocados nas colunas, gerando matrizes de tamanho $n \times p$.

Os descritores tipicamente utilizados em QSAR podem ser agrupados em classes relaci-

onadas à representação molecular. A tabela 2.1 ilustra, mas não esgota, os descritores utilizados em QSAR.

2.7.3 Validação

Validação em QSAR desempenha um papel para aplicação da previsão de novos compostos. As validações são divididas em validação interna e validação externa. A fim de realizar estas validações o conjunto de dados é dividido em um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento é composto pelas moléculas que serão utilizadas na construção dos modelos e participam da validação interna dos modelos. O conjunto de teste não participa da construção dos modelos de QSAR [29].

2.7.3.1 Validação cruzada - interna

Na validação cruzada, o conjunto de dados é dividido em grupos de x tamanhos e vários modelos são gerados, restando um desses grupos de fora do modelo. O modelo de regressão obtido é usado para prever a variável dependente (atividade biológica ou propriedade físico-química) das amostras deixadas de fora da análise. Esse processo é repetido até que todas as amostras tenham sido excluídas da análise uma vez.

2.8 Docagem

A "docagem"refere-se ao processo de acoplar uma molécula a um sítio alvo em uma proteína ou receptor biológico. O termo "docagem"é muitas vezes utilizado de forma intercambiável com *"docking"* em inglês [34].

O QSAR é uma abordagem computacional usada na área de química para entender e prever as relações entre a estrutura química de compostos e suas atividades biológicas. A docagem é uma etapa que envolve a predição da conformação espacial otimizada de uma molécula em relação ao sítio de interação da enzima ou proteína alvo [34], mas nem sempre é utilizada em QSAR.

Durante a docagem no QSAR, os métodos computacionais são usados para avaliar a interação entre a molécula de interesse e o sítio de interação da enzima ou proteína. Isso é importante para entender como diferentes estruturas moleculares interagem com enzima ou proteína específicas e como essas interações afetam a atividade biológica [29].

Essa abordagem é especialmente útil em design de medicamentos, onde os pesquisadores buscam identificar compostos candidatos que possam ter atividade biológica desejada, como a inibição de uma enzima específica. A docagem em QSAR permite prever a afinidade e a interação entre moléculas e proteínas, ajudando no desenvolvimento de novos fármacos [29].

2.9 Aprendizado de máquina (AM)

O aprendizado de máquina ou *Machine Learning*, é um dos campos da inteligência artificial que faz uso de modelos estatísticos para desenvolver previsões, onde os dispositivos computacionais são programados para aprender uma forma de modelagem preditiva ou análise descritiva.[99]

Através de um banco de dados de treinamento o algoritmo induz modelos preditivos

que podem predizer o valor de seu atributo alvo. O AM pode ser dividido em duas categorias: métodos o supervisionados (cujas variáveis dependentes são rotuladas em categorias) e os não supervisionados [99]. Neste trabalho serão analisados os modelos de predição que se encontram na categoria de métodos dos supervisionados. Além disto, será empregada uma análise de componentes principais, PCA, que é um método de análise exploratória não supervisionada.

A aplicação do AM no campo farmacêutico tem ganhado destaque nos últimos anos devido ao seu potencial para acelerar o processo de descoberta e desenvolvimento de novos medicamentos, bem como otimizar outras etapas do ciclo de vida de um fármaco [30, 31].

Através de um grupo de estruturas químicas são retirados cálculos físicos - químicos, os tipos de ligações, o número de conformações e demais cálculos de caráter químico, ambos são identificados como descritores, sendo variáveis independentes. O banco de dados de descritores é analisado juntamente com o valor do atributo alvo (variável dependente) e através de métodos de AM pode-se prever padrões numéricos a partir de um conjunto de treinamento, como por exemplo os valores de atividade inibitória (IC_{50}) e estrutura dos compostos ativos, resultando em uma análise dos respectivos resultados .

A partir da correlação e análise de dados, são verificados quais modelos preditivos que podem ser abordados para realizar uma boa predição. Dentre os métodos de aprendizado de máquina linear estão a análise de componentes principais - PCA, análise dos mínimos quadrados parciais - PLS, regressão por componentes principais (PCR) ou regressão linear múltipla (MLR) e etc... com um objetivo de desenvolver equação de regressão [32] [99].

Os Algoritmos empregados podem indicar novas moléculas com características desejadas, otimizando a estrutura química do fármaco para melhorar a eficácia e minimizar efeitos colaterais, através de uma análise de banco de dados, se identifica os compostos promissores que merecem investigação adicional, economizando tempo e recursos.

Quanto a predição não linear por classificação, pode-se citar a máquina de vetor suporte (SVM), a Rede neural e a Floresta aleatória [99].

3 Objetivo

Com as moléculas sintetizadas, analisadas biologicamente por [43] e os resultados de atividade de inibição da enzima N-miristoiltransferase do parasita *L. donovani* disponíveis, foi realizada uma análise do melhor método de aprendizado de máquina para construção de um modelo de previsão da atividade inibitória sobre essa enzima.

O trabalho envolve o desenvolvimento de modelos de relações quantitativas entre a estrutura e a atividade, de sigla QSAR do nome em inglês *Quantitative Structure Activity Relationships*, para gerar uma matriz de dados a ser analisada estatisticamente, juntamente com a aplicação da aprendizagem de máquina.

O objetivo geral deste trabalho é o estudo competitivo entre diferentes algoritmos de aprendizado de máquina, na modelagem preditiva da inibição do parasita *Leishmania donovani*, em comparação com a regressão linear múltipla, adotado como *benchmark*. Os objetivos específicos são:

1. Estudo competitivo entre diferentes algoritmos de aprendizado de máquina na predição de pIC_{50}, em uma base de dados de tamanho pequeno, contendo amostras de um grupo de classes química, através de diferentes métodos:

    a) Métodos lineares:

    - Regressão linear Múltipla, MLR
    - Regressão por mínimos quadrados parciais, PLS
    - Regressões regularizadas, como a Regressão Ridge, RR, e a Rede Elástica. ENet.

    b) Métodos não-lineares:

    - Regressão por máquina de vetor suporte, com os *kernels* função de base radial, RBF, e Linear
    - Regressão por Floresta Aleatória, RF.

A escolha do melhor modelo, caso dois ou mais modelos sejam aprovados em todos os requisitos de QSAR, o melhor modelo será aquele com o menor RMSE, nas etapas de calibração e validação cruzada *leave-one-out*, conforme indicado por [79], [87] e [90].

4 Metodologia

4.1 Base de dados

A base de dados utilizada neste trabalho foi extraída de [43]. Este banco de dados contém 77 estruturas moleculares que foram selecionadas a partir dos compostos sintéticos que exibem atividade inibitória em baixas concentrações de IC_{50} (concentração necessária para inibir 50% da atividade da enzima) da N-miristoiltransferase encontrada na *L. donovani* .

O IC_{50} é uma medida da potência de uma substância na inibição de uma função biológica ou bioquímica específica. O componente biológico neste trabalho é a enzima N-miristoiltransferase de *L. donovani*.

Os valores de IC_{50} foram convertidos para a escala:

$$pIC_{50} = -\log_{10}(IC_{50}) \tag{4.1}$$

Devido ao sinal de menos, valores mais altos de pIC_{50} indicam inibidores exponencialmente mais potentes. IC_{50} é geralmente medido como uma concentração molar e sua unidade é, portanto, mol/L ou M, seus submúltiplos, como mM, uM, nm, etc...[2].

A escolha das estruturas iniciou com a análise das respostas da IC_{50}, para o protozoário, *L. donovani*. Desta forma, foram excluídas estruturas com valores IC_{50} faltantes.

As atividades inibitórias sobre a N-miristoiltransferase contra a *L. donovani* em estágio ativo aeróbio serão inicialmente expressas em mol/L e depois convertidas para unidade logarítmica, usando para aumentar a linearidade e aproximar a distribuição normal da atividade valores[45].

4.2 Modelagem Molecular

Os Programas para modelagem molecular podem ser usados para o cálculo de descritores para análise de QSAR, neste trabalho utilizaram-se os programas Spartan'20. [46] e GOLD versão 2022.3.0 [47].

4.2.1 Programa Spartan

O programa 20 [48] foi utilizado para modelagem das moléculas, com o uso da otimização de cada geometria, por meio de método de mecânica quântica e criação de descritores. O programa Spartan é um aplicativo de modelagem molecular e química computacional da Wavefunction, contendo um código para mecânica molecular, métodos semi-empíricos, modelos ab

initio, modelos funcionais de densidade, modelos pós-Hartree-Fock, e cálculos termoquímicos [46]. As funções primárias são fornecer informações sobre estruturas, estabilidades relativas e outras propriedades de moléculas isoladas. Os cálculos de mecânica molecular em moléculas complexas são comuns na comunidade química. Cálculos químicos quânticos, incluindo cálculos de orbitais moleculares pelo método Hartree-Fock, mas especialmente cálculos que incluem correlação eletrônica, são mais demorados e mais precisos.

Cálculos de química quântica podem fornecer um conteúdo para complementar dados experimentais existentes ou substituí-los completamente, por exemplo, cargas atômicas para análises quantitativas de relação estrutura-atividade (QSAR) e potenciais intermoleculares para cálculos de mecânica molecular e dinâmica molecular [50]. Esse programa foi escolhido por conta da licença adquirida pela UFRRJ, o mesmo apresenta um módulo de cálculo com propriedades que podem ser usadas para buscar de modelos de QSAR e também por calcular propriedades eletrônicas, com a possibilidade de escolher quais parâmetros quer obter [5]. As energias foram minimizadas através do campo de força MMFF (*Merck Molecular Force Field*). Após isto, as energias foram novamente minimizadas com o método semi-empírico PM6 (*Parametric Model 6*).

Para cada molécula modelada, foram obtidos 24 descritores, sendo estes de variáveis independentes (apêndice 1).

No programa Spartan'20 é possível construir tanto as estruturas (em duas ou três dimensões) quanto obter informações das mesmas, a respeito das propriedades estruturais e eletrônicas (descritores químico-quânticos), que podem ser exploradas na tentativa de obter modelos de correlação, no que se chama QSAR.

O desenho de estruturas candidatas a moléculas bioativas pode ser realizado por vários métodos diferentes, um dos quais é a relação de atividade de estrutura quantitativa (QSAR), que relaciona os dados biológicos à estrutura química. Para a busca de modelos de QSAR é necessário um conjunto ou matriz de dados incluindo a geometria de modelos que requer a representação de um conjunto ou matriz de dados incluindo a medida quantitativa da atividade biológica e os parâmetros físico-químicos e estruturais que descrevem as propriedades dos compostos químicos. Ou seja, o conjunto de dados contém os valores da atividade biológica e variáveis descritivas referentes a compostos gerados [5].

A base de dados original contém 25 variáveis, sendo 24 independentes (os descritores) e uma dependente (o pIC_{50}), apresentando uma matriz de dimensão 77(amostras extraídas de [43])×25. Todavia, as variáveis X6 e X16 foram removidas por apresentarem dados constantes. Foram avaliadas medidas numéricas descritivas para cada variável descritora e a variável dependente).

Os descritores extraídos do programa do Software Spartan '20 foram nomeados de 1 a 24, na identificação de colunas nulas foram excluídas as colunas 6 e 16, restando 22 descritores para análise.

Tabela 4.1 – Tabela de Descritores - Software Spartan

Variável	Descrição e unidade	Definição
X1	Energia (kJ/mol)	Energia da estrutura obtida com o modelo selecionado.
X2	Energia em fasea aquosa (kJ/mol)	Energia aquosa (calor de formação)baseada no modelo SM5.4
X3	Energia de Solvatação (kJ/mol)	Diferença de energia entre a energia em fase aquosa e a energia no vácuo.
X4	Energia HOMO eV	A energia do último orbital molecular ocupado por elétrons.
X5	Momentos de dipolo são dados em Debye	Medida da polaridade de um sistema de cargas elétricas.
X6	Tautômeros	Isomerismo conformacional
X7	Peso amu	Massa molecular
X8	Energia LUMO eV	Menor energia de ocupação no orbital molecular
X9	Confôrmeros	Número de Confôrmeros
X10	Área Â2	Área de superfície molecular
X11	Volume Â3	Volume molecular
X12	Acc. Área Â2	área acessível da superfície
X13	Min ElPot (kJ/mol)	Potencial eletrostático mínimo
X14	Min LocIonPot kJ/mol	Mínimo do potencial de ionização local
X15	Log P	O valor do coeficiente de partição (log P) -medida quantitativa da lipofilicidade de qualquer compostos. P representa a partição de uma substância entre uma fase apolar (n-octanol) e a fase aquosa.
X16	HBD Count	Contagem de doadores de ligações de hidrogênio
X17	PSA Â2	Superfície polar
X18	Ovality	Ovalidade obtida a partir de um modelo de preenchimento de espaço
X19	P-Area(75) Â2	Área polar acessível
X20	Acc. P-Area(75) Â2	área polar (Acc. P-Area) de um mapa de potencial eletrostático*
X21	Max ElPot (kJ/mol)	Máximo potencial eletrostático
X22	Polarizability	Polarizabilidade
X23	HBA Count	Contagem de receptores de ligações de hidrogênio
X24	Energia (kJ/mol)	Energia interação intramolecular

4.2.2 Programa GOLD

O programa GOLD versão 2022.3.0 [47] utiliza o algoritmo genético para propor diversos modos de interação, para os quais são calculadas pontuações com as funções escolhidas. O processo de docagem neste programa é o acoplamento de estruturas moleculares nos sítios de ligação de macromoléculas, identificando as formas de ligação mais suscetíveis quanto a afinidade de ligação. Onde a estrutura se liga a proteína rígida e é analisada a interação [51]. As informações geradas pelos algoritmos de busca são então avaliadas pelas funções de pontuação, utilizadas para ordenar os ligantes para determinar a ordem de aptidão, dando uma estimativa da afinidade dos ligantes pelo alvo molecular.

Este programa realiza o docagem de proteína-ligante validado e configurável que pode ser usado na busca de candidatos a fármacos, da triagem virtual até a otimização de *leads*.

O GOLD apresenta restrições para orientar os resultados em relação a recursos ou comportamentos conhecidos, ele avalia o impacto das moléculas de água na docagem e obtém resultados rápidos de docagem do ligante na proteína.

A captura de tela na figura 4.1 apresenta a captura de tela de uma estrutura molecular a ser analisada.

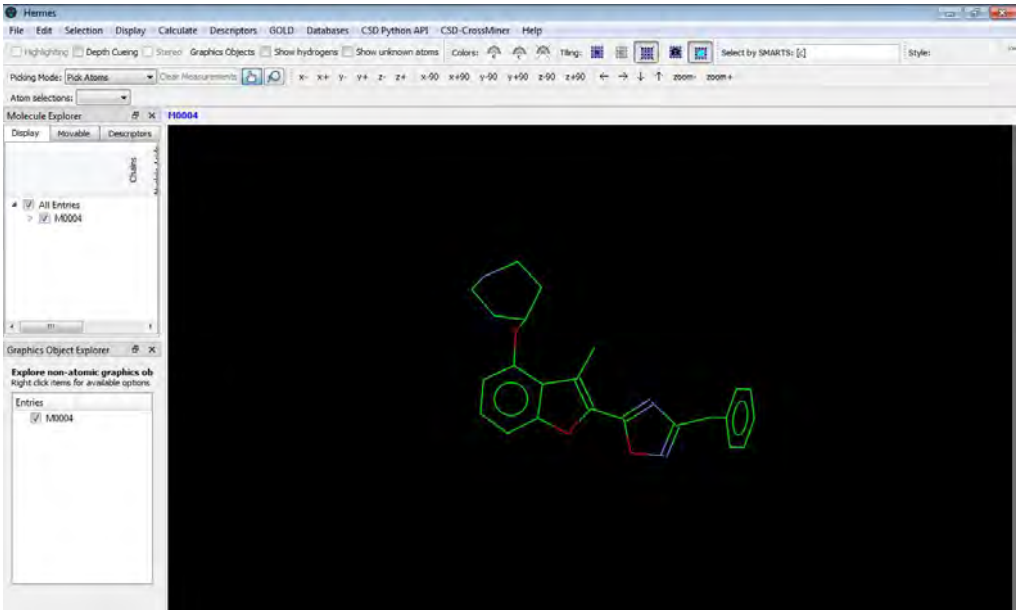


Figura 4.1 – Captura de tela da área de trabalho do *software* GOLD.
Fonte: O autor, 2023.

4.3 Identificação de descritores no programa GOLD

As modelagem das amostras criadas no programa Spartan foram salvas no formato mol2 para utilizar no programa GOLD. O programa GOLD apresentou restrições para orientar os resultados em relação a recursos ou comportamentos conhecidos, pois ele avalia o impacto das moléculas de água no acoplamento e obtém resultados rápidos de ancoragem do ligante na proteína.

Com o banco de dados pronto, foram aplicadas as 4 funções do programa GOLD, utilizadas nas 77 amostras, para o código (2wuu), que originou 25 descritores:

Tabela 4.2 – Codificações, funções e variáveis utilizadas

Codificação	Função	Variáveis independentes
A	CHEMSCORE	oito
B	ASP	seis
C	GOLDSTORE	cinco
D	CHEMPLP	seis

Elaborada pelo autor, 2023.

1 - A função Chemscore estima a variação da energia livre total do sistema. Nessa etapa acontece a ligação do ligante com a proteína através da soma ponderada da energia livre, que incluem interações lipofílicas, interações com metais e ligações de hidrogênio [53].

Tabela 4.3 – Tabela da função A - Chemscore

A1	CS_Score	Contribuição para ligações CH...O H fracas
A2	DG	Contribuição da variação de energia livre (que ocorre na ligação do ligante) para o valor do ChemScore
A3	$CS_S(hbond)$	Contribuição da ligação H do ligante proteína para o valor do ChemScore
A4	S(lipo)	Contribuição lipofílica de proteína-ligante para o valor ChemScore
A5	H(rot)	Contribuição do termo de congelamento de ligação rotativa para o valor ChemScore
A6	$CS_DE(clash)$	Penalidade de confronto proteína-ligante para o valor ChemScore
A7	$CS_DE(int)$	Penalidade de deformação torcional do ligante interno para o valor ChemScore
A8	intcor	Energia interna do ligante

2 – A função ASP (Astex Statistical Potential) é identificada como a geração de potenciais estatísticos, a partir da frequência de interação entre pares de átomos da proteína e do ligante que foram observadas em uma coleção de complexos ligante-proteína [54].

Tabela 4.4 – Tabela de função B - ASP

B1	Score	função de Pontuação ASP
B2	ASP	Potencial estatístico calculado mais o termo de colisão ChemScore e o termo de energia interna
B3	S(Map)	O potencial estatístico total calculado é uma soma de todas as combinações de átomos de proteína e ligante
B4	DE(clash)	Penalidade de colisão proteína-ligante para o valor ASP
B5	DE(int)	Contribuição da ligação H intramolecular do ligante para o valor de ASP
B6	intcor	Variação de energia interna do ligante

3 - A função Goldscore através de parâmetros empíricos, calcula a soma de termos de energia para as interações de Van der Waals e de ligações de hidrogênio entre a proteína e o ligante, juntamente com termos de energia torcional para os ligantes [34].

Tabela 4.5 – Tabela da função C - Goldscore

C1	Fitness	Valor total de aptidão GoldScore do ligante acoplado
C2	S(hb$_ext$)	Contribuição da ligação H do ligante proteína para o valor GoldScore
C3	S(vdw$_ext$)	Contribuição da energia de Van der Waals entre a proteína e o ligante para o valor GoldScore
C4	S(int)	Contribuição da ligação H intramolecular do ligante para o valor GoldScore
C5	intcor	Variação de energia interna do ligante

4 – A função ChemPLP modela a complementariedade estérica entre ligante e proteína,através dos ângulos e das distâncias de interações de hidrogênio e interações envolvendo metais. A sigla PLP (Piecewise Linear Potential) representa um potencial específico que modela a atração e repulsão de átomos diferentes de hidrogênio entre proteína e ligante e potenciais internos para os ligantes [55].

Tabela 4.6 – Tabela de função D - ChemPLP

D1	Score	Contribuição para ligações CH...O H fracas
D2	S(PLP)	Potenciais calculados mais o termo de colisão ChemScore e o termo de energia interna
D3	S(hbond)	Contribuição da ligação H do ligante de proteína ChemScore
D4	DE(clash)	Penalidade de colisão proteína-ligante para o valor PLP
D5	DE(tors)	Penalidade de deformação torcional do ligante interno para o valor PLP
D6	intcor	Variação de energia interna do ligante

Uma matriz com descritores x amostras foi confeccionada. Essa matriz tem 22 descritores do programa Spartan, mais 25 descritores do programa GOLD e 77 amostras.

4.4 Análise de Dados

4.4.1 *Hardware*

A análise foi realizada em um notebook Lenovo Legion, com processador 12th Intel Core i7- 2700H a 2,30 GHz; 32 GB de memória RAM; placa de vídeo Nvidia GeForce RTX com 6 GB de memória; HD Samsung SSD 500 GB.

4.4.2 Programa

Para aplicação de aprendizado de máquina, utilizou-se o programa R para manipulação, análise e predição de dados .

Foi utilizada a versão 4.3.0 [56] e seu ambiente de desenvolvimente integrado, IDE (*integrated development environment*), o RStudio versão 2023.03.1+446 [57].

Este ambiente de programação, é também uma linguagem de programação amplamente utilizada em análises estatísticas, visualização de dados e aprendizado de máquina. O R é gratuito e tem código aberto, o que significa que qualquer pessoa pode baixá-lo, usar e modificar o código-fonte [58].

O ambiente R é um conjunto integrado de recursos para manipulação de dados, cálculo e gráficos [58]. Entre suas características, destacam-se a eficaz de manipulação e armazenamento de dados; ter um conjunto de operadores para cálculos em matrizes, em particular matrizes; apresentar uma coleção grande, coerente e integrada de ferramentas intermediárias para análise de dados; oferecer recursos gráficos para análise e exibição de dados diretamente no computador ou em cópia impressa, e tem condicionais, loops, funções recursivas definidas pelo usuário e recursos de entrada e saída.

O software R trabalha com o sistema de pacotes. Desta forma, ele não é um programa monolítico, mas sim um programa flexível para que os usuários escrevam rotinas e publiquem em repositórios, como GitHub e a CRAN (esta após verificação pelo R Core Team). Desta forma, diferentes pacotes foram utilizados neste trabalho. Eles são listas na tabela 5.1, assim como as funções relacionadas.

4.5 Análise Exploratória de Dados, EDA

A análise de dados em QSAR envolve forte uso de ferramentas estatísticas, tanto descritivas quanto inferenciais [42, 76]. A análise começa pelo conceito de Análise Exploratória de Dados (EDA), onde são observadas estatísticas univariadas, tais como os cinco números de Tukey (mínimo, primeiro quartil, mediana, terceiro quartil e máximo); a média aritmética, o desvio-padrão amostral, o coeficiente de variação [88], a amplitude, a assimetria e a curtose.[86].

Existem diferentes equações na literatura sobre assimetria e curtose. A importância da assimetria se dá na etapa de modelagem preditiva pelo método dos mínimos quadrados ordinários, pois espera-se que cada variável preditora tenha distribuição normal. Desta forma, utilizou-se a assimetria amostral G_1 (equação 4.2) e o excesso de curtose amostral G_2 (equação 4.3) [86]

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{4.2}$$

$$G_2 = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right) - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{4.3}$$

A interpretação da assimetria G_1 se dá de acordo conforme Bulmer (1979) *apud* [86]:

Tabela 4.7 – Interpretação do coeficiente G_1.

Condição	**Resultado**
$G_1 > \lvert \pm 1 \rvert$	Altamente assimétrica
$-1 \leq G_1 \leq -1/2$ ou $1/2 \leq G_1 \leq 1$	Moderada assimetria
$-1/2 \leq G_1 \leq 1/2$	Aproximadamente simétrica
$G_1 = 0$	Perfeita simetria

Já para a interpretação do excesso de curtose amostral, se $G_2 < 0$ os dados são platicúrticos; se $G_2 = 0$ os dados são mesocúrticos e, for fim, se $G_2 > 0$, os dados são leptocúrticos.

Além destas estatísticas univariadas, [87, pág. 31-33] sugerem que, se a razão máximo/mínimo for maior que 30, que existe uma forte assimetria presente e sugerem a transformação Box-Cox, que apesar do nome, é uma família de transformações que são indexadas pelo parâmetro λ:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log() & \text{se } \lambda = 0 \end{cases} \tag{4.4}$$

A presença de valores discrepantes, ou *outliers*, pode também impactar na etapa de modelagem preditiva. De forma univariada, pode-se avaliar a presença de *outliers* através de uma métrica simples, o escore-z [88]:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{4.5}$$

onde z_{ij} é o i-ésimo valor, da j-ésima coluna, transformado; x_{ij} é o i-ésimo valor original, da j-ésima coluna; s_j é o desvio-padrão da j-ésima coluna; e $\bar{x}$ é a média aritmética da

j-ésima coluna. Ao realizar esta transformação, os dados passam a ter média zero e desVio-padrão unitário. Desta forma, valores acima de $|\pm 3,0|$ são considerados *outliers*.

Outra forma de identificar *outliers* univariados ocorre através da inspeção visual de amostras em um gráfico de caixas. Com dados originais, amostras que estão acima (ou abaixo) dos limites mostrados nas equações 4.6 e 4.7 são tidas como *outliers* suspeitos:

$$LI = Q_1 - 1,5 \times (Q_3 - Q_1) \tag{4.6}$$

$$LS = Q_3 + 1,5 \times (Q_3 - Q_1) \tag{4.7}$$

Além disso, um *outlier* univariado não necessariamente é um *outlier* multivariado. Para avaliar isso, foi usada a distância de Mahalanobis [89, pág. 105].

A elipse de confiança baseada na distância de Mahalanobis é uma ferramenta estatística utilizada para representar a variabilidade multivariada de um conjunto de dados e é especialmente útil quando se lida com dados correlacionados.

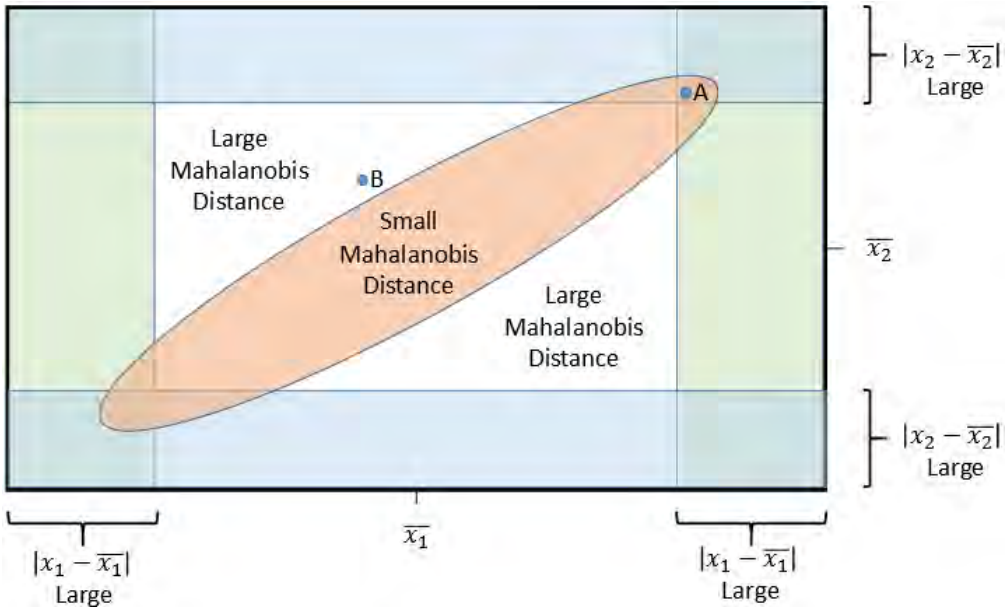


Figura 4.2 – Elipse da distância de Mahalanobis

Fonte: Extraído de [91].

A distância de Mahalanobis é uma medida que leva em consideração as correlações entre as variáveis e é definida como a distância entre um ponto de dados e o centro dos dados, levando em conta a matriz de covariância dos dados. Neste trabalho realizou-se a construção da distância de Mahalanobis clássica [90]).

A equação da distância de Mahalanobis para um ponto x em relação ao vetor médio μ e a matriz de covariância $\sum$ é dada por [89]:

$$d(x,i) = (x - \mu_i)^T \sum_i^{-1} (x - \mu_i) \tag{4.8}$$

Se os dados multivariados seguem uma distribuição normal multivariada, com dada matriz de média e covariância, então uma amostra será considera um *outlier* multivariado se o valor da distância de Mahalanobis ao quadrado for maior que o valor crítico, calculado a partir de uma distribuição qui-quadrado [90]

$$\text{Valor crítico} = \sqrt{\chi^2_{\left(1-\frac{\alpha}{2};\, m\right)}} \tag{4.9}$$

onde α é o nível de significância adotado e m, o número de graus de liberdade (aqui, m representa o número de variáveis preditoras).

Por último, mas não menos importante, é a análise da relação entre as variáveis preditoras e também entre as variáveis preditoras e a variável dependente. Para isso, caso a relação entre as variáveis seja linear, pode-se utilizar o coeficiente de correlação linear de Pearson:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}} \quad \text{com } -1 \leq r \leq 1 \tag{4.10}$$

4.6 Modelagem preditiva

Uma vez que o banco de dados seja considerado adequado, a próxima etapa na criação e validação de modelos em QSAR consiste no emprego da regressão linear múltipla por mínimos quadrados ordinários [42]. Como este método é o *benchmark* na área de QSAR, será dado uma maior ênfase à sua descrição, bem como aos pressupostos que devem ser atendidos ao utilizá-lo.

4.6.1 Regressão linear múltipla, MLR

A regressão linear múltipla (MLR, *multiple linear regression*) por mínimos quadrados ordinários (OLS) consiste na relação entre uma variável dependente, y, e duas ou mais variáveis independentes (ou, em QSAR, preditoras). O modelo de regressão linear múltipla, para k variáveis independentes, é dado por [96]:

$$\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{4.11}$$

e a resposta estimada é obtida a partir da equação de regressão amostral:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \varepsilon_i \tag{4.12}$$

onde $\mu_{Y|x}$ é a resposta média de y dado x; β_0 é o intercepto de Y; β_1 é a inclinação de Y em relação à variável x_1, mantendo-se constantes as variáveis x_2, x_3; $\ldots$; x_k e assim por diante; e ε_i é o erro aleatório em Y, para cada i-ésima observação.

A regressão linear múltipla pode ser representada em uma forma matricial [75, 96, 97, 98]:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.13}$$

onde $\mathbf{Y}$ é o vetor com os dados da variável dependente, β é o vetor de coeficientes e $\boldsymbol{\varepsilon}$ é o vetor com erros aleatórios de cada i-ésima amostra, ou seja:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (4.14)$$

Desta forma, o método de mínimos quadrados para estimação de β envolve encontrar $\mathbf{b}$ para o qual a soma quadrática dos erros (ou resíduos) é minimizada:

$$\text{SQE} = \left(\mathbf{y} - \mathbf{Xb}\right)^{'} \left(\mathbf{y} - \mathbf{Xb}\right) \quad (4.15)$$

Este processo de minimização envolve resolver $\mathbf{b}$ na equação:

$$\frac{\partial}{\partial \mathbf{b}} (\text{SQE}) = 0 \quad (4.16)$$

O resultado reduz a solução de $\mathbf{b}$ em:

$$\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)\mathbf{b} = \mathbf{X}^{'}\mathbf{y} \therefore \mathbf{b} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\mathbf{X}^{'}\mathbf{y} \quad (4.17)$$

Uma vez obtida a equação de regressão, pode-se obter o erro-padrão da regressão, s^2, que é um estimador não enviesado de σ^2, que é uma medida de variação nos erros de predição.

$$s^2 = \frac{\text{SQE}}{n-k-1} \quad (4.18)$$

onde k é o número de variáveis preditoras e a soma quadrática dos resíduos é:

$$\text{SQE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \quad (4.19)$$

onde e_i é i-ésimo resíduo. Uma análise de variância (ANOVA) pode empregada, para avaliar a qualidade da equação de regressão. Neste caso, a decomposição da variância leva a:

$$\begin{gathered} SQT = SQR + SQE \\ \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \end{gathered} \quad (4.20)$$

Sob a hipótese $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, podemos verificar se a quantidade de variação explicada pelo modelo é significativa ou não. Para isto, construímos uma tabela ANOVA:

Tabela 4.8 – Tabela ANOVA da regressão.

Fonte	**Soma dos quadrados**	**Graus de liberdade**	**Média quadrática**	**F**
Regressão	SQR	k	MQR	MQR/MQE
Erro	SQE	$n-k-1$	MQE	
Total	SQT	$n-1$	MQT	

Se o valor da estatística F exceder o valor crítico $F\,(1-\alpha,\ k,\ n-k-1)$, então a equação da regressão difere de uma constante e ao menos uma variável regressora é significativa.

A próxima etapa consiste em avaliar a significância estatística de cada coeficiente de regressão. Para isto, um teste t é empregado. Sob as hipóteses nula $H_0 : \beta_j = \beta_{j0}$ e alternativa $H_1 : \beta_j \neq \beta_{j0}$, temos [112]:

$$t = \frac{b_j - \beta_{j0}}{\sqrt{s_{b_j}}} \quad (4.21)$$

onde s_{b_j} é o erro-padrão do coeficiente b_j, que é obtido como:

$$s_{b_j} = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{4.22}$$

onde $\hat{\sigma}^2$ é o estimador da variância e C_{jj} são os elementos da diagonal da matriz de covariância $\mathbf{C}$ [97]

$$C_{jj} = \text{diag}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-\mathbf{1}} \tag{4.23}$$

O coeficiente será significativo estatisticamente se $|t| > t\left(1-\frac{\alpha}{2}; n-p\right)$, onde p é o número de parâmetros incluindo o intercepto.

Para avaliar a qualidade do modelo, podem ser empregadas duas métricas: o coeficiente de determinação múltiplo, r^2 (equação 4.24), e o coeficiente de determinação ajustado, r^2_{aj} (equação 4.25) [112]. O primeiro explica a porção da variação total na variável dependente que é explicada pela variação nas variáveis independentes. Já o segundo penaliza o modelo caso mais variáveis sejam atribuídas, com o intuito de melhorar o ajuste.

$$r^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = 1 - \frac{SQE}{SQT} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{4.24}$$

$$r^2_{aj} = 1 - \frac{SQE/\left(n-k-1\right)}{SQT/\left(n-1\right)} \tag{4.25}$$

Vale ressaltar que tanto um r^2 quanto o r^2_{aj} elevados não implicam necessariamente com uma relação linear entre as variáveis independentes e a variável dependente. Todo modelo deve ser examinado visualmente.

Todas as métricas indicadas acima referem-se a um modelo inicial. Em QSAR, e também na área de modelagem preditiva, devido ao tamanho do banco de dados, costuma-se particionar a base de dados em duas: treino (ou calibração) e teste (ou validação) [87, 89, 102]. Essa partição, em geral, ocorre nas proporções 80%/20%, 70%/30% ou 60%/40% [99, pág. 151].

Uma vez que a base de dados tenha sido dividida em treino e teste, algumas métricas são necessárias para avaliar a qualidade do modelo e a qualidade preditiva do modelo em um conjunto de dados independente. Além disso, para observar se o modelo está enviesado, utiliza-se um método de reamostragem denominado *leave-one-out*, onde uma amostra é removida do conjunto de treino e tem seu valor previsto. Isso é feito até que esse processo tenha sido aplicado a todas as amostras [99] [100, pág. 14]. Essa estratégia é muito importante para se ter uma ideia a respeito da capacidade preditiva e da robustez do modelo e é chamada *leave-one-out* [32].

Desta forma, diferentes métricas de qualidade foram utilizadas neste trabalho, para avaliar cada etapa (calibração, validação cruzada e validação).

Na etapa de calibração, foram avaliados o coeficiente de determinação múltiplo (equação 4.24), o coeficiente de determinação ajustado (equação 4.25), a raiz do erro médio quadrático (equação 4.26), o erro médio absoluto (equação 4.27) e o viés (equação 4.28) [101, 102].

$$RMSEC = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}\left(y_i - \hat{y}_i\right)^2} \tag{4.26}$$

$$MAE = \frac{1}{n_c}\sum_{i=1}^{n_c}\left|y_i - \hat{y}_i\right| \tag{4.27}$$

$$BIAS = \frac{1}{n_c}\sum_{i=1}^{n_c}\left(y_i - \hat{y}_i\right) \tag{4.28}$$

A vantagem destas métricas é que permitem a comparação entre diferentes modelos, com diferentes números de variáveis. A interpretação das equações 4.26, 4.27 e 4.28 é bem simples, na comparação entre diferentes modelos: quanto menor a métrica obtida, melhor.

Para a etapa de validação cruzada, foram utilizadas as seguintes métricas de qualidade: soma dos quadrados do erro de predição (equação 4.29), [101, pág. 185-186] [96] [102, pág. 71] [76]. Em específico a área de QSAR, foram usadas o coeficiente de correlação de validação cruzada (equação 4.30), o coeficiente de determinação permutado médio corrigido (equação 4.31) [76], e coeficiente de determinação de predição (equação 4.32) [96].

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 \tag{4.29}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \tag{4.30}$$

$${}^{c}r_p^2 = r \times \sqrt{r^2 - r_p^2} \tag{4.31}$$

$$r^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \tag{4.32}$$

onde y_i é o i-ésimo valor experimental; $\hat{y}_{(i)}$ é o valor estimado de y_i sem que a amostra tenha sido incluída no modelo; r_p^2 é o coeficiente de determinação médio permutado, após a aleatorização em y. Vale notar que $Q^2 \equiv r^2_{pred}$, onde a notação difere somente em relação à área aplicada (QSAR e Estatística, respectivamente). Um outro teste é a aleatorização em y [35], onde são feitas permutações nas variáveis dependentes mantendo as variáveis independentes fixadas. Com isso, novos modelos de regressão são criados. Se os valores destes novos r^2 e Q^2 forem muito mais baixos que os originais, então há um motivo para confiar no modelo. Os autores [103, pág. 50] indicam que, na aleatorização em y devemos usar um coeficiente de determinação médio das permutações. Neste trabalho foram feitas 1.000 permutações em y.

Na etapa de validação, são utilizadas mesmas métricas utilizadas na etapa calibração (RMSE, MAE, BIAS e r^2); porém os valores preditos $\hat{y}_i$ são obtidos não através da construção de um modelo mas sim a partir da aplicação do modelo construído sobre os valores de teste. Esta etapa é importante para avaliar quão bem o modelo se comporta na modelagem de amostras que não foram incluídas na etapa de calibração.

Conforme [39] e [35] indicam, na validação cruzada LOO, as seguintes relações devem ser preservadas: $r^2 > Q^2$ e RMSEC < RMSEP. Uma condição aceitável é $Q^2 > 0,5$ e $r^2 > 0,6$. Eles alertam que se a diferença entre o coeficiente de determinação e o de validação cruzada for maior que intervalo de 0,2 a 0,3, ou seja, $r^2 - Q^2 > (0,2 \text{ a } 0,3)$, o modelo tem sobreajuste.

Além disso, na construção de modelos de previsão por mínimos quadrados, é importante avaliar a multicolinearidade [104], pois esta pode aumentar a variância dos coeficientes obtidos, tornando-os instáveis. Além da correlação linear entre as variáveis, onde [105] indicam que a correlação entre as variáveis preditoras maior que 0,7 pode apresentar problemas, pode-se avaliar também o fator de inflação da variância (equação 4.33). Em geral, se o valor de VIF_j for maior que 5 ou 10, existe a variável apresenta colinearidade [97, pág. 168]. Vale observar que [104] indica a regra de Lein na avaliação da multicolinearidade. Esta regra diz que a multicolinearidade torna-se um problema quando o coeficiente de determinação do modelo obtido de uma regressão auxiliar for maior que o r^2 obtido via y-original versus X-originais.

$$VIF_k = \frac{1}{1 - r_k^2} \tag{4.33}$$

Na construção de modelos, de acordo com [106, pág. 281], existem métricas para seleção de variáveis, na construção do modelo. Entre eles, destacam-se o critério de informação de Akaike, o critério de informação de Schwarz e o Cp de Mallows [88, 107].

$$AIC = \left[\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n} \right] \times e^{2k/n} \tag{4.34}$$

$$BIC = \left[\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n} \right] \times n^{k/n} \tag{4.35}$$

$$Cp = \frac{\left(1 - r_k^2\right)(n - T)}{1 - r_T^2} - (n - 2\,(k+1)) \tag{4.36}$$

onde T é o número total de parâmetros (incluindo o intercepto), k é o número de variáveis preditoras, r_k^2 é o coeficiente de determinação com k variáveis preditoras e r_T^2 é coeficiente de determinação com todas as variáveis preditoras. No estudo comparativo entre modelos candidatos, o modelo "ideal"será aquele com os menores valores de AIC ou BIC; enquanto que, ao se adotar o critério de Mallows, será aquele com Cp menor ou igual ao número de variáveis preditoras mais intercepto.

Além disso, caso o modelo esteja adequado, avalia-se também as premissas de normalidade dos resíduos, de homocedasticidade, de ausência de autocorrelação serial e de linearidade [112, 88]. Não obstante, a influência e resíduos estudentizados externamente são também avaliados, na identificação de possíveis *outliers*.

O teste de normalidade adotado foi o de Shapiro-Wilk. Sob a hipótese nula $H_0 : e_i \sim N\left(0, \sigma^2\right)$ e um nível de significância α, o teste será rejeitado se o valor do teste for maior que o valor crítico tabelado. Vale ressaltar que a função *shapiro.test()* utilizada na verdade é aplicada de acordo com a modificação de Royston, que expandiu o método original, limitado em até 50 amostras, para até 5000 amostras [108].

Para testar a homocedasticidade, foi empregado o teste de Breusch-Pagan [109, pág. 196]. Sob a hipótese nula $H_0 : \sigma_{e_i}^2 = 0$, dado um nível de significância α, esta será rejeitada se a estatística do teste for maior que o valor crítico tabelado. Neste teste, após a obtenção do modelo de regressão linear, é construído um novo modelo; desta vez, dos resíduos ao quadrado em função dos valores previstos de y (equação 4.37):. Desta forma, a estatística de teste é obtida através de um teste qui-quadrado (equação 4.38). A hipótese nula será rejeitada se a estatística do teste for maior que o valor crítico (equação 4.39):

$$\varepsilon_i^2 = \delta_0 + \delta_1 \hat{y}_i \tag{4.37}$$

$$\chi^2 = n \times R_{\text{aux}}^2 \tag{4.38}$$

$$\chi_{\text{crit}}^2 \left(1 - 0{,}05; 1\right) \cong 3{,}841 \tag{4.39}$$

Para testar a ausência de autocorrelação serial, foi empregado o teste de Durbin-Watson [88] (equação 4.40). A estatística do teste está delimitada entre zero e quatro, onde existem duas regiões de rejeição da hipótese nula e uma região de não-rejeição. Infelizmente, este teste apresenta duas regiões inconclusivas, calculadas a partir do número de amostras presentes na

regressão e o número de variáveis preditoras. A hipótese nula será rejeitada se o valor calculado cair na região de não-rejeição da hipótese nula.

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} \left(e_i^2\right)} \tag{4.40}$$

4.6.2 Regressão por Mínimos Quadrados Parciais, PLS

A regressão por mínimos quadrados parciais (PLSR, *partial least squares regression*), apesar de suas origens na Economia [87, pág. 113], viu sua popularidade na Química, em particular na Química Analítica [23, pág. 289]. Este método apresenta como vantagens poder trabalhar em bancos de dados **X**, cujo número de variáveis p é maior que o tamanho amostral n (com $\mathbf{X} \in R^{n \times p}$), bem como a presença de variáveis preditoras altamente correlacionadas. Por ser um método iterativo, diversos algoritmos estão disponíveis na literatura, Em termos gerais, o PLS visa construir variáveis latentes (novas variáveis formadas a partir de combinações lineares das variáveis originais) de forma a capturar a maior variância em **X** e **Y** e a maximizar a correlação entre essas matrizes, ou seja, a maximização da covariância entre **X** e **Y** [89, pág. 156]. Desta forma, a matriz **X** é decomposta em três matrizes: a matriz **T**, de escores (*scores*); a matriz de pesos **P** (*loadings*) e a matriz de resíduos **E**. Já a matriz **Y** também é decomposta em três matrizes: uma de escores (**T**), uma de pesos **Q** e uma de resíduos, **F** [23, 64, 89, 90, 102, 140]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \sum_{i=1}^{A} \mathbf{t_i}\mathbf{p_i}' + \mathbf{E} \tag{4.41}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}` + \mathbf{F} = \sum_{i=1}^{A} \mathbf{t_i}\mathbf{q_i}' + \mathbf{f} \tag{4.42}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}\left(\mathbf{P}'\mathbf{W}\right)^{-1} \tag{4.43}$$

Essa decomposição da matriz **X** e de, no caso, um vetor **y**, é ilustrada na figura 4.3, extraída de [23, pág. 290]:

A matriz X, com I amostras e J variáveis ou preditores, é decomposta nas matrizes **T**, **P** e **E**, onde I é o número de amostras; J é o número de variáveis originais e A é o número de variáveis latentes. Já o vetor de respostas **y** (no exemplo original chamado de **c**, em virtude de ser um vetor de concentrações), é também decomposto em uma matriz de escores **T**, um vetor **q** de pesos e um vetor de resíduos **f**. Por fim, **W** é uma matriz de pesos.

Do ponto de vista geométrico, os escores e os pesos representam as coordenadas no plano cartesiano e a contribuição relativa de cada variável preditora original, respectivamente [141, 23]

A escolha do número de variáveis latentes é um processo crítico, na construção do modelo de regressão. Em geral, é utilizado como apoio no processo de escolha o uso de validação cruzada [23, pág. 215]. Pode ser tentador encontrar um modelo com alto valor de r^2, entretanto, ele provavelmente apresentará *overfitting*, o que prejudicará na etapa de validação, por inflar os coeficientes e aumentar os erros. Desta forma, ao realizar a validação cruzada e adotar uma métrica para minimizar a ocorrência desse fenômeno, em geral a raiz quadrada do erro médio (RMSE), é obtido um modelo mais realista.

Um modelo PLS pode conter centenas ou milhares de variáveis [137]. Algumas estratégias de seleção de variáveis podem ser utilizadas, tanto para aumentar a performance da

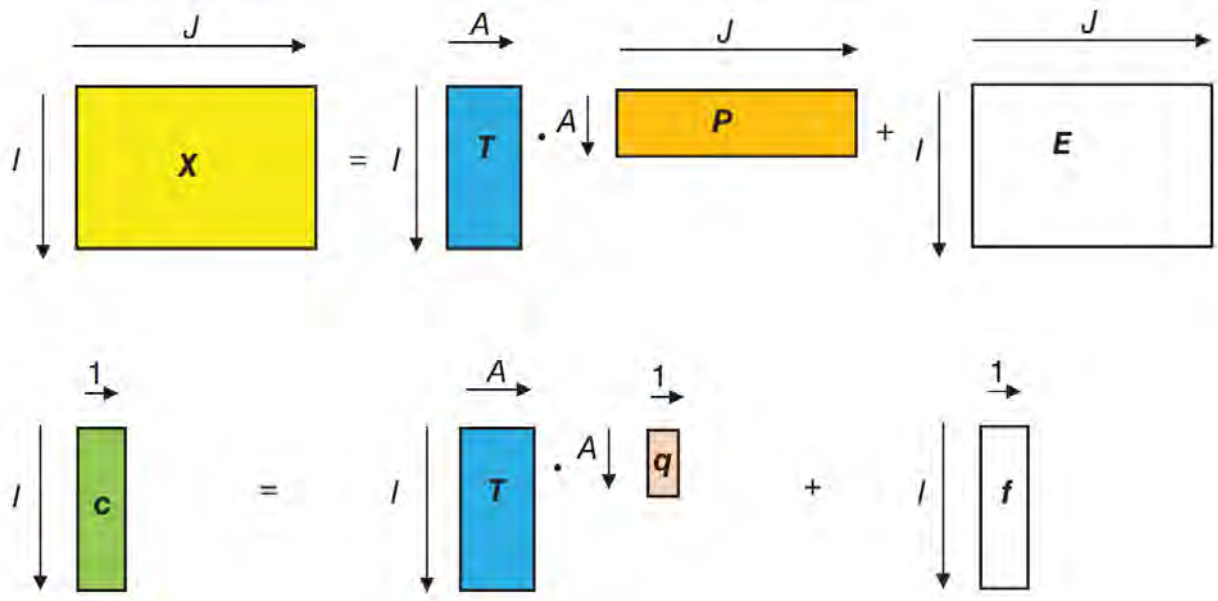


Figura 4.3 – Modelo PLS

Fonte: extraído de [23].

capacidade preditiva do modelo quanto para diminuir o custo computacional. Neste trabalho, serão utilizados dois métodos de seleção de variáveis para o PLS.

4.6.2.1 Projeção da importância da variável, VIP-PLS

Este nétodo de seleção de variáveis foi proposto por [138]. A ideia básica é reter variáveis latentes que estejam acima de um limiar estabelecido. A importância de cada variável é obtida de acordo com a equação 4.44 [140]:

$$v_j = \sqrt{\frac{p\sum_{a=1}^{A}\left[SS_a\left(w_{aj}/||w_a||\right)^2\right]}{\sum_{a=1}^{A} SS(w_a)}} \tag{4.44}$$

onde SS_a é a soma dos quadrados explicada pelo a-ésimo componente e $\left(w_{aj}/||w_a||\right)^2$ representa a importância da j-ésima variável. Uma medida de limiar para remoção da j-ésima variável é $v_j < u$, onde $u \in [0,\infty)$. É comum aceitar como limiar o valor $u = 1$. Desta forma, espera-se que as variáveis candidatas finais permitam um incremento na capacidade preditiva do modelo.

4.6.2.2 Eliminação de variável não-informativa, UVE-PLS

Neste método, proposto por [139], a importância de cada váriável, para o modelo final, é comparada através de um índice de confiabilidade, uma função dos coeficientes de regressão, com os das variáveis aleatórias artificiais. A partir da matriz de dados X original, é criada uma matriz aleatória artificial com a mesma dimensão. A seguir, um modelo PLS é construído, e a matriz de coeficientes B é retida. O índice de confiabilidade é então obtido [140]:

$$
\begin{aligned}
c_j &= \frac{m(b_j)}{s(b_j)} \\
m(b_j) &= \frac{\sum_{i=1}^{n} b_{ij}}{n} \\
s(b_j) &= \sqrt{\frac{\left(b_{ij} - m(b_j)\right)^2}{n-1}}
\end{aligned}
\tag{4.45}
$$

onde c_j é o índice de confiabilidade da j-ésima variável; $m(b_j)$ e $s(b_{ij})$ são, respectivamente, a média e o desvio-padrão do coeficiente de regressão da j-ésima variável obtida via validação cruzada *leave-one-out*, b_{ij}. Se as variáveis latentes originais com índice de confiabilidade absoluta forem menores que o valor de corte, elas são então eliminadas.

4.6.3 Métodos de regularização

Embora a regressão PLS seja popular em Química [23], ela não é a única alternativa quando a multicolinearidade está presente. Alguns métodos preditivos populares na área de Estatística podem ser utilizados: a regressão ridge, a regressão LASSO e a regressão por Rede Elástica (...)".

4.6.3.1 Regressão Ridge

Quando existe *overfitting* ou então a multicolinearidade está presente, uma forma de contornar esses obstáculos é através da regressão ridge[1] [142], que adiciona um termo de penalidade sobre a soma dos quadrados dos coeficientes (equação 4.46):

$$
SQE_{L_2} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{4.46}
$$

onde SQE_{L_2} significa uma penalidade de segunda ordem. De acordo com [87, pág. 123], as estimativas dos parâmetros só podem tornar-se grandes se a redução na soma quadrática do erro for proporcional. Desta forma, esta abordagem reduz as estimativas para zero a medida que o termo de penalidade lambda se torna grande.

As estimativas dos coeficientes são obtidas da seguinte forma [145, pág. 450 - material suplementar]:

$$
\beta^{\star}(\lambda) = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{4.47}
$$

Os autores [145, pág. 450 - material suplementar] indicam, ainda, que as variáveis preditoras devem ser transformadas, antes da construção do modelo.
Nem sempre é intuitivo escolher o valor adequado do parâmetro λA, pois ele está situado em na região $0,0 \leq \lambda \leq 1,0$, onde $1,0$ é a solução completa.. A escolha do melhor parâmetro λ pode então ser feita através de validação cruzada.

[1] Nota: Embora a tradução seja apontada como *regressão corrigida* por [112, pág. 263] e por *regressão de cumeeira* por [146] e [147], optou-se por manter a expressão mais comum da literatura.

4.6.3.2 Regressão LASSO

Um outro método de regularização é a regressão LASSO (*Least Absolute Shrinkage and Selection Operator* [143]. Uma grande vantagem da regressão LASSO é a sua capacidade de selecionar variáveis, eliminando aquelas que são irrelevantes ou redundantes [143]. A regularização L_1 controla o *overfitting* e melhora o desempenho do modelo em conjuntos de dados de tamanho limitado [144, pág. 219].

$$SQE_{L_1} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \tag{4.48}$$

4.6.3.3 Regressão por Rede Elástica

A regressão por rede elástica (Enet) também é utilizada na regularização, porém ela combina as duas penalidades a regressão Lasso e a Ridge:

$$SQE_{Enet} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j| \tag{4.49}$$

4.6.4 Regressão por Máquina de Vetor Suporte, SVM

A Máquina de Vetor de Suporte (SVM, do inglês Support Vector Machine) é um algoritmo de aprendizado de máquina supervisionado que pode ser utilizado para tarefas de classificação e regressão. O objetivo desse método procura encontrar um hiperplano que melhor separa os dados em diferentes classes [99]. O SVM tem capacidade de generalização, implementação simples, poucos parâmetros livres, e independência dimensional. A flexibilidade na regressão e a capacidade de função contínua aproximada torna os SVMs muito adequados para estudos QSAR[92]. As típicas funções de kernel, originadas da equação 4.50, são lineares, polinomiais, gaussianas e sigmóides [93]. Com essas funções ocorre o relacionamento linear entre os preditores e o resultado, que podem ser usadas para generalizar o modelo de regressão e abranger funções não lineares dos preditores. Neste trabalho se utilizou a função a kernel de base radial - RBF (equação 4.52), a função linear (equação 4.51) e a função com tangente hiperbólica (equação 4.53).

$$f(\mathbf{u}) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(\mathbf{x_i}, \mathbf{u}) \tag{4.50}$$

onde $K(.)$ é chamado de função kernel.

$$\text{função linear} - K(mx_i, u) = \sum_{j=1}^{P} x_{ij} u_j = \mathbf{x_i'u} \tag{4.51}$$

$$\text{função de base radial} = \exp\left(-\sigma||\mathbf{x} - \mathbf{u}||^2\right) \tag{4.52}$$

$$\text{tangente hiperbólica} = \tanh\left(\phi\left(\mathbf{x_i'u}\right) + 1\right) \tag{4.53}$$

onde α_i , ϕ e σ são parâmetros de escala [87], os parâmetros α_i serão exatamente zero, indicando que eles não têm impacto na equação de predição. O conjunto de dados de treinamento, ou seja, o xij são necessários para novas previsões e $\mathbf{x_i u}$ corresponde ao produto escalar.

4.6.5 Regressão Floresta Aleatória, RF

Random Forest ou Floresta Aleatória, é um método desenvolvido por [94], que através de um conjunto de Árvores de Decisão realiza tarefas de classificação ou regressão. Em vez de construir uma única árvore para realização específica, serão construídas muitas árvores, diferentes entre si, com isso a previsão final será a média da previsão de todas as árvores [95].

O erro associado a cada modelo linear é usado no lugar de SD(S) na equação abaixo em redução da taxa de erro para a próxima divisão.

$$\text{redução} = SD(S) \sum_{i=1}^{P} \frac{n}{n_i} \times SD(S_i) \tag{4.54}$$

O processo de crescimento da árvore continua ao longo dos galhos da árvore até que não haja mais melhorias na taxa de erro ou não haja amostras suficientes para continuar o processo. Uma vez a árvore está totalmente desenvolvida, existe um modelo linear para cada nó da árvore [87].

5

Estudo comparativo entre diferentes algoritmos de aprendizado de máquina na predição de pIC_{50} de inibidores de N-miristoiltransferase de *L. donovani*

5.1 Introdução

Neste capítulo serão avaliados tanto modelos lineares quanto não lineares, na predição de atividade atividade inibitória pIC_{50} sobre a enzima N-miristoiltransferase, em um pequeno banco de dados de moléculas com a mesma classe química.

Neste capítulo, foi utilizada a linguagem R, com diversos pacotes, para análise exploratória, transformação de dados, modelagem preditiva e análise gráfica.

Os pacotes utilizados neste capítulo, bem como um resumo sucinto da sua utilização, e as respectivas referências, são listados na tabela 5.1.

5.2 Análise Exploratória de Dados

Tabela 5.1 – Lista de pacotes, referências e aplicações usadas neste trabalho.

Pacote R	**Função(ões)**	**Referência**
readxl	Importação de arquivos xlsx	[59]
pryr	Tamanho dos objetos (MB)	[60]
e1071	Assimetria e Curtose	[61]
GGally	Gráficos	[62]
tiff	Exportação de figuras	[63]
caret	Treinamento de modelos de classificação e regressão	[64]
corrplot	Mapa de correlações	[65]
stats	Estatística descritiva e inferencial	[56]
lmtest	Análise da regressão	[66]
olsrr	Análise da regressão	[67]
ggplot2	Gráficos	[68]
tidyr	Manipulação de dados	[69]
gridExtra	Gráficos	[70]
chemometrics	Distância de Mahalanobis	[71]
robustbase	Determinante de covariância mínima	[72]
leaps	Regressão por melhores subconjuntos	[73]
car	Fator de inflação da variância	[75]
psych		[77]
FactoMineR	PCA	[78]
pls	PLS	[79]
plsVarSel	VIP-PLS; UVE-PLS	[80]
kernlab	SVM Radial; SVM Linear	[81]
elasticnet	Ridge Regression; LASSO	[82]
neuralnet	ANN	[83]
RSNNS	MLP	[84]
randomForest	Random Forest	[85]

O número de pacotes utilizados está condizente com a filosofia de trabalho da linguagem R, que não é monolítica

A partir das 77 amostras extraídas de [43] foi feita uma análise com base em grupos funcionais dessas estruturas químicas. De acordo com [74], as interações entre fármaco e alvo são bem específicas e a alteração de um grupo funcional pode bloquear a capacidade de a molécula interagir com o alvo, com isso é importante analisar os grupos funcionais. A princípio foi selecionado desse conjunto 51 amostras que continham grupos funcionais de éter, amina, amida, éster e oxadiazol. Ao analisarmos os critérios de QSAR para essas 51 amostra, identificamos que não atendia as exigências. Optou-se então, por excluir o grupo funcional oxadiazol, restando assim, 21 estruturas químicas a serem estudas, entre elas estão as amostras 2, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24 e 25 que são encontradas na tabela do primeiro apêndice 8.
A tabela 5.2 mostra a codificação das estruturas utilizadas extraídas do apêndice 8 e a codificação da ordem das amostras em linguagem R.

Tabela 5.2 – Codificação.

Compostos	Índice R	Compostos	Índice R	Compostos	Índice R
2	1	10	8	18	15
4	2	11	9	19	16
5	3	13	10	21	17
6	4	14	11	22	18
7	5	15	12	23	19
8	6	16	13	24	20
9	7	17	14	25	21

A primeira etapa consistiu em verificar se haviam valores faltantes. Para esta base de dados, não havia casela em branco. Todas as variáveis presentes eram numéricas contínuas e/ou discretas.

Foram removidas variáveis que apresentassem o valor zero em sua composição. Neste caso, as variáveis A8, B6, C2, D3, D4, D6, X6 e X16.

A seguir, foi avaliado o sumário estatístico dos dados, que foram divididos em duas tabelas. Para cada variável, foram obtidos os cinco números de Tukey (o mínimo, o primeiro quartil, a mediana, o terceiro quartil e máximo), a média aritmética e o desvio-padrão amostral, mostrados na tabela 5.3.

Tabela 5.3 – Estatística descritiva da base de dados: resumo dos cinco número de Tukey, média e desvio-padrão amostral.

Var	Mín	Q_1	$\bar{x}$	$\tilde{x}$	Q_3	Máx	s
A1	34,96	39,39	40,84	40,89	43,09	50,61	3,42
A2	-53,06	-46,43	-42,79	-43,14	-41,06	-35,12	3,99
A3	1,00	1,84	2,71	2,542	3,19	3,74	0,80
A4	248,2	268,9	277,20	278,80	285,5	340,50	20,58
A5	1,12	1,15	1,32	1,35	1,51	1,73	0,21
A6	0,04	0,16	0,90	1,169	1,91	2,84	0,97
A7	0,05	0,81	1,07	1,078	1,43	2,35	0,54
B1	38,82	48,97	51,12	50,50	53,75	60,69	6,06
B2	40,23	49,89	53,89	52,33	56,31	62,45	6,23
B3	-312,30	-281,50	-269,40	-261,60	-249,50	-201,10	31,16
B4	0,05	0,17	0,37	0,7167	1,20	2,46	0,74
B5	0,03	0,75	1,02	1,113	1,39	2,90	0,63
C1	63,78	67,93	71,76	71,33	73,25	79,73	4,18
C3	42,76	47,37	51,47	50,72	53,78	59,07	4,68
C4	-9,19	-6,44	-5,35	-5,451	-4,23	-2,01	1,99
C5	-3,88	-1,28	-0,35	-0,4605	0,95	2,62	1,83
D1	73,21	78,76	80,39	81,88	83,60	104,14	6,19
D2	-99,01	-78,42	-76,35	-76,57	-72,58	-66,37	6,97
D5	0,09	0,42	0,55	0,6324	0,78	1,34	0,34
X1	-183,04	44,49	148,16	137,77	233,43	351,46	144,7
X2	-368,9	-127,41	-27,87	-57,64	9,46	168,98	136,84
X3	-275,90	-191,10	-182,80	-195,40	-175,70	-166,20	34,09
X4	-12,45	-12,15	-11,83	-11,82	-11,68	-10,42	0,47
X5	4,56	5,89	6,96	9,283	8,91	24,14	5,66
X7	311,40	326,40	328,50	335,30	342,40	372,40	19,24
X8	-4,09	-3,79	-3,58	-3,62	-3,44	-3,05	0,26
X9	36,00	72,00	72,00	154,30	216,00	648,00	157,80
X10	334,9	349,60	358,10	362,20	376,00	394,20	17,38
X11	332,60	345,00	352,20	356,60	367,10	387,70	16,92
X12	209,30	222,80	239,70	240,30	247,40	335,30	26,18
X13	-114,88	24,01	34,37	35,63	56,99	241,20	67,32
X14	50,96	64,60	66,54	66,49	70,48	72,03	4,79
X15	1,55	2,63	2,83	2,846	3,37	3,88	0,53
X17	24,83	27,65	34,59	33,78	36,87	51,54	6,86
X18	1,44	1,47	1,49	1,493	1,52	1,57	0,03
X19	315,40	329,4	335,60	339,10	351,90	366,30	15,65
X20	208,30	222,6	232,20	231,40	238,80	255,40	12,34
X21	473,20	514,50	530,80	547,80	565,70	662,80	57,13
X22	66,33	67,23	67,96	68,31	69,21	70,91	1,43
X23	3,00	3,00	3,00	3,524	4,00	5,00	0,75
X24	38,97	148,99	159,46	170,26	197,13	301,12	63,29
pIC_{50}	4,63	5,37	5,796	5,731	6,155	6,77	0,60

Observa-se, pelos dados da tabela 5.3, que os dados apresentam unidades distintas, em

função dos valores mínimos e máximos. As variáveis **X1**, **X2**, **X9** e **X24** apresentaram alta dispersão nos dados, conforme visto pela amplitude. O coeficiente de variação, a amplitude, a amplitude interquartil, o coeficiente de assimetria amostral e o excesso de curtose amostral são mostrados na tabela 5.4.

Tabela 5.4 – Estatística descritiva da base de dados: coeficiente de variação, amplitude, amplitude interquartil, assimetria amostral e excesso de curtose amostral.

Var	CV%	A	AIQ	Ass	Curt
A1	8,35	15,65	3,70	0,700	2,34
A2	-9,26	17,94	5,37	-0,2848	0,89
A3	31,33	2,74	1,35	-0,2379	-1,17
A4	7,38	92,26	16,56	1,3105	3,23
A5	15,65	0,61	0,36	0,4166	-1,41
A6	83,21	2,80	1,75	0,3184	-1,29
A7	49,65	2,30	0,62	0,2737	0,58
B1	11,99	21,87	4,78	-0,4461	-0,17
B2	11,91	22,22	6,42	-0,5196	-0,34
B3	-11,91	111,14	32,07	0,5198	-0,34
B4	103,45	2,41	1,03	1,1831	0,19
B5	56,94	2,87	0,64	0,9599	2,04
C1	5,86	15,95	5,32	0,0657	-0,26
C3	9,23	16,31	6,41	0,0902	-0,75
C4	-36,45	7,18	2,21	-0,1493	-0,42
C5	-396,45	6,50	2,23	-0,2882	-0,49
D1	7,56	30,93	4,84	2,3454	8,25
D2	-9,10	32,64	5,84	-1,6294	4,63
D5	53,96	1,25	0,36	0,5563	-0,22
X1	105,03	534,5	188,94	-0,7531	0,01
X2	-237,43	537,88	136,87	-0,6644	0,45
X3	-17,45	109,66	15,44	-1,5786	1,07
X4	-3,97	2,03	0,47	1,4037	2,82
X5	60,96	19,58	3,02	1,7854	2,08
X7	5,74	61,04	16,00	0,5386	-0,53
X8	-7,29	1,04	0,35	0,0191	-0,21
X9	102,28	612,00	144,00	1,9252	3,86
X10	4,80	59,32	26,40	0,3695	-0,90
X11	4,75	55,09	22,09	0,4727	-0,63
X12	10,89	126,07	24,56	2,4322	8,56
X13	188,95	356,08	32,98	0,6192	4,77
X14	7,21	21,07	5,88	-1,7169	4,51
X15	18,65	2,33	0,74	-0,3851	0,76
X17	20,32	26,71	9,21	0,7305	0,60
X18	2,28	0,13	0,05	0,4166	-0,42
X19	4,61	50,94	22,57	0,1663	-0,97
X20	5,33	47,12	16,13	-0,0037	-0,62
X21	10,43	189,63	51,17	0,9221	-0,14
X22	2,1	4,58	1,98	0,4932	-0,80
X23	21,27	2,00	1,00	1,0919	-0,20
X24	37,17	262,15	48,14	0,0167	0,83
pIC_{50}	10,55	2,14	0,79	-0,0166	-0,83

As variáveis **A1**, **A4**, **B1**, **B2**, **C1**, **C3** , **D2**, **X7**, **X10**, **X11**, **X14**, **X18**, **19** e **X20**, **X22**

apresentaram um CV(%) menor que 15%, indicando baixa dispersão. As variáveis **A3**, **A6**, **B4**, **B5**, **D5**, **X1**, **X5**, **X9**, **X13** e **X24** apresentaram um CV(%) maior que 50%, indicando alta dispersão.

As variáveis **B3**, **X1**, **X2**, **X3**, **X9**, **X12**,**13**, **X21** e **X24** apresentaram as maiores amplitudes nas variáveis estudadas. Entretanto, apenas pela amplitude, não podemos inferir sobre a distribuição dos dados.

Em relação ao coeficiente de assimetria amostral, as variáveis **A2**, **A3**, **A5**, **A6**, **A7**, **B1**, **C1**, **C3**, **C4**, **C5**, **X8**, **X10**, **X11**, **X15**, **X18**, **X19**, **X20**, **X22**, **X24** e **pIC**$_{50}$ são aproximadamente normais. As variáveis **A1**, **B3**, **B5**, **D5**, **X7**, **X13**, **X17** e **X21** apresentaram moderada assimetria positiva enquanto as variáveis **B2**, **X1** e **X2** moderada assimetria negativa. Por fim, as variáveis **A4**, **B4**, **D1**, **X4**, **X5**, **X9**, **X12** e **X23** apresentaram forte assimetria positiva, enquanto que as variáveis **D2**, **X3** e **X14** apresentaram forte assimetria negativa.

Em relação ao excesso de curtose amostral, **A3**, **A5**, **A6**, **B1**, **B2**, **B3**, **C1**, **C3**, **C4**, **C5**, **D5**, **X7**, **X8**, **X10**, **X11**, **X19**, **X20**, **X21**, **X22**, **X23** e **pIC**$_{50}$ são platicúrticas. As demais, são leptocúrticas.

A distribuição dos dados das variáveis pode ser visualizado através do gráfico de caixas mostrado na figura 5.1.

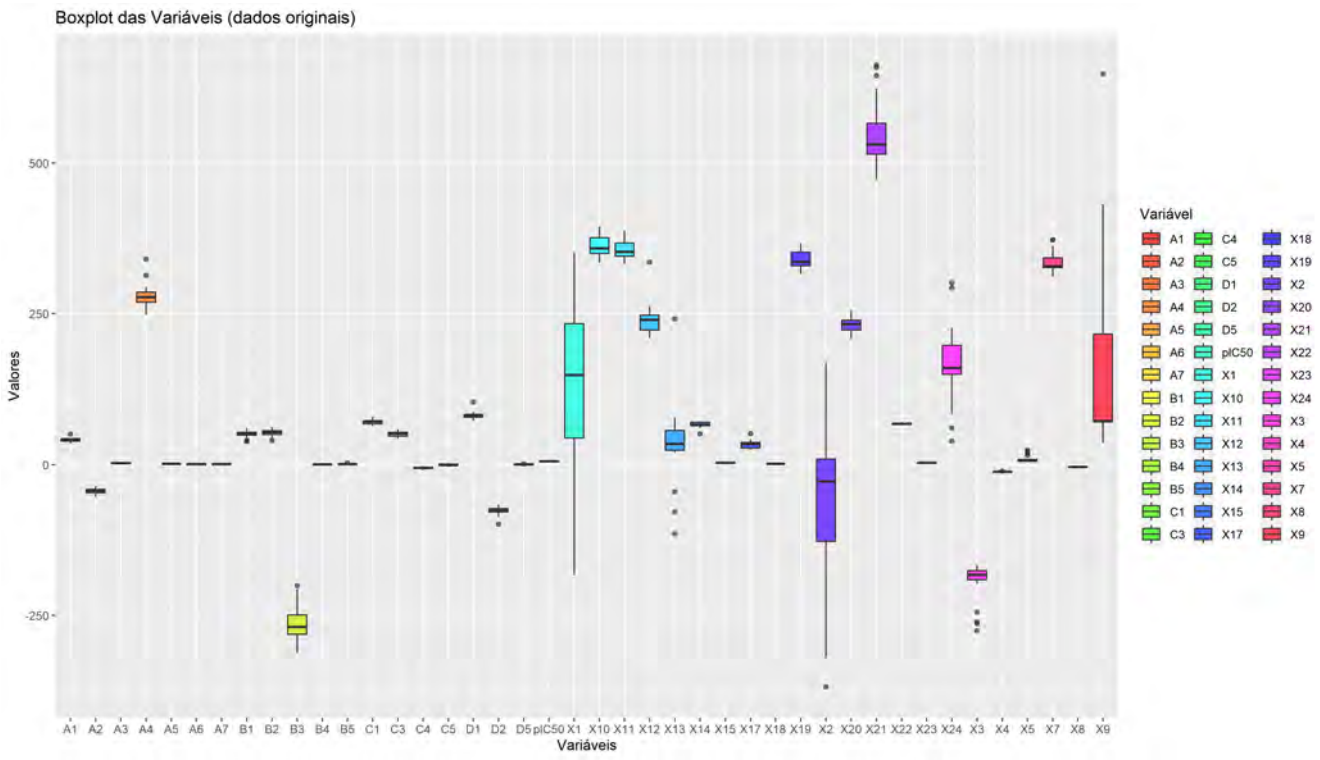


Figura 5.1 – Gráfico de caixas, dados originais

Fonte: A autora, 2023.

Pela magnitude de cada variável, não é possível observar a distribuição dos dados de mais da metade das variáveis. Desta forma, foi construído um gráfico de caixas com dados padronizados através do escore-z, conforme figura 5.2.

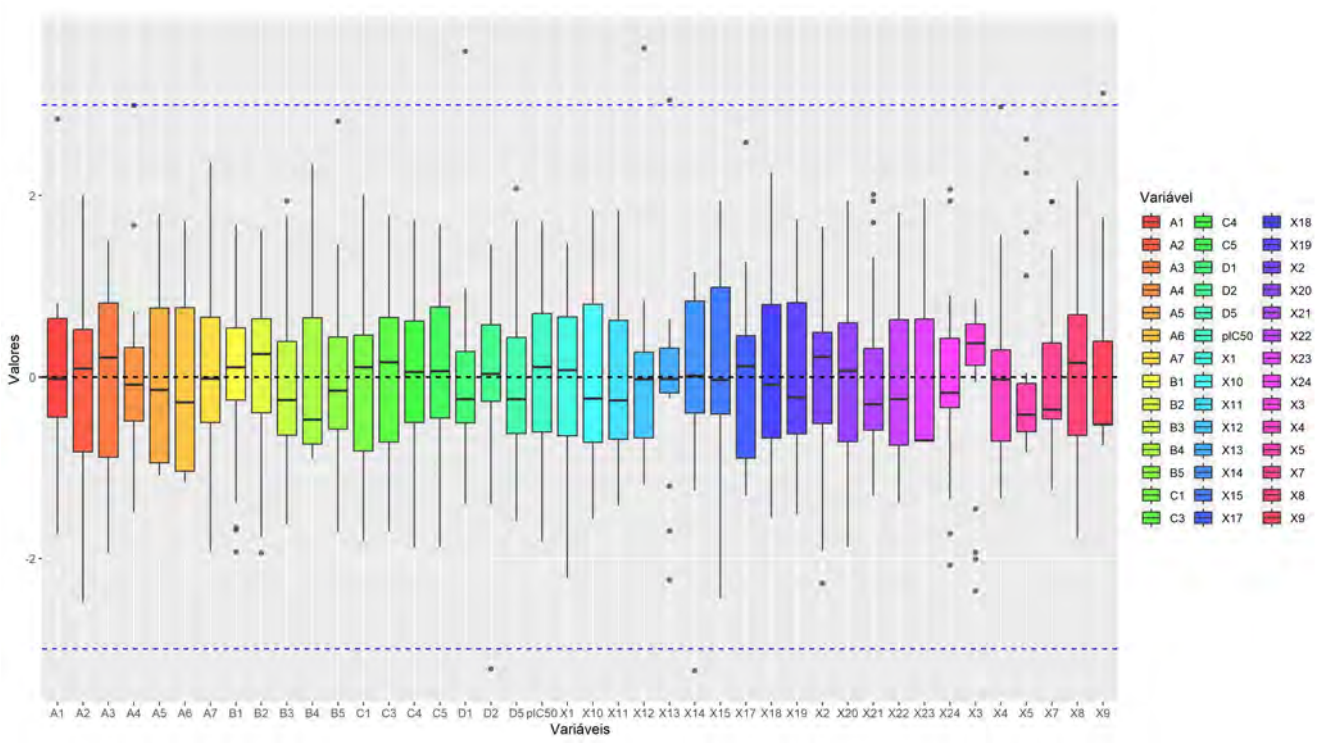


Figura 5.2 – Gráfico de caixas, dados padronizados

Fonte: A autora, 2023.

Na inspeção do gráfico de caixas com dados padronizados, observa-se que as variáveis **D2** e **X14** apresentaram amostras com valores abaixo do limiar de -3,0 unidades de desvio-padrão. A amostra 18 da variável **D2** apresentou um valor de escore-z $Z \cong -3,2197$. Já na variável **X14**, a amostra 12 apresentou um escore-z de $Z \cong -3,2409$. O limite de $\pm 3,0$ unidades de desvio-padrão é representado pela linha tracejada azul horizontal.

As variáveis **D1**, **X9**, **X12** e **X13** apresentaram amostras com valores maiores que o limiar de +3,0 unidades de desvio-padrão. As amostras 18 da variável **D1**, 20 da variável **X9**, 1 da **X12** e 1 da variável **X13** apresentaram, respectivamente, os seguintes valores de escore-z: $Z \cong 3,5930$, $Z \cong 3,1287$, $Z \cong 3,6292$ e $Z \cong 3,0537$,

Por se tratar de uma análise univariada, não é possível saber, de antemão, se as amostras são *outliers* multivariados ou não. Para avaliar, e, por conseguinte, decidir pela permanência ou não destas amostras, será empregada a distância de Mahalanobis para a detecção de *outliers* multivariados [87, pág. 31-33], após a remoção das variáveis preditoras que tenham as maiores correlações lineares (adotou-se o valor de corte $r > |\pm 0,75|$).

Após o emprego deste critério ($r > |\pm 0,75|$) apenas entre as variáveis preditoras, a base de dados reduziu, de 41 variáveis para 23 variáveis candidatas. A figura 5.3 mostra a matriz de correlações desta nova base de dados:

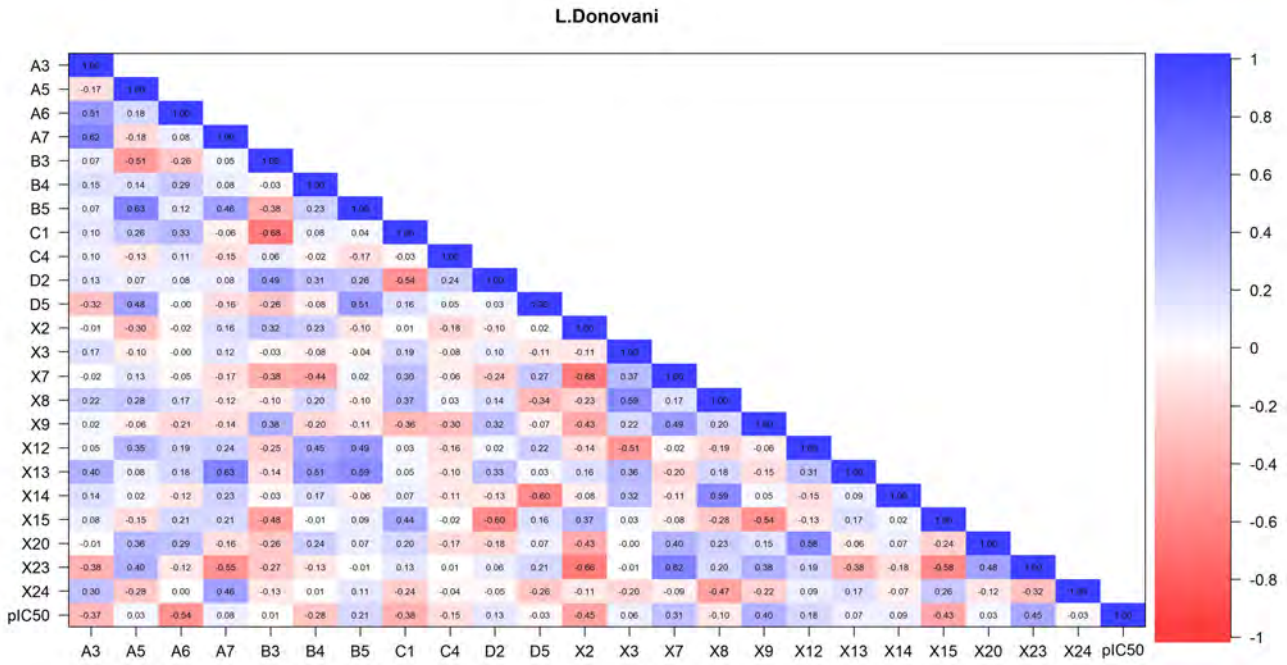


Figura 5.3 – Matriz de correlações (método de Pearson)

Fonte: A autora, 2023

A matriz de correlações contém, no triângulo inferior, os valores das correlações lineares entre todas as variáveis, e na diagonal, o resultado da correlação linear da variável com ela própria, apresentando, desta forma, o valor 1,0 e a cor azul escuro. Nesta representação, quanto mais escura a cor, mais forte a relação. De maneira similar, quanto mais clara, mais fraca a relação. As variáveis **B5** e **A5** apresentam uma correlação linear positiva forte, enquanto as variáveis **X2** e **X7** apresentam uma correlação linear negativa forte. Já as variáveis **X2** e **C1** apresentam uma correlação linear desprezível.

Baseado no resultado da matriz de correlações, deu-se prosseguimento à exclusão de variáveis cujo valor do coeficiente de correlação entre a variável preditora e a variável dependente fosse $r < |\pm 0,3|$ [111].

Foram excluídas as variáveis **A4**, **A5**, **A7**, **B4**, **B5**, **C2**, **C4**, **D2**, **D3**, **D5**, **X5**, **X8**, **X12**, **X13**, **X14**, **X20** e **X24**, por apresentarem uma correlação desprezível. A base de dados passou para oito variáveis independentes (**A3**, **A6**, **C1**, **X2**, **X7**, **X9**, **X15** e **X23**) e uma variável dependente ($\mathbf{pIC_{50}}$).

Passou-se para a investigação dos outliers multivariados. A figura 5.4 mostra, após a construção da distância de Mahalanobis, qual amostra poderia ser um *outlier* multivariado.

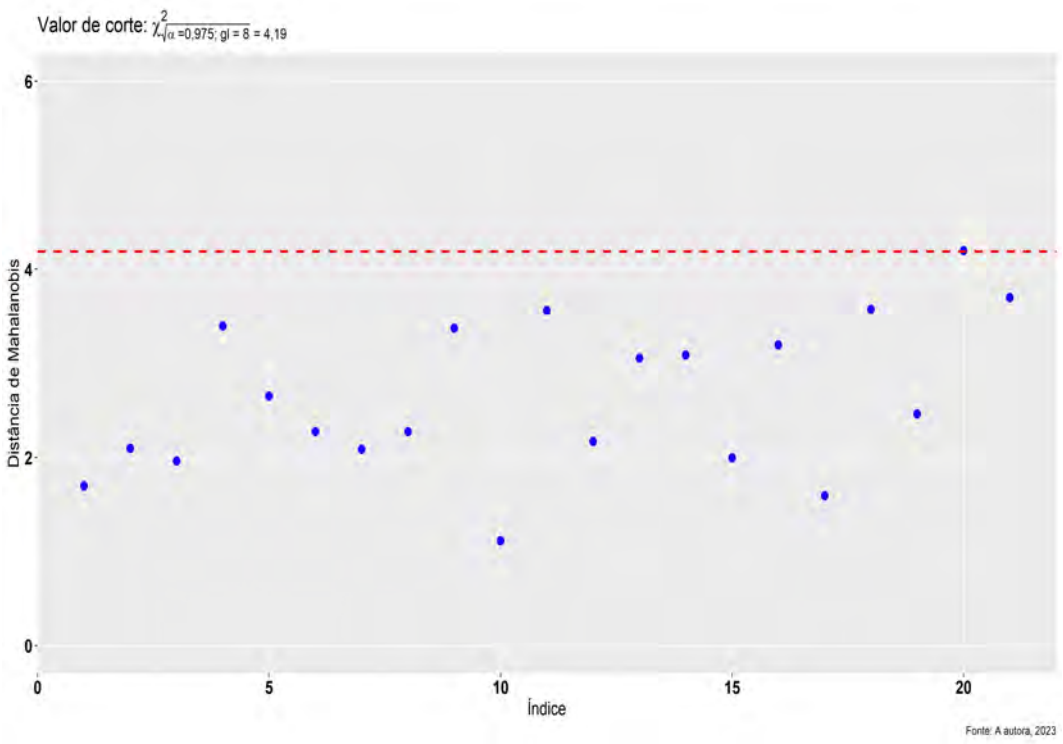


Figura 5.4 – Distância de Mahalanobis

Fonte: A autora, 2023.

A partir de uma distribuição qui-quadrado, usando a equação 4.9, foi adotado um valor de corte:

$$\begin{aligned}\chi^2 &= \sqrt{1-\frac{\alpha}{2};\mathrm{gl}} \\ &= \sqrt{1-\frac{0,05}{2};\ 8} \\ &= \sqrt{(17,53455)} \\ &= 4,187427 \cong 4,19.\end{aligned}$$

A tabela 5.5 mostra os valores das distâncias de Mahalanobis, clássica e robusta, para as amostras:

Tabela 5.5 – Distância de Mahalanobis das amostras.

Indice	Clássico	Robusto	Indice	Clássico	Robusto	Indice	Clássico	Robusto
1	1,6988	0,9077	8	2,274	1,5022	15	1,9974	1,3439
2	2,0961	1,4001	9	3,371	1,7919	16	3,1925	1,7554
3	1,9628	1,2324	10	1,1134	1,3669	17	1,5923	1,5514
4	3,3928	1,7334	11	3,5581	1,7635	18	3,5728	10,4615
5	2,6531	1,8521	12	2,1719	5,8796	19	2,4642	1,2625
6	2,2738	7,8353	13	3,0506	7,971	20	4,1972	9,8293
7	2,0864	1,5392	14	3,0842	8,5408	21	3,695	1,7661

A partir dos resultados obtidos na tabela 5.5, observa-se que a amostra 20 ($d = 4,197155 \cong 4,20$) foi considerada um *outlier* multivariado e removida do banco de dados . Para a base de dados **B**, a remoção de seis amostras (figura 5.5) reduziria o tamanho dos futuros subconjuntos treino e teste (adotando uma partição 70%/30%) para apenas 11 amostras, sem considerar possíveis remoções na etapa de análise de resíduos. Portanto, para este trabalho, o banco de dados **B** não foi considerado.

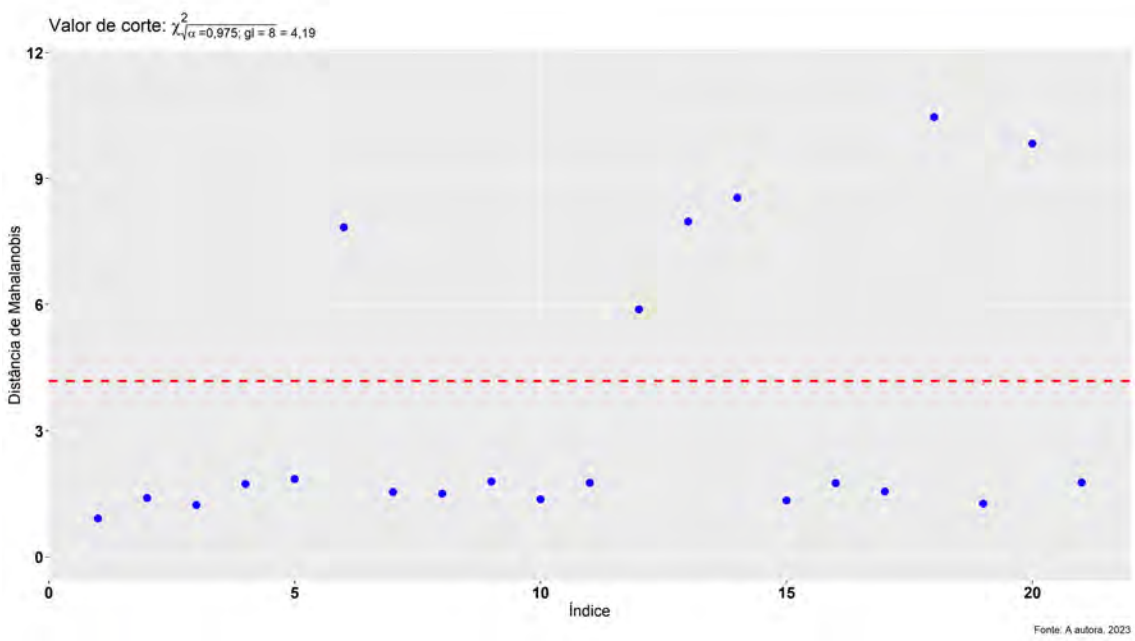


Figura 5.5 – Distância de Mahalanobis (robusta)

Fonte: A autora, 2023.

5.3 Modelagem preditiva

5.3.1 Regressão linear múltipla

Após a exclusão da amostra 20, foi feita a preparação final para a construção do modelo de regressão. Uma das limitações do uso do método dos mínimos quadrados ordinários reside na multicolinearidade, o que pode tornar os coeficentes β's instáveis [112, pág. 261]. Assim, investigou-se o fator de inflação da variância (VIF, *variance inflation factor*). Como valor de corte, apesar da literatura apontar tanto 5 quanto 10, neste trabalho adotou-se o valor 5. A análise VIF foi realizada de acordo com o mostrado em [88, pág. 1310], [107, pág. 460] e [110].

Tabela 5.6 – Fator de inflação da variância: variáveis retidas na base de dados **A**.

iteração	**A3**	**A6**	**C1**	**X2**	**X7**	**X9**	**X15**	**X23**
1ª	2,91	1,79	2,55	2,89	13,30	15,36	13,30	9,15
2ª	2,84	1,79	2,25	2,88	3,71	x	4,35	8,28
3ª	1,38	1,63	1,75	2,52	2,60	x	1,52	x

Na primeira iteração, apesar das variáveis **X7**, **X9**, **X15** e **X23** apresentarem valores de VIF maiores que 5, de acordo com [134, pág. 240], foi removida a variável com o maior valor de VIF. Na segunda iteração, apenas uma apresentou um valor de VIF maior que 5 (variável **X23**).

Na terceira iteração, nenhuma variável apresentou um valor de VIF maior que 5. Desta forma, as variáveis **A3**, **A6**, **C1**, **X2**, **X7** e **X15** foram consideradas aptas para o início da modelagem preditiva. Desta forma, o banco de dados inicial, de tamanho 21×42 apresenta uma dimensão 20×7.

Na base de dados com distância de Mahalanobis robusta (base de dados **B**), a tabela 5.7 mostra os valores de VIF das variáveis preditoras, em cada iteração:

Tabela 5.7 – Fator de inflação da variância: variáveis retidas na base de dados **B**.

	Fator de Inflação da Variância, VIF							
Iteração	**A3**	**A6**	**C1**	**X2**	**X7**	**X9**	**X15**	**X23**
1ª	2,71	2,53	2,52	60,70	65,17	25,43	10,08	25,13
2ª	2,68	2,35	2,21	34,69	x	11,03	6,83	24,63
3ª	2,68	2,09	1,94	x	x	6,99	6,01	3,53
4ª	2,65	2,06	1,85	x	x	x	2,95	2,31

Na primeira iteração, as variáveis **X2**, **X7**, **X9**, **X15** e **X23** apresentaram um valor de VIF maior 5,0. A variável **X7** foi então removida, por apresentar o maior valor.

Na segunda iteração, as variáveis **X2**, **X9**, **X15** e **X23** apresentaram um valor de VIF maior que cinco. A variável **X2** foi removida.

Na terceira iteração, as variáveis **X9** e **X15** apresentaram um valor VIF maior que cinco. A variável **X9** foi removida.

Na quarta iteração, nenhuma variável excedeu o valor de corte VIF. Desta forma, as variáveis preditoras finais, para uso em um modelo de regressão linear múltipla, foram **A3**, **A6**, **C1**, **X15** e **X23**.

Assim, duas bases de dados foram mantidas, denominadas **A** (20×7) e **B** (15×6).

Após a partição do conjunto de dados **A** em treino (70%) e teste (30%), as seguintes amostras foram alocadas para cada conjunto (tabela 5.8).

Tabela 5.8 – Amostras retidas para os conjuntos treino e teste.

Conjunto	Amostras selecionadas	Total
Treino	2, 3, 5, 7, 8, 9, 10, 11, 12 14, 15, 16, 17, 18, 19 e 21	16
Teste	1, 4, 6 e 13	4

A próxima etapa consistiu na obtenção da equação da regressão por mínimos quadrados, bem como realizar testes de hipóteses sobre os coeficientes obtidos e avaliar algumas métricas de qualidade. Na construção do modelo, o método de amostragem aplicado foi a validação cruzada do tipo *leave-one-out*.

A tabela 5.9 mostra os coeficientes β's obtidos, com seus respectivos erros-padrão; os valores dos testes t de cada coeficiente, assim como o valor-p associado a este teste:

Tabela 5.9 – Inferência sobre o primeiro modelo de regressão.

Variável	$\hat{\beta}_j$	$ep(\hat{\beta}_j)$	Teste t	$Pr(>\|t\|)$
Intercepto	6,5717739	2,5597662	2,567	0,0303
A3	-0,3856899	0,1902072	-2,028	0,0732
A6	-0,1271935	0,1531815	-0,830	0,4278
C1	-0,0344850	0,0322111	-1,071	0,3122
X2	-0,0008589	0,0012743	-0,674	0,5172
X7	0,0100563	0,0086193	1,167	0,2733
X15	-0,2282747	0,2539041	-0,899	0,3920

O modelo apresentou um coeficiente de determinação $r^2 = 0,7603$ e um coeficiente de determinação ajustado $r^2_{aj} = 0,6004$. O erro-padrão da regressão foi $\hat{\sigma} = 0,4138$.

A análise da regressão mostrou que, na realização do teste t dos coeficientes angulares, sob a hipótese nula $H_0 : \beta_j = 0$, a um nível de significância $\alpha = 0,05$, com exceção do intercepto, nenhum dos coeficientes β's foi considerado significativo estatisticamente (apesar dos valores r^2 e r^2_{aj}. O teste F geral, obtido a partir da tabela ANOVA, apresentou uma estatística $F = 4,757$, que excede o valor crítico $F(1 - 0,05; \nu_1 = 6; \nu_2 = 9) \cong 3,37$.

A partir dos dados disponibilizados na tabela 5.9, a equação para a regressão linear múltipla com seis variáveis preditoras é:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 \\ \hat{y}_i &= 6,57 - 0,39\mathbf{A3} - 0,13\mathbf{A6} - 0,03\mathbf{C1} - 0,00009\mathbf{X2} + 0,01\mathbf{X7} - 0,23\mathbf{X15} \end{aligned} \tag{5.1}$$

A partir da equação 5.1, foi construída uma reta de previsão, com os dados de treinamento (círculo preenchido em azul turquesa) e de teste (círculo preenchido em vermelho).

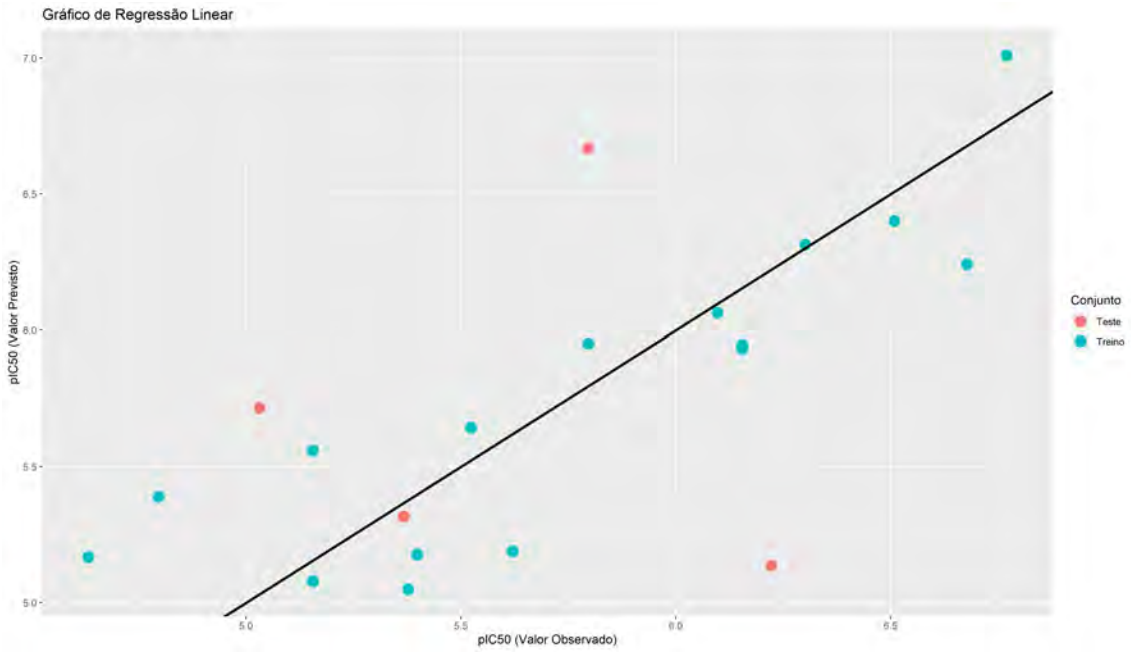


Figura 5.6 – Valores observados versus valores preditos, modelo ML

Fonte: A autora, 2023.

O gráfico mostra que, enquanto na etapa calibração, as amostras apresentaram um ajuste razoável, na etapa teste o ajuste não foi tão bom, onde apenas a amostra 6 (pIC_{50} = 5,366532) apresentou um resíduo pequeno.

Neste ponto, ressalta-se que, na etapa de análise de resíduos, foram construídos modelos de regressão linear múltipla após a exclusão de amostras baseados nos resíduos estudentizados externamente. A tabela 5.10 mostra as métricas de qualidade, o índice da base de dados original, as amostras restantes em cada base de dados e as amostras removidas, nessa investigação. Esta tabela não apresenta os valores de obtidos de cada coeficiente, tampouco os testes de hipóteses associados.

Partindo da base de dados **A**, a análise de resíduos estudentizados externamente indicou que a amostra 12 era um *outlier*. A sua remoção levou ao banco de dados **A(a)**, com 15 amostras. A análise dos resíduos após a construção deste modelo indicou que a amostra 5 era um *outlier*. Este procedimento, de remoção de amostras cujo resíduo estudentizado externamente fosse maior que $|\pm 3,0|$ e construído um novo modelo sem essa amostra, foi repetido até que nenhuma amostra fosse considerada um *outlier*, resultando na base de dados **A(i)**, com apenas 8 amostras.

Tabela 5.10 – Resultados das análises de resíduos dos modelos MLR construídos.

Parâmetros	#	A	A(a)	A(b)	A(c)	A(d)	A(e)	A(f)	A(g)	A(h)	A(i)
	1	2	2	2	2	2	2	2	2	- - -	- - -
	2	3	3	3	3	3	3	3	3	3	3
	3	5	5	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
	4	7	7	7	7	7	- - -	- - -	- - -	- - -	- - -
	5	8	8	8	- - -	- - -	- - -	- - -	- - -	- - -	- - -
	6	9	9	9	9	9	9	9	9	9	9
	7	10	10	10	10	- - -	- - -	- - -	- - -	- - -	- - -
	8	11	11	11	11	11	11	11	11	11	11
	9	12	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -
	10	14	14	14	14	14	14	14	14	14	14
	11	15	15	15	15	15	15	15	15	15	15
	12	16	16	16	16	16	16	- - -	- - -	- - -	- - -
	13	17	17	17	17	17	17	17	- - -	- - -	- - -
	14	18	18	18	18	18	18	18	18	18	18
	15	19	19	19	19	19	19	19	19	19	19
	16	21	21	21	21	21	21	21	21	21	21
Ntreino		16	15	14	13	12	11	10	9	8	8
R2 (C)		0,7603	0,8172	0,8419	0,9141	0,9589	0,9876	0,9959	0,9997	0,999990	0,999925
R2aj		0,6004	0,6802	0,7063	0,8282	0,9095	0,9689	0,9878	0,9988	0,9999	0,9997
RMSE (C)		0,3103	0,2575	0,2106	0,1481	0,0974	0,0558	0,0331	0,0094	0,0018	0,0049
MAE (C)		0,2578	0,2224	0,1735	0,1027	0,0755	0,0465	0,0247	0,0075	0,0016	0,0041
R2 (CV)		0,3906	0,2954	0,4033	0,7663	0,8434	0,9005	0,9346	0,9949	0,9982	0,9977
RMSE (CV)		0,5702	0,6095	0,4715	0,2576	0,1917	0,1596	0,1344	0,0406	0,0246	0,0287
MAE (CV)		0,4859	0,5121	0,4075	0,2166	0,1657	0,1385	0,1040	0,0346	0,0212	0,0228
PRESS		853,3394	5,5728	3,1129	0,8625	0,4412	0,2800	0,1805	0,0148	0,0048	0,0066
Q2		-131,7565	-0,0240	0,2076	0,7402	0,8406	0,8984	0,9329	0,9943	0,9981	0,9974
R2 (V)		0,0058	0,0995	0,0155	0,0511	0,0683	0,0671	0,1135	0,0997	0,0997	0,0936
RMSE (V)		0,7759	0,8635	0,7214	0,5807	0,5452	0,5431	0,5089	0,5252	0,5257	0,5297
MAE (V)		0,6727	0,8101	0,6417	0,5319	0,5024	0,5041	0,4695	0,4795	0,4795	0,4815

Como era de se esperar, a remoção contínua de amostras elevou o valor dos coeficientes r^2 e Q^2, mas piorou a capacidade preditiva (também não se pode descartar, neste momento, que o problema esteja também nas amostras do conjunto teste, visto o resíduo apresentado na figura 5.6). É tentador mostrar um trabalho com um modelo preditivo com valores altíssimos de r^2_{cal} e Q^2, porém, como alertam [88, pág. 1370], seria extremamente antiético, dado que os dados foram *torturados*[2] até confessarem o que queríamos saber!

Chama a atenção que cerca de 50% das amostras do conjunto treino original foram removidas.A inspeção dos coeficientes de regressão do modelo **A(i)** mostrou que apenas a variável **X7** não apresentou significância estatística. A remoção dessa amostra levou a um modelo com duas amostras com resíduos fora dos limites. Essa estratégia mostrou-se inadequada, visto que, caso essas duas amostras fossem removidas, não seria possível obter tanto o erro-padrão quanto calcular o teste t de significância dos coeficientes betas do modelo, bem como não poder calcular o teste F e o r^2_{aj}.

Como são poucas variáveis candidatas, optou-se por uma seleção de variáveis, a partir da base de dados **A**, sem eliminação de nenhuma amostra ($n = 20$, através do método de regressão de melhores subconjuntos. A tabela 5.11 mostra os resultados obtidos. Sete métricas foram avaliadas globalmente, na escolha do melhor modelo: r^2, $r2_{aj}$, Cp de Mallows, critérios de informação de Akaike (AIC) e Bayesiano ou de Schwarz (BIC) e média quadrática do erro de predição, MSEP.

Tabela 5.11 – Resultado das Regressões de Melhores Subconjuntos

	Parâmetros de mérito						
Modelos	r^2	r^2_{aj}	r^2_{pred}	Cp	AIC	BIC	MSEP
A3	0,39	0,35	0,23	10,77	28,81	31,13	4,46
A3 + X2	0,60	0,54	0,43	4,99	24,12	27,21	3,18
A3 + C1 + X7	0,70	0,63	0,54	3,18	21,43	25,30	2,59
A3 + C1 + X7 + X15	0,74	0,64	0,51	3,91	21,51	26,14	2,52
A3 + A6 + C1 + X7 + X15	0,75	0,62	0,45	5,45	22,75	28,16	2,68
A3 + A6 + C1 + X2 + X7 + X15	0,76	0,60	0,19	7,00	23,96	30,14	2,87

Cinco modelos candidatos foram selecionados. Os coeficientes de determinação variaram de 0,39 a 0,76, mostrando que mesmo com todas as variáveis, o modelo só conseguiria explicar 76% da variabilidade na variável dependente $\mathbf{pIC}_{50}$. Tomando como ferramentas de apoio à decisão os critérios de informação de Akaike (AIC) e de Schwarz (BIC), ebm como o MSEP, o modelo a ser trabalhado contém as variáveis **A3**, **C1** e **X7**.

Assim, o modelo de regressão linear múltiplo assume a equação de regressão:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(A3) + \hat{\beta}_2(C1) + \hat{\beta}_3(X7) \tag{5.2}$$

Neste ponto, é sempre bom lembrar da sabedoria de genial George Box dita em 1976: *All models are wrong, some are useful* [113].

Com as três variáveis retidas, foi construído um novo modelo de regressão, com o intuito de se avaliar as variáveis selecionadas apresentariam coeficientes significativos (conforme tabela 5.12). O modelo apresentou um coeficiente de determinação $r^2 = 0,7022$, um coeficiente

[2] Trecho adaptado da entrevista realizada por: Denise Becker, D.; Lima, S. M. Marcelo Soares: "Torturados, os números dizem qualquer coisa". objETHOS, Edição 1085, de 28/04/2020. Disponível em: <https://objethos.wordpress.com/2020/04/23/marcelo-soares-torturados-os-numeros-dizem-qualquer-coisa/>. Acesso: 05/02/2024.

de determinação ajustado $r^2_{aj} = 0,6277$ e o erro-padrão da regressão $\hat{\sigma} = 0,3994$. A figura 5.7 mostra a reta de regressão, com dados dos conjuntos treino e teste, a partir do modelo selecionado.

Tabela 5.12 – Inferência sobre o modelo de regressão selecionado.

Variável	$\hat{\beta}_j$	$ep(\hat{\beta}_j)$	Teste t	$Pr(>\|t\|)$
Intercepto	6,109719	1,990828	3,0690	0,00974
A3	-0,476665	0,139443	-3,418	0,00509
C1	-0,063694	0,024275	-2,624	0,02223
X7	0,016207	0,005214	3,1090	0,00904

Ao realizar o teste t para cada coeficiente, em todos os testes o valor-p obtido foi menor que o nível de significância adotado, o que indica que os coeficientes são estatisticamente significativos. O teste F geral da regressão apresentou uma estatística de teste $F = 9,431$, que excede o valor crítico $F(1 - 0,05; \nu_1 = 3; \nu_2 = 12) = 3,49$, o que indica que a regressão é significativa. Os coeficientes de determinação e determinação ajustado foram listados na tabela 5.12.

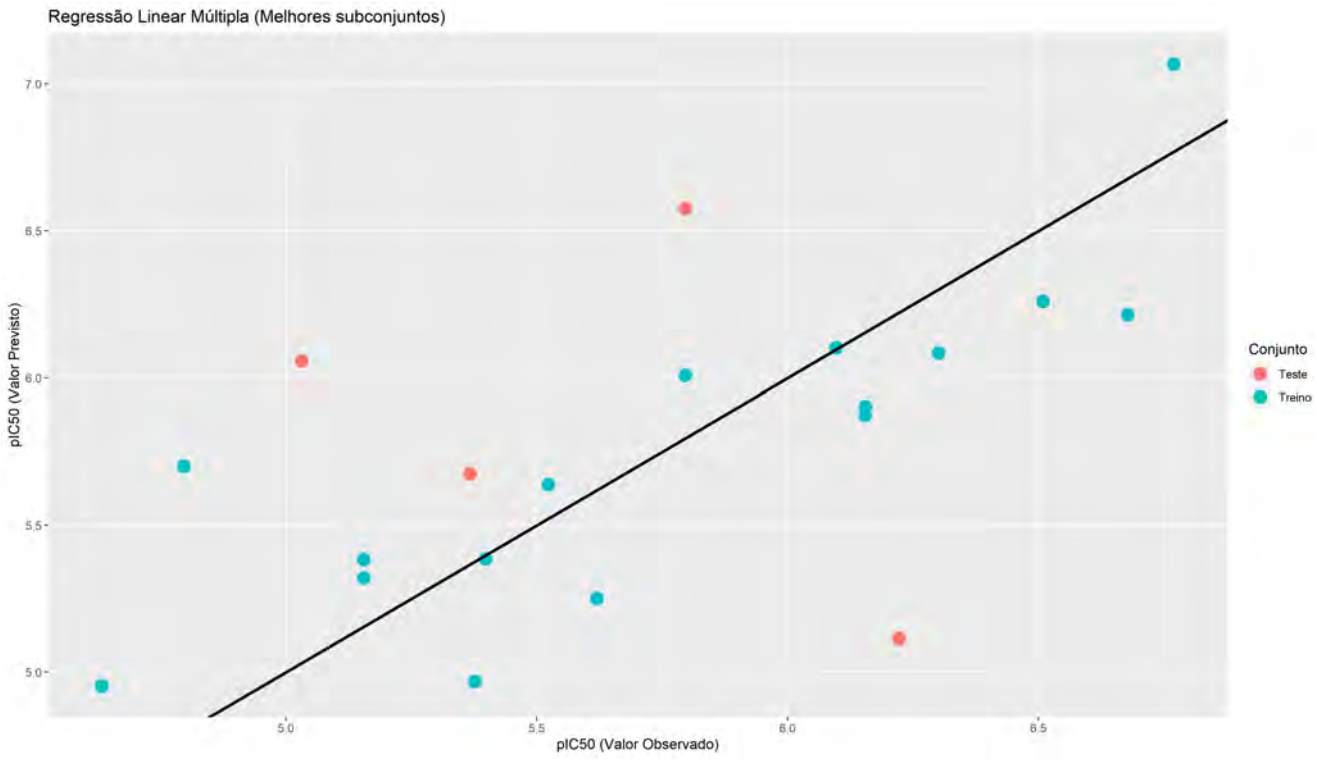


Figura 5.7 – Valores observados versus resíduos observados, no modelo de regressão linear múltipla com melhores subconjuntos.

Fonte: A autora, 2023.

Foi investigada, após o estabelecimento do modelo candidato, se os pressupostos do modelo de mínimos quadrados ordinários foram violados ou não [88, 112]. Para todos os testes de hipóteses, adotou-se um nível de significância $\alpha = 0,05$.

Para a normalidade de resíduos, foi empregado o teste de Shapiro-Wilk com modificação de Royston [114]. Sob a hipótese nula $H_0 : e_i \sim N(0, \sigma^2)$ e a hipótese alternativa H_1 : c.c., a estatística do teste foi $W = 0,91883$, com valor-p = $0,1615$. Desta forma, este pressuposto não foi violado.

A distribuição dos resíduos e a aderência à distribuição normal podem ser verificadas através de gráficos quantil-quantil, mostrados nas figuras 5.8 (resíduos ordinários) e 5.9 (resíduos estudentizados). Em ambos os gráficos, a maior parte das amostras estão próximas da linha teórica. Uma amostra encontra-se afastada da linha, entretanto, está dentro da área hachurada que representa o intervalo de confiança. A amostra **12**, mostrada na figura 5.9, mostra que, como resíduo ordinário, a amostra encontra-se dentro de um intervalo de confiança de 95%, porém, como resíduo estudentizado externamente, não.

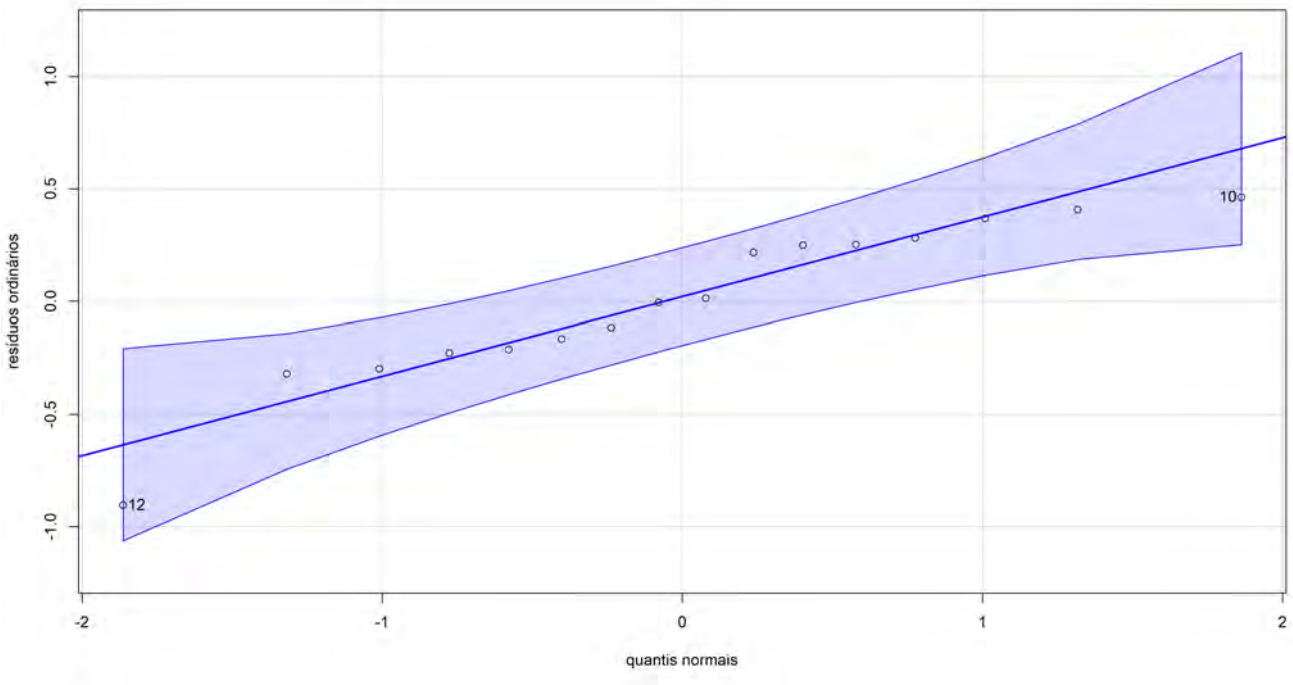


Figura 5.8 – Gráfico quantil-quantil dos resíduos ordinários

Fonte: A autora, 2023.

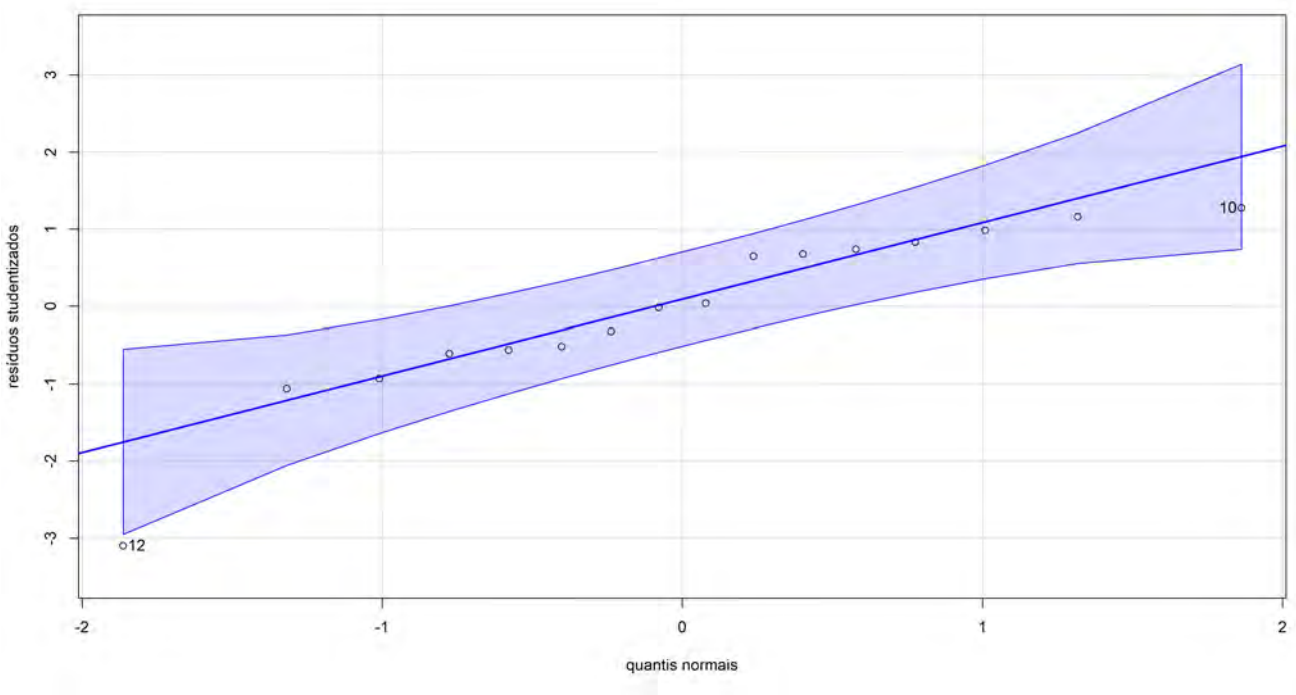


Figura 5.9 – Gráfico quantil-quantil dos resíduos estudentizados externamente

Fonte: A autora, 2023.

Para a homocedasticidade, foi empregado o teste studentizado de Breusch-Pagan [109]. A estatística do teste foi $BP = 0,63149$, com três graus de liberdade e valor-p = $0,8892$. Desta forma, os resíduos são homocedásticos, ou seja, apresentam variância constante.

Para a ausência de autocorrelação serial, foi empregado o teste Durbin-Watson [88]. A estatística do teste foi $DW = 1,582$, com valor-p = $0,1551$. Entretanto, o teste de Durbin-Watson apresenta duas regiões inconclusivas, e apenas a guiagem pelo valor-p pode, neste caso, não fornecer a decisão correta. Para um modelo de regressão com três variáveis independentes ($k = 3$) e $\alpha = 0,05$, temos o limite inferior da região inconclusiva $d_{Inf} = 0,86$ e o limite superior da região inconclusiva $d_{Sup} = 1,73$. Desta forma, não podemos concluir sobre se este pressuposto foi violado ou não [33, pág. 207].

Avaliou-se também a influência (*leverage*) versus os resíduos de Student. Caso uma amostra apresentasse tanto um alto resíduo quanto uma forte influência, esta amostra seria removida do conjunto.

Conforme pode-se observar na figura 5.10 , a amostra **8** apresenta uma alta influência ($d_i =$) sobre o modelo, mas não deve ser excluída, pois está dentro do intervalo de resíduos studentizados. Já a amostra 9 apresenta um alto valor de resíduo ($t_i = -3,1006$) mas um valor de leverage ($d_i = 0,0846$) abaixo do limiar de 0,5. Cabe aqui ressaltar que as amostras 8 e 9 sequenciais da partição treino são, respectivamente, as amostras rotuladas 11 e 12, da base de dados antes da partição. Essa representação gráfica é oriunda do pacote olsrr.

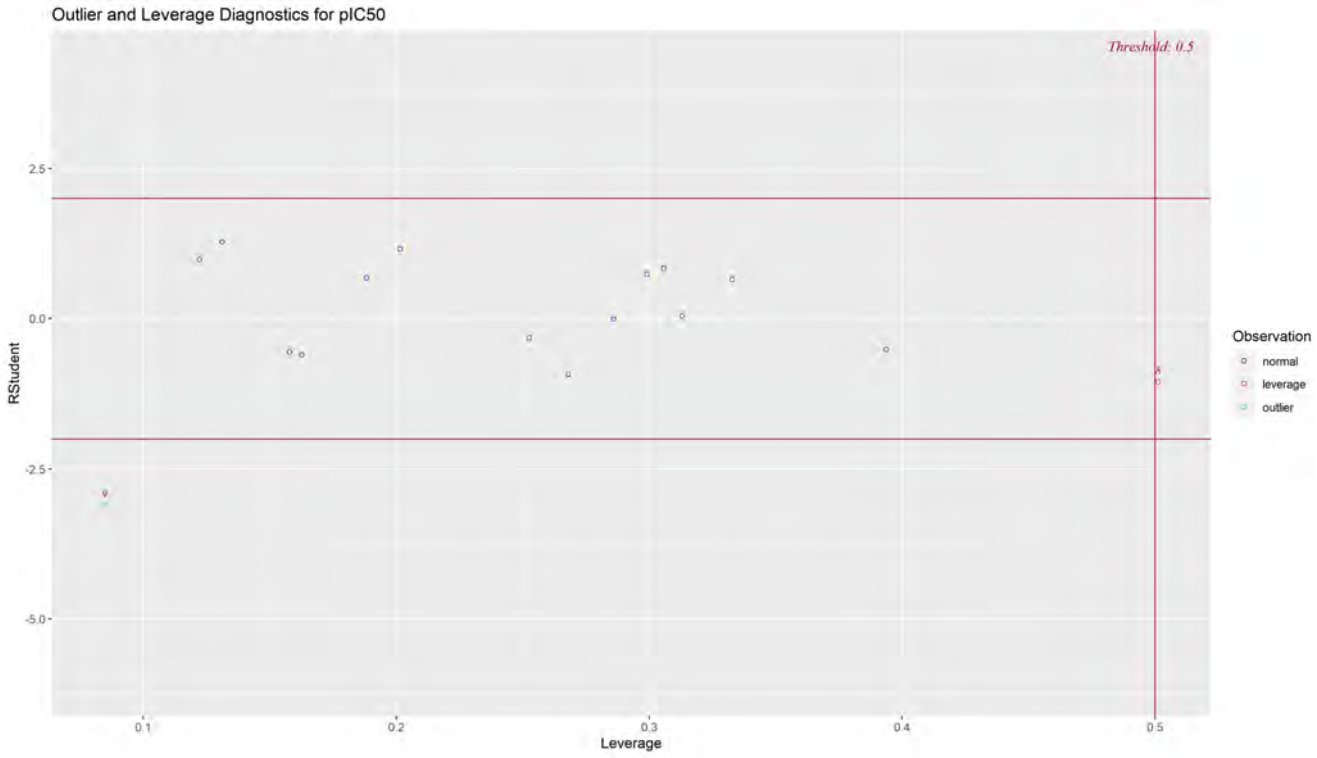


Figura 5.10 – Leverage versus resíduos de estudentizados externamente

Fonte: A autora, 2023.

A partir do gráfico, observa-se que a amostra 08 apresenta um alto valor de leverage, pois encontra-se acima do limiar de 0,5, o que indica que ela pode ser uma amostra influente. Já a amostra 09 aparece abaixo do limite de -3,0 unidades de desvio-padrão, indicando que pode ser um *outlier*. Entretanto, nenhuma amostra caiu no córner superior direito ou no córner inferior direito, o que seria algo problemático para o modelo. Dessa forma, todas as amostras foram mantidas, nesta investigação preliminar.

Essas amostras podem ser melhor identificadas a partir das figuras 5.11 e 5.12. A figura 5.11 mostra que a nona amostra, o gráfico de barra vermelho, está abaixo do limite de três desvios-padrão, o que indica que é um outlier. Já a figura 5.12, a distância de Cook, indica que a amostra 8 é uma amostra influente no modelo.

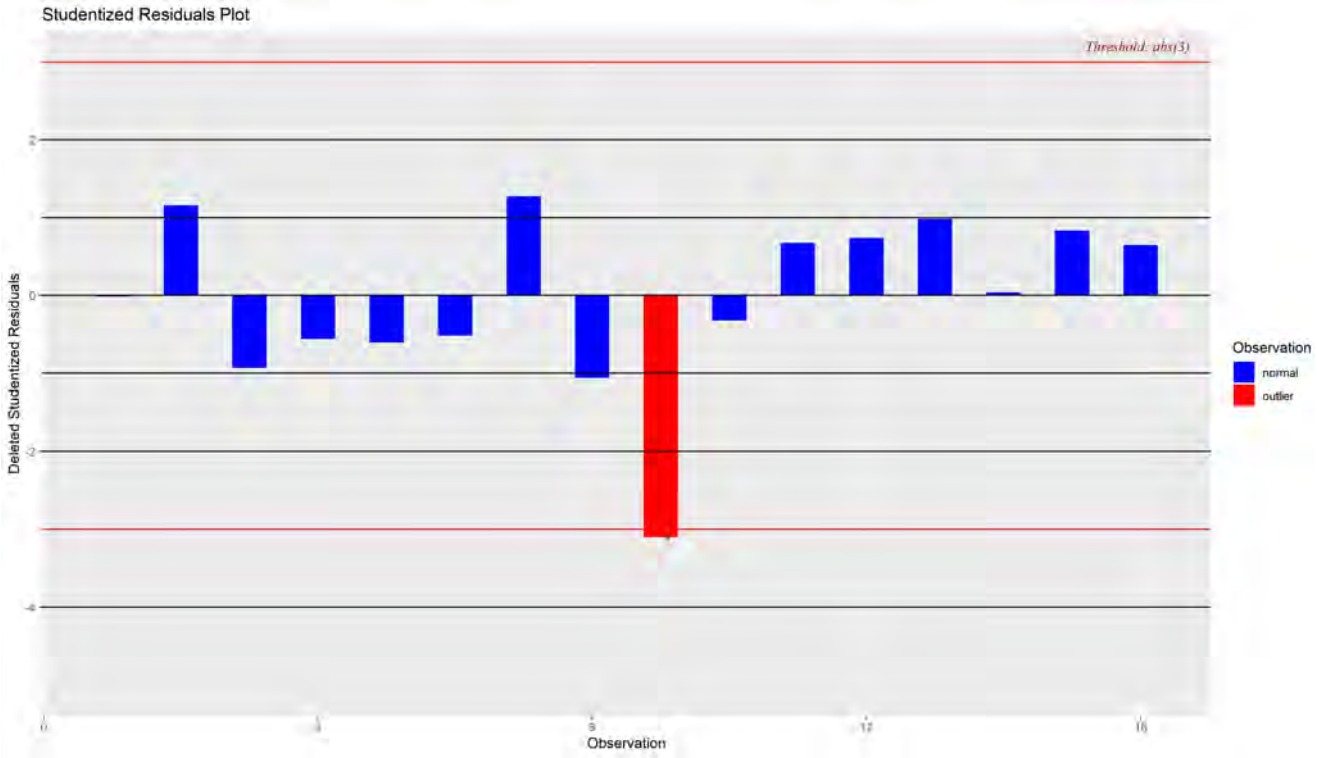


Figura 5.11 – Resíduos estudentizados externamente

Fonte: A autora, 2023.

Após todas as etapas, foi feita a aleatorização em y, permutando-se os valores e calculando-se novas regressões. Cada coeficiente de determinação permutado foi armazenado, para se obter o coeficiente de determinação permutado médio. Foram gerados mil modelos permutados. O resultado do coeficiente de determinação permutado médio, r_p^2, bem como outras figuras de mérito, são mostrados na tabela 5.13.

Tabela 5.13 – Figuras de mérito para o modelo regressão por melhores subconjuntos.

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,7022	r^2_{CV}	0,5589	r^2	0,1685
r^2_{aj}	0,6277	- - -	- - -	- - -	- - -
RMSE	0,3459	RMSE	0,4311	RMSE	0,8633
MAE	0,2813	MAE	0,3663	MAE	0,8054
Bias	3,33E-16	Bias	-0,0067	Bias	-0,2517
- - -	- - -	Q^2	0,5373	- - -	- - -
- - -	- - -	PRESS_{CV}	2,9740	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{CV}}$	0,1437	- - -	- - -
- - -	- - -	r_p^2	0,1604	- - -	- - -
- - -	- - -	${}^c r_p^2$	0,6168	- - -	- - -

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $R_p^2 = 0,160399739 \cong 0,1604$ e o coeficiente de determinação permutado médio corrigido, ${}^cR_p^2 = 0,616786284 \cong 0,6179$. De acordo com [76], os modelos com ${}^cR_p^2 > 0,5$ são

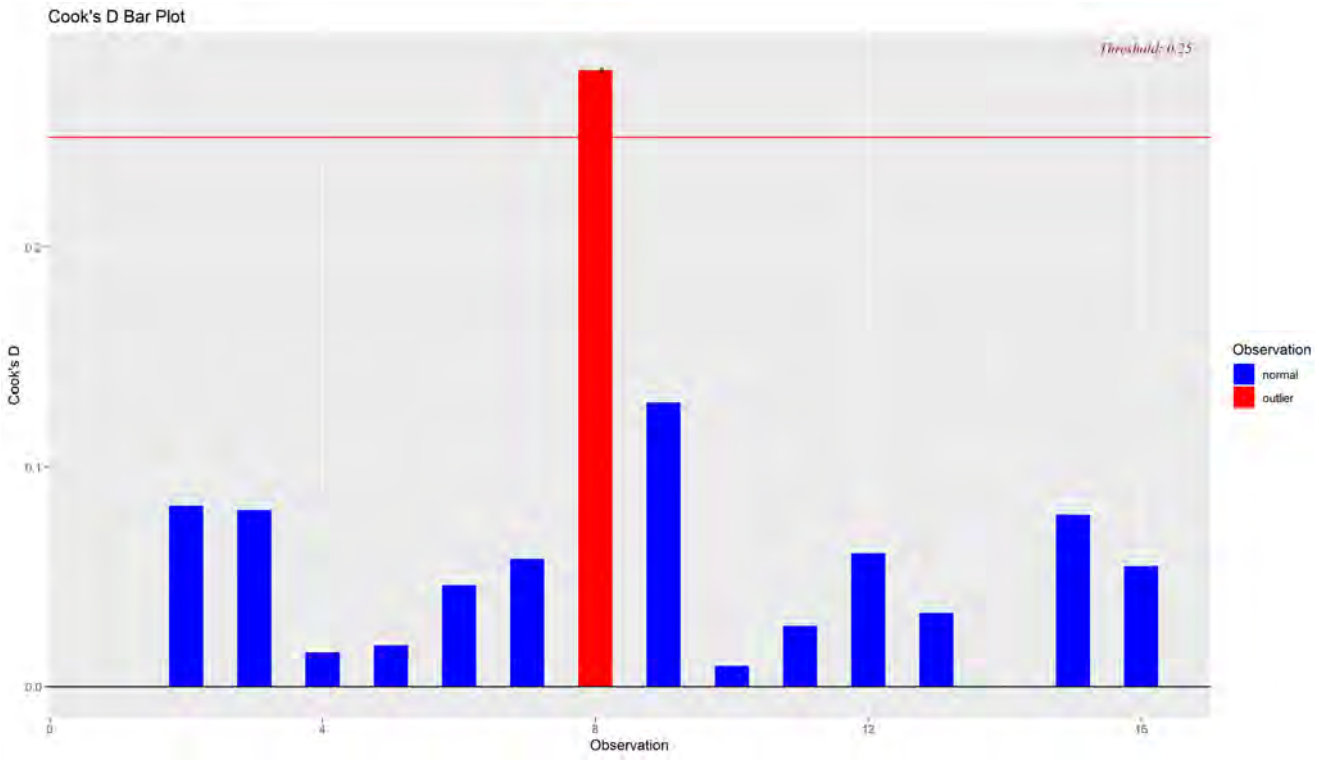


Figura 5.12 – Distância de Cook

Fonte: A autora, 2023.

considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso.

Além das métricas avaliadas, em específico da área de QSAR, vale ressaltar que o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_{\text{cal}} = 0,5669 > Q^2 = 0,3148$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC $= 0,4079 <$ RMSEP $= 0,6568$. Por fim, a diferença entre o r^2 e o Q^2 ($r^2_c - Q^2 = 0,2521$) foi menor que o intervalo de 0,2-0,3, o que demonstra que o modelo não tem sobreajuste [111]. O modelo obtido não passou nos critérios $Q^2 > 0,5$ e $r^2_c > 0,6$.

Desta forma o modelo obtido e os parâmetros estatísticos referente a ele são:

$$\text{pIC}_{50} = 6,1097 - 0,4767 \times \text{CL} - 0,0637 \times \text{VTA} - 0,0162 \times \text{P}$$

$$n = 16 \quad r^2 = 0,7022 \quad Q^2 = 0,5373 \quad p = 0,0208$$

Onde *CL* é a contribuição lipofílica; *VTA* é o valor total de aptidão GoldScore do ligante docado; *P* é o Peso em amu ; n é o tamanho amostral utilizado na construção do modelo; r^2 é o coeficiente de determinação do modelo; Q^2 é o coeficiente de determinação da validação cruzada; e p é o valor-p do teste de significância do modelo de regressão.

Pode chamar a atenção o valor do r^2 da etapa de validação ser baixo, entretanto, deve-se lembrar que o conjunto contém *apenas* quatro amostras. Caso o conjunto fosse maior, ou mesmo contivesse outras amostras, a modelagem poderia ter um melhor ajuste.

A seguir, investigou-se o impacto que a amostra **9** (rótulo original), conforme mostrado na figura 5.10, teria sobre o modelo, ao ser removida. A tabela 5.14 mostra os resultados das etapas de calibração, de validação cruzada e de validação, após a exclusão dessa amostra. O

modelo apresentou um erro-padrão da regressão $\hat{\sigma}^2 = 0,3047$. O teste F de qualidade do modelo apresentou uma estatística $F = 15,8668$, que excede o valor crítico $F(1-0,05;\nu_1 = 3;\nu_2 = 11) = 3,5874$. O valor-p deste teste foi 0,0003.

Tabela 5.14 – Figuras de mérito para o segundo modelo BS-MLR

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,8123	r^2	0,6530	r^2	0,1660
r^2_{aj}	0,7611	r^2_{aj}	- - -	r^2_{aj}	- - -
RMSE	0,2610	RMSE	0,3702	RMSE	0,8674
MAE	0,2328	MAE	0,3232	MAE	0,8188
Bias	0,0000	Bias	-0,0272	Bias	-0,2869
- - -	- - -	Q^2	0,6222	- - -	- - -
- - -	- - -	PRESS_{cv}	2,0561	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	0,1304	- - -	- - -
- - -	- - -	r^2_p	0,1705	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,7220	- - -	- - -

Neste modelo, vale ressaltar que o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_{\text{cal}} = 0,8123 > Q^2 = 0,6222$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,2610 < RMSEP = 0,8674. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,1901$) foi menor que o intervalo de 0,2-0,3 [111], o que demonstra que este modelo também não tem sobreajuste. A figura a seguir mostra os valores observados versus os valores predutos de pIC_{50}, através deste modelo.

Na etapa de avaliação dos pressupostos, o teste de Shapiro-Wilk-Royston apresentou uma estatística $W = 0,94886$, com valor-p = 0,5066. A um nível de significância $\alpha = 0,05$, não há evidência para rejeitar a hipótese nula de que os resíduos seguem uma distribuição normal.

O teste de Durbin-Watson apresentou uma estatística $DW = 2,1952$, com valor-p = 0,6081. A primeira região inconclusiva apresentou o valor crítico inferior $d_i = 0,82$ e valor crítico superior $d_s = 1,75$. A segunda região inconclusiva apresentou o valor inferior $4 - d_s = 2,25$ e o valor superior $4 - d_i = 3,18$. A um nível de significância $\alpha = 0,05$ e $k = 3$, não há evidência para rejeitar a hipótese nula de que os resíduos não apresentam autocorrelação.

O teste de Breusch-Pagan apresentou uma estatística $BP = 0,56095$, com valor-p = 0,9053. A um nível de significância $\alpha = 0,05$, não há evidência para rejeitar a hipótese nula de que os resíduos são homocedásticos.

A figura 5.14 mostra o gráfico quantil-quantil dos resíduos estudentizados, com intervalo de confiança de 95%. Observa-se que todos os resíduos estudentizados externamente estão dentro do intervalo de confiança e que tem aderência à distribuição normal.

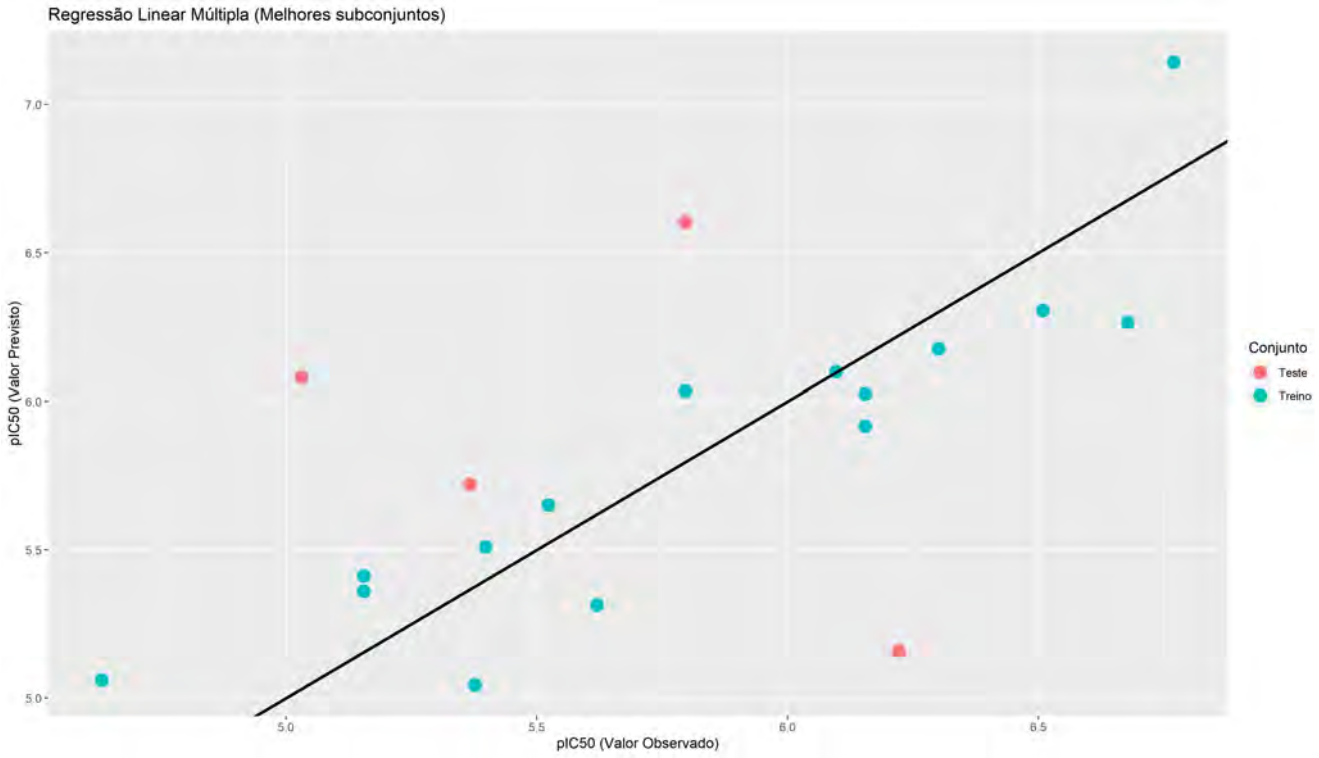


Figura 5.13 – Valores observados vs preditos de pIC_{50}

Fonte: A autora, 2023.

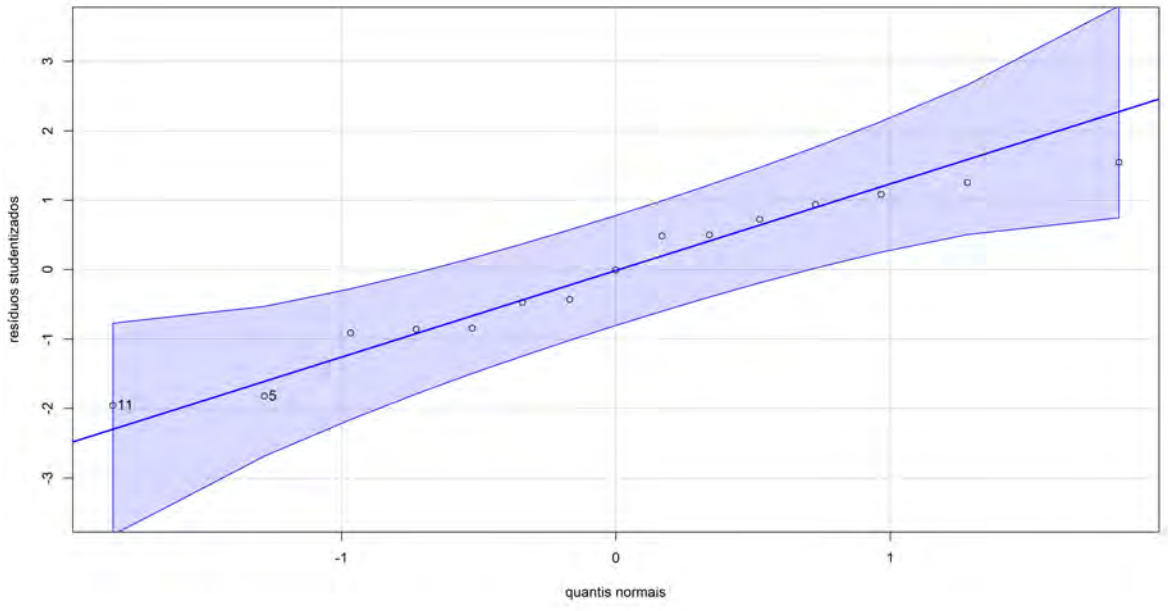


Figura 5.14 – Gráfico quantil-quantil dos resíduos estudentizados externamente

Fonte: A autora, 2023.

A figura 5.15 mostra o gráfico de leverage versus resíduos de estudentizados externamente. Nenhuma amostra excedeu os valores de corte de mais ou menos dois desvios-padrão.

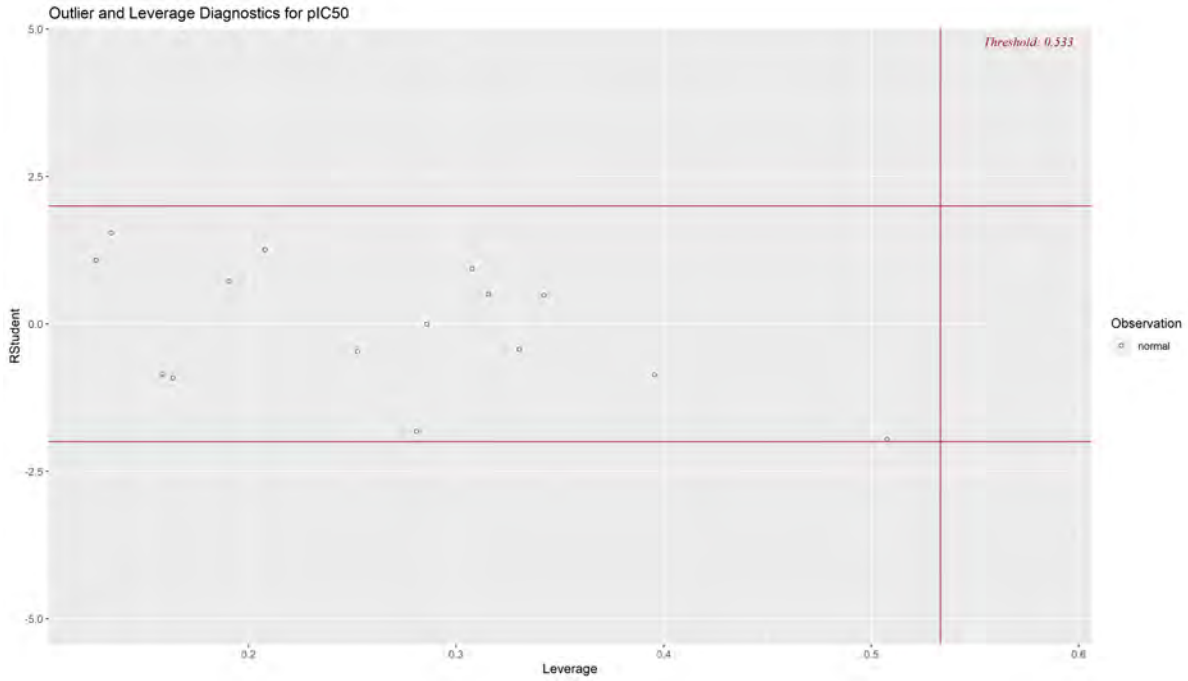


Figura 5.15 – Leverage versus resíduos estudentizados externamente

Fonte: A autora, 2023.

A figura 5.16 mostra o gráfico de resíduos estudentizados externamente. Observa-se que todos os resíduos estão dentro do intervalo de até dois desvios-padrão (linhas horizontais sólidas em preto) e nenhuma amostra ultrapassa o limiar de $\pm 3,0$ desvios-padrão.

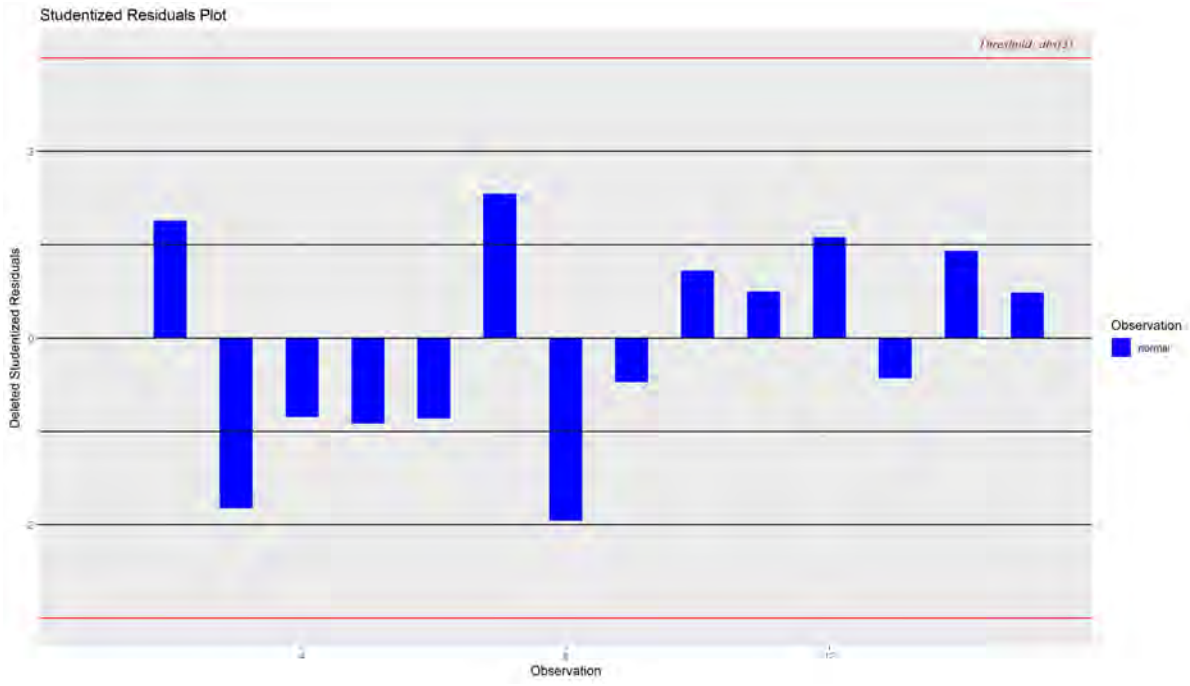


Figura 5.16 – Resíduos estudentizados externamente

Fonte: A autora, 2023.

A figura 5.17 mostra o gráfico da Distância de Cook. Duas amostras ultrapassaram o limiar: as amostras 3 e 8. Isso significa que estas amostras possuem uma grande influência no modelo.

De acordo com Breiman (2001) *apud* [100] existem duas culturas na modelagem preditiva: a primeira, denominada *data modeling culture*, com forte apelo na comunidade estatística, com aplicações de testes de hipóteses e intervalos de confiança para os parâmetros estimados; e a segunda, *algorithmic modeling culture*, que impera na comunidade de aprendizado de máquina. Como observam os autores:

> *Não se assume que o modelo utilizado para os dados é correto; o modelo é utilizado apenas para criar bons algoritmos para prever bem novas observações. Muitas vezes não há nenhum modelo probabilístico explícito por trás dos algoritmos utilizados.*

Para prever bem novas observações, é necessário ter boas amostras, pois não existe modelo preditivo que consiga modelar dados ruins de forma satisfatória a atender diferentes requisitos.

O presente modelo proposto foi extremamente afetado tanto pela quantidade quanto a qualidade do banco de dados utilizado na etapa teste. A seguir, serão investigados outros modelos preditivos, sem as restrições requeridas pelo método dos mínimos quadrados (exceto, obviamente, pelos métodos de regularização, como a regressão ridge, por exemplo).

5.3.2 Regressão por mínimos quadrados parciais

Com o intuito de contornar algumas restrições da regressão por mínimos quadrados ordinários, como ausência de multicolinearidade, foi empregada a regressão por mínimos qua-

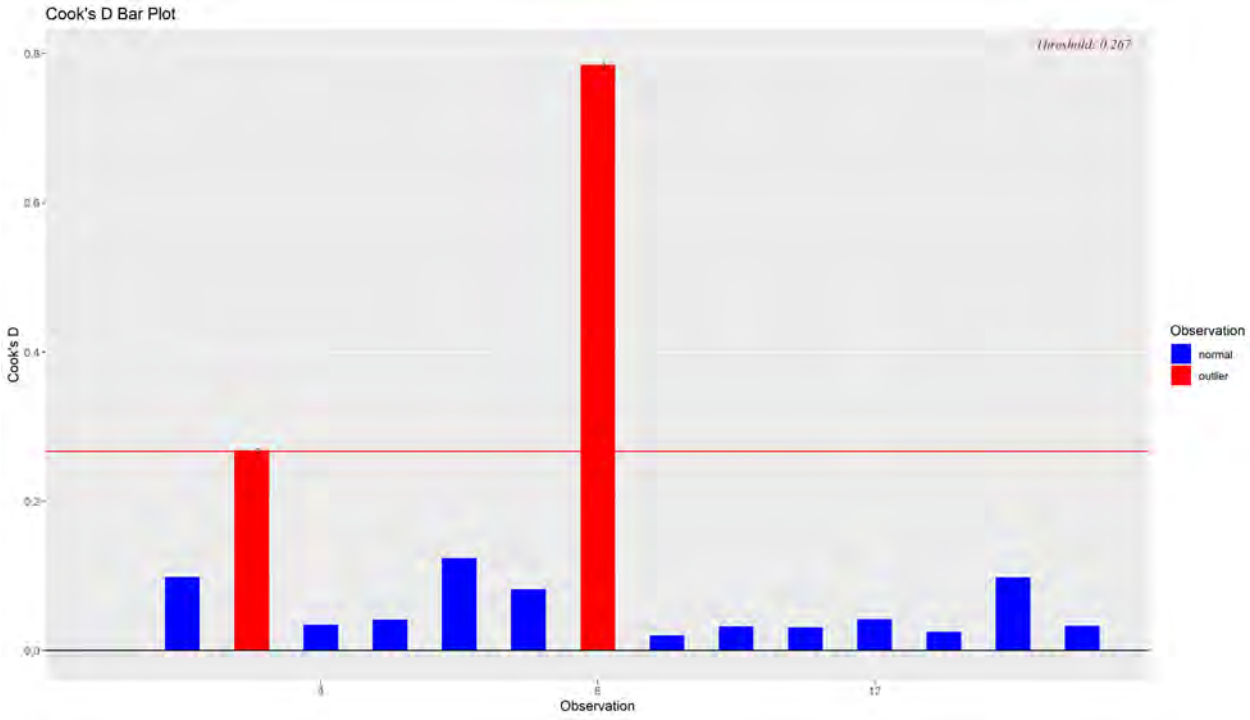


Figura 5.17 – Distância de Cook

Fonte: A autora, 2023.

drados parciais [102]. Porém, algumas transformações foram efetuadas na base de dados, de tamanho 21x42: as variáveis cujas razões máxímo/mínimo fossem maiores que 30 [87], foram transformadas através da transformação Box-Cox. A tabela 5.15 mostra as razões obtidas:

As variáveis **A6**, **A7**, **B4** e **B5** ultrapassaram o valor de corte, conforme [87] (respectivamente, 71; 47; 49,2; e 96,67). Foi então aplicada a transformação Box-Cox nessas variáveis. Os valores de lambda adotados foram, respectivamente: 0,4; 0,8; 0,1 e 0,6. A figura 5.18 mostra o histograma das variáveis antes (linha superior) e após a aplicação da transformação Box-Cox (linha inferior) das variáveis selecionadas.

Tabela 5.15 – Razões máximo/mínimo maiores que 30 nas variáveis preditoras e dependente [87]

Var	R	Var	R	Var	R	Var	R
A1	1,45	**B5**	96,67	**X4**	0,84	**X17**	2,08
A2	0,66	**C1**	1,25	**X5**	5,29	**X18**	1,09
A3	3,74	**C3**	1,38	**X7**	1,2	**X19**	1,16
A4	1,37	**C4**	0,22	**X8**	0,75	**X20**	1,23
A5	1,54	**C5**	-0,68	**X9**	18	**X21**	1,4
A6	71	**D1**	1,42	**X10**	1,18	**X22**	1,07
A7	47	**D2**	0,67	**X11**	1,17	**X23**	1,67
B1	1,56	**D5**	14,89	**X12**	1,6	**X24**	7,73
B2	1,55	**X1**	-1,92	**X13**	-2,1	**pIC$_{50}$**	1,46
B3	0,64	**X2**	-0,46	**X14**	1,41		
B4	49,2	**X3**	0,6	**X15**	2,5		

A seguir, após a aplicação da transformação Box-Cox, a matriz de dados com apenas as variáveis preditoras foi investigada em relação a presença de *outliers* multivariados. Nesta etapa, foi aplicada a análise de componentes principais (PCA), para redução de dimensionalidade antes de aplicar a distância de Mahalanobis. Os dados foram centrados na média e autoescalados. A figura 5.19 mostra a variância capturada para os componentes principais.

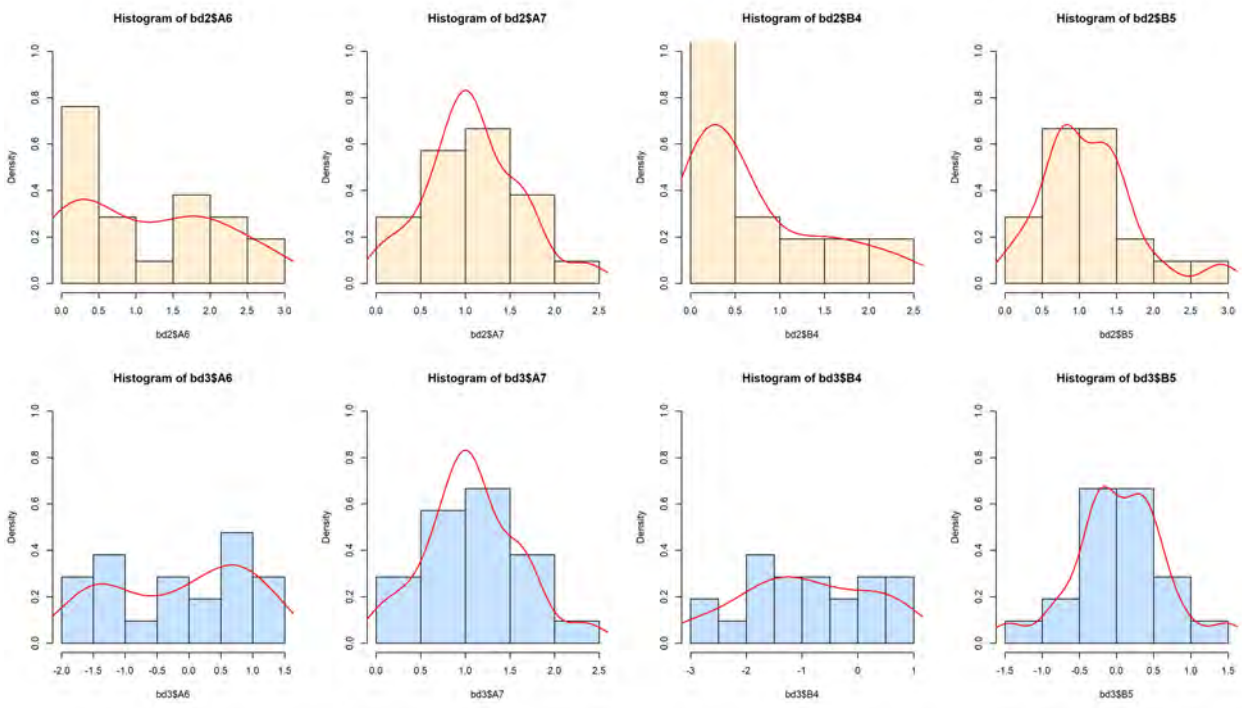


Figura 5.18 – Transformação Box-Cox nas variáveis, antes (bege) e depois (azul)

Fonte: A autora, 2023.

O critério de escolha de retenção do número de componentes principais foi a variância capturada acumulada. A tabela 5.16 mostra, para cada componente principal, o autovalor da matriz de covariância, a variância percentual e a variância percentual acumulada.

A escolha do número de componentes principais pode incluir diferentes critérios, tais como o *scree plot*, o método de Kaiser, o procedimento de Horn, etc [115, pág. 91-93]. Neste trabalho, optou-se pela regra *scree plot*, com a retenção de quatro componentes principais, que correspondiam a quase 67% da variância capturada acumulada. Desta forma, com quatro componentes principais, o valor de corte da distância de Mahalanobis [71] tornou-se:

$$\sqrt{\chi^2_{\left(1-\frac{0,05}{2},4\right)}} = 11,14329$$

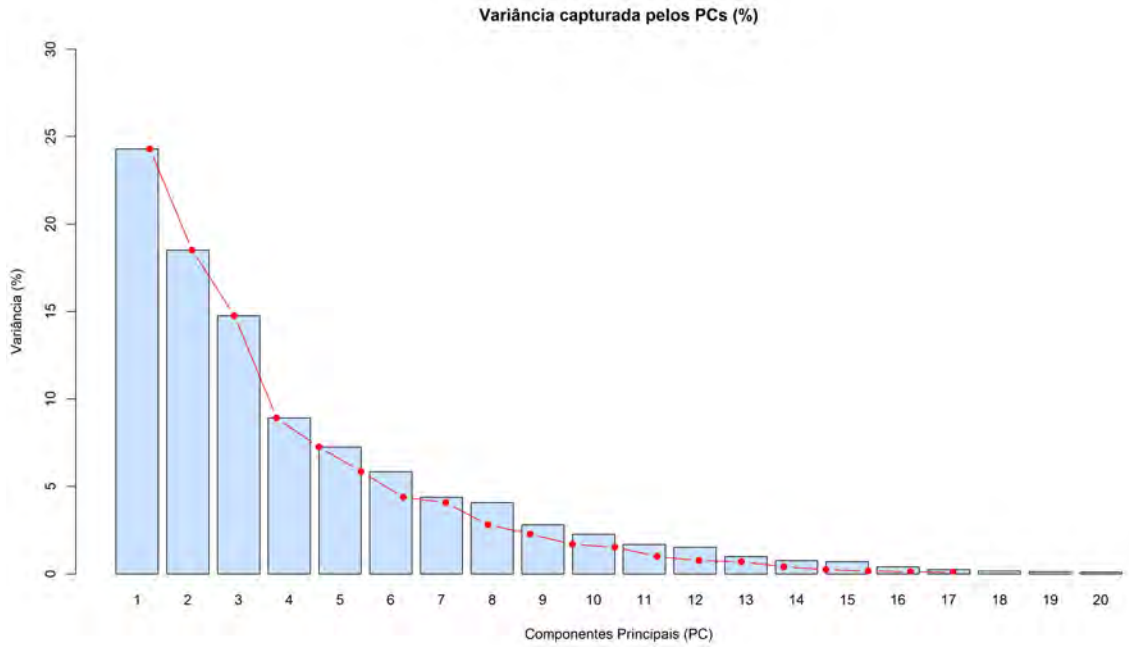


Figura 5.19 – Variância capturada por cada componente principal

Fonte: A autora, 2023.

Baseado nos dados apresentados na tabela 5.17, observa-se que a amostra 18 excede o valor de corte, e pode ser considera um outlier multivariado. Ela será então removida do conjunto.

O modelo de regressão por mínimos quadrados parciais, PLS, foi construído com a matriz de $\mathbf{X}$ com dados centrados na média e autoescalados, além de ter dimensão 20×42. Este modelo foi empregado para contornar o problema da multicolinearidade presente. Desta forma, foram utilizadas 41 variáveis independentes na modelagem. As amostras alocadas nos conjuntos de calibração (treino) e validação (teste) foram, respectivamente, 16 e 4.

A etapa de validação cruzada *leave-one-out* mostrou que apenas uma variável latente (VL) era necessária.

A tabela 5.18 mostra as figuras de mérito para as etapas de calibração, de validação cruzada e de validação, para o modelo construído.

Tabela 5.16 – Análise de componentes principais.

PC	Autovalor de cov(X)	Variância capturada (%)	Variância capturada total (%)
1	9,9621	24,2978	24,2978
2	7,5895	18,5110	42,8088
3	6,0552	14,7689	57,5777
4	3,6594	8,9255	66,5031
5	2,9773	7,2616	73,7648
6	2,4011	5,8564	79,6211
7	1,7995	4,3890	84,0102
8	1,6769	4,0900	88,1002
9	1,1552	2,8176	90,9178
10	0,9355	2,2817	93,1995
11	0,6947	1,6945	94,8940
12	0,6319	1,5411	96,4351
13	0,4109	1,0022	97,4373
14	0,3183	0,7762	98,2135
15	0,2893	0,7057	98,9192
16	0,1710	0,4172	99,3364
17	0,1050	0,2561	99,5925
18	0,0717	0,1750	99,7674
19	0,0534	0,1303	99,8977
20	0,0419	0,1023	100,0000

Tabela 5.17 – Distância de Mahalanobis.

Amostra	Valor	Amostra	Valor	Amostra	Valor
1	7,93	8	2,54	15	0,55
2	1,64	9	4,93	16	5,02
3	1,52	10	0,14	17	2,00
4	1,54	11	2,99	18	11,53
5	3,09	12	0,55	19	4,62
6	2,30	13	9,74	20	3,20
7	6,82	14	1,46	21	5,90

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r_p^2 = 0,1417$ e o coeficiente de determinação permutado médio corrigido, $cr_p^2 = 0,7381$. De acordo com [76], os modelos com $cr_p^2 > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso. Além das métricas avaliadas, em específico da área de QSAR, vale ressaltar que o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r_c^2 = 0,5399 > Q^2 = 0,0461$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,4204 < RMSEP = 0,6345.

Por fim, a diferença entre o r_c^2 e o Q^2 ($r_c^2 - Q^2 = 0,4938$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobreajuste [111]. O modelo não passou neste critério e também nos critérios $Q^2 > 0,5$ e $r_{\text{cal}}^2 > 0,6$.

A figura 5.20 mostra os dados medidos e previstos, de pIC_{50}, dos conjuntos de treino (círculo azul) e teste (círculo vermelho).

Como o modelo não passou em dois critérios, foi investigado se o uso de seleção de

Tabela 5.18 – Figuras de mérito para o modelo PLS

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,5399	r^2	0,1272	r^2	0,1132
RMSE	0,4204	RMSE	0,6053	RMSE	0,6345
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
MAE	0,3039	MAE	0,4582	MAE	0,5457
Bias	3,97E-16	Bias	-0,0346	Bias	0,1582
- - -	- - -	Q^2	0,0461	- - -	- - -
- - -	- - -	PRESS	5,8627	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	-0,0931	- - -	- - -
- - -	- - -	r^2_p	0,1417	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,7381	- - -	- - -

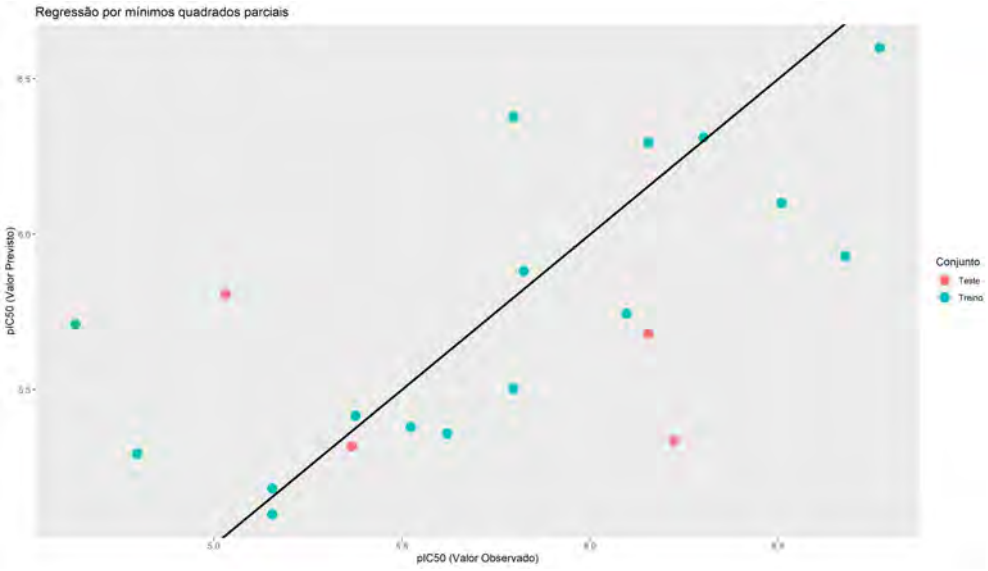


Figura 5.20 – Treino e teste, PLS

Fonte: A autora, 2023.

variáveis melhoraria a modelagem com o PLS. Dois métodos foram utilizados acoplados ao método PLS: a importância da variável na predição (VIP-PLS) e eliminação de variável não informativa (UVE-PLS).

5.3.2.1 VIP-PLS

Como critério de corte, todas as variáveis cuja importância fossem maior que 1, foram selecionadas. A figura 5.21 mostra as variáveis selecionadas, em azul. Os valores de cada projeção são mostradas na tabela 5.19

De posse dessas variáveis, foi construído um novo modelo PLS. A validação cruzada indicou que apenas uma variável latente era necessária. A figura 5.22 mostra os valores preditos e observados de pIC_{50}, para os conjuntos treino (círculo azul) e teste (círculo vermelho).

A tabela 5.20 mostra as figuras de mérito desse modelo construído.

Tabela 5.19 – Valores VIP

Var.	VIP	Var.	VIP	Var.	VIP	Var.	VIP
A1	1,5070	B5	0,6779	X4	0,2690	X17	0,1479
A2	1,6270	C1	1,0405	X5	0,3600	X18	0,1123
A3	1,8897	C3	0,5442	X7	1,4813	X19	1,0477
A4	0,2632	C4	1,0099	X8	0,1649	X20	0,4317
A5	0,2074	C5	0,5369	X9	1,2565	X21	1,3113
A6	1,7180	D1	0,2885	X10	1,2250	X22	1,7246
A7	0,2266	D2	0,2920	X11	1,7215	X23	1,5831
B1	0,5663	D5	0,0515	X12	0,4386	X24	0,2212
B2	0,4626	X1	1,5981	X13	0,2388	- - -	- - -
B3	0,4625	X2	1,5774	X14	0,6059	- - -	- - -
B4	1,0212	X3	0,6085	X15	1,1062	- - -	- - -

Tabela 5.20 – Figuras de mérito para o modelo VIP-PLS

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,5669	r^2	0,3462	r^2	0,0551
RMSE	0,4079	RMSE	0,5131	RMSE	0,6568
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
MAE	0,3124	MAE	0,3897	MAE	0,5627
Bias	3,75E-16	Bias	-0,0202	Bias	0,2244
- - -	- - -	Q^2	0,3148	- - -	- - -
- - -	- - -	PRESS	4,2117	- - -	- - -
- - -	- - -	s_{PRESS}	-0,6841	- - -	- - -
- - -	- - -	r^2_p	0,1652	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,4772	- - -	- - -

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $R^2_p = 0,1652$ e o coeficiente de determinação permutado médio corrigido, $cR^2_p = 0,4772$. De acordo com [76], os modelos com $cR^2_p > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido pode ter sido obtido meramente ao acaso, não passando, portanto, neste critério. Além das métricas avaliadas, em específico da área de QSAR, vale ressaltar que o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_c = 0,5669 > Q^2 = 0,3148$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,4079 < RMSEP = 0,06568. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,1649$) foi menor que o intervalo de 0,2-0,3, o que demonstra que o modelo não tem sobreajuste [111].

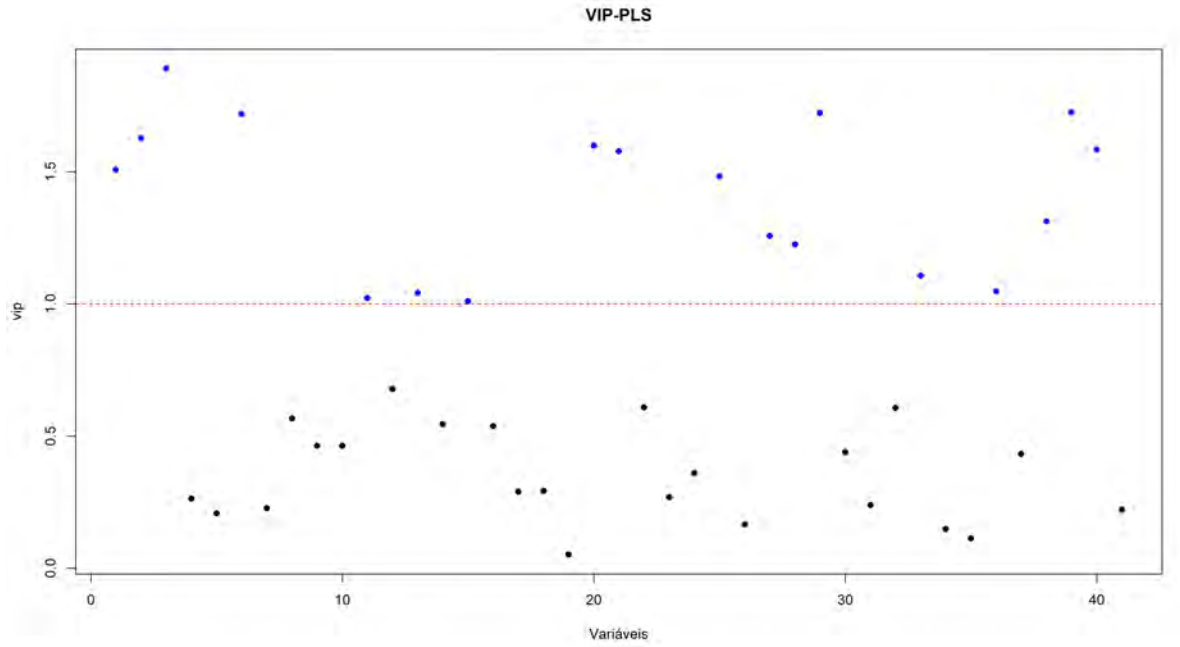


Figura 5.21 – Variáveis selecionadas (azul), VIP-PLS

Fonte: A autora, 2023.

O modelo não passou em três critérios $\left(Q^2 > 0,5;\ r_c^2 > 0,6 \text{ e } cr_p^2 > 0,5\right)$. Foi então investigado o uso de outro método de seleção de variáveis acoplado ao método PLS.

5.3.2.2 UVE-PLS

As variáveis preditoras selecionadas foram: **B2**, **B3**, **X2**, **X7**, **X9**, **X11**, **X21**, **X22**, **X23** e **X24**, com uma dimensão 16×11 no subconjunto treino. Através do procedimento de validação cruzada *leave-one-out*, foi determinado que o parâmetro ótimo de variáveis latentes era 1. A tabela 5.21 mostra as figuras de mérito obtidas. A figura 5.23 mostra os valores preditos e observados, para o pIC_{50}, dos conjuntos treino (círculo azul) e teste (círculo vermelho).

Tabela 5.21 – Figuras de mérito para o modelo UVE-PLS

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,3032	r^2	0,1189	r^2	0,2005
RMSE	0,5174	RMSE	0,5962	RMSE	0,5824
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
MAE	0,3908	MAE	0,4710	MAE	0,5180
Bias	3,87E-16	Bias	9,19E-4	Bias	0,1761
- - -		Q^2	0,0745	- - -	- - -
- - -	- - -	PRESS	5,6882	- - -	
- - -	- - -	$s_{\text{PRESS}_{cv}}$	0,4770	- - -	
- - -	- - -	r^2_p	0,1408	- - -	
- - -	- - -	${}^c r^2_p$	0,2219	- - -	

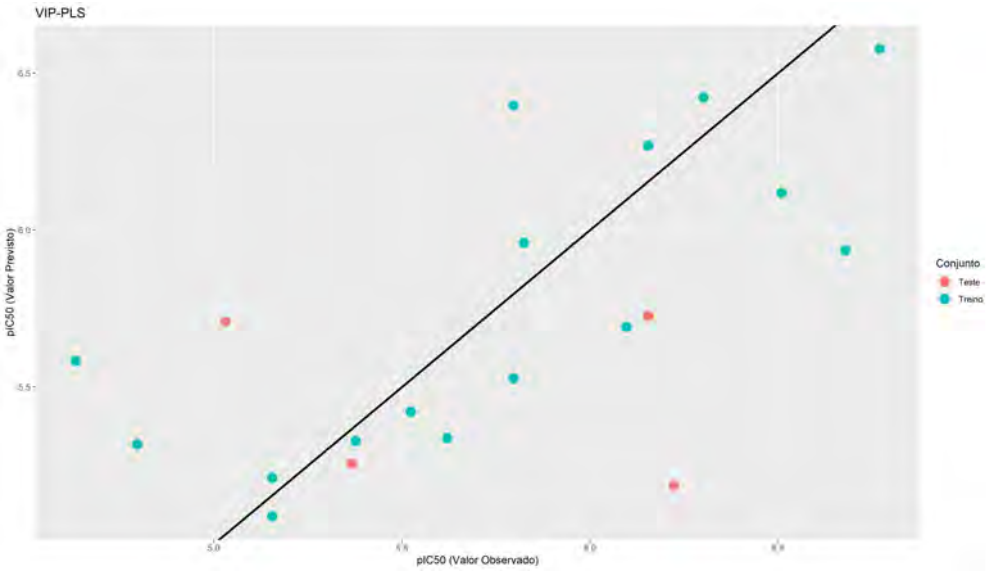


Figura 5.22 – Treino e teste, VIP-PLS

Fonte: A autora, 2023.

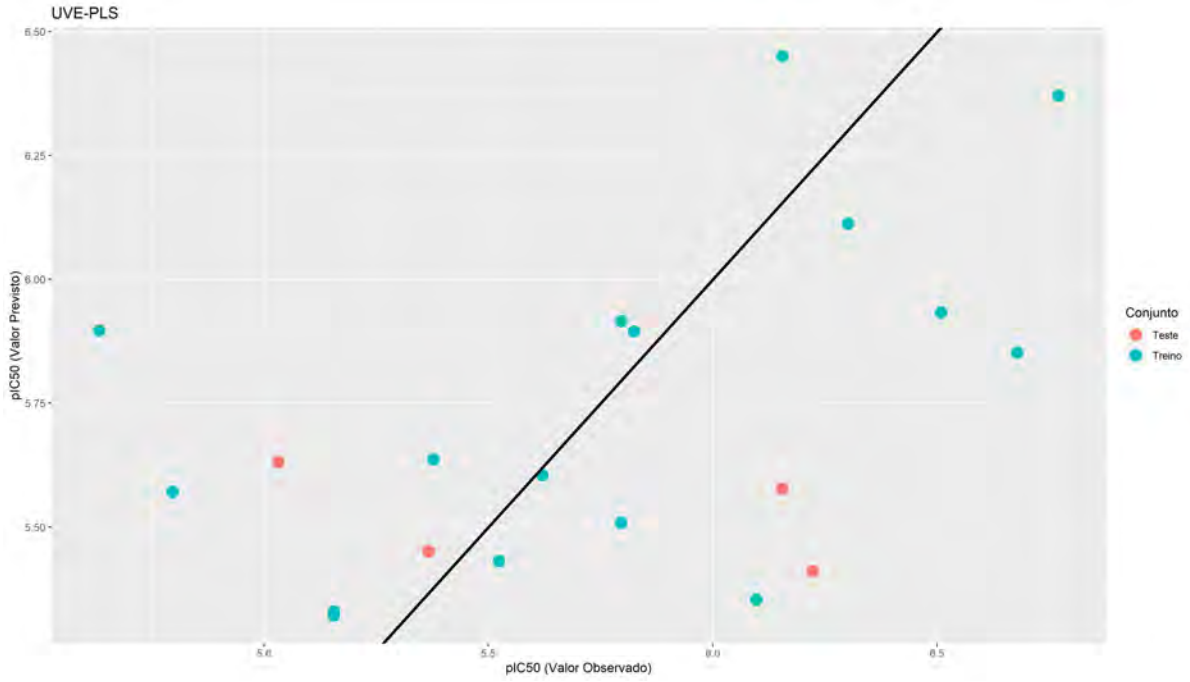


Figura 5.23 – Treino e teste, UVE-PLS

Fonte: A autora, 2023.

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r_p^2 = 0,1408$ e o coeficiente de determinação permutado médio corrigido, $cr_p^2 = 0,2219$. De acordo com [76], os modelos com $cr_p^2 > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido pode ter sido obtido meramente ao acaso. Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r_c^2 = 0,3032 > Q^2 = 0,0745$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,5174 < RMSEP = 0,5824. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,2287$) foi menor que o intervalo de 0,2-0,3, o que demonstra que o modelo não tem sobreajuste [111]. O presente modelo não passou em dois critérios da área de QSAR $\left(Q^2 > 0,5 \text{ e } r_c^2 > 0,6\right)$. Assim, serão investigados métodos de regularização.

5.3.3 Regressão Ridge

Na construção do modelo de regressão ridge, o valor de lambda foi selecionado através de validação cruzada ($\lambda = 0,4545$). Foi aplicada a padronização por escore-z nas amostras. A tabela 5.22 mostra as figuras de mérito do modelo obtido, enquanto a figura 5.24 mostra os valores preditos e observados de pIC_{50} dos conjuntos treino (círculo azul) e teste (círculo vermelho).

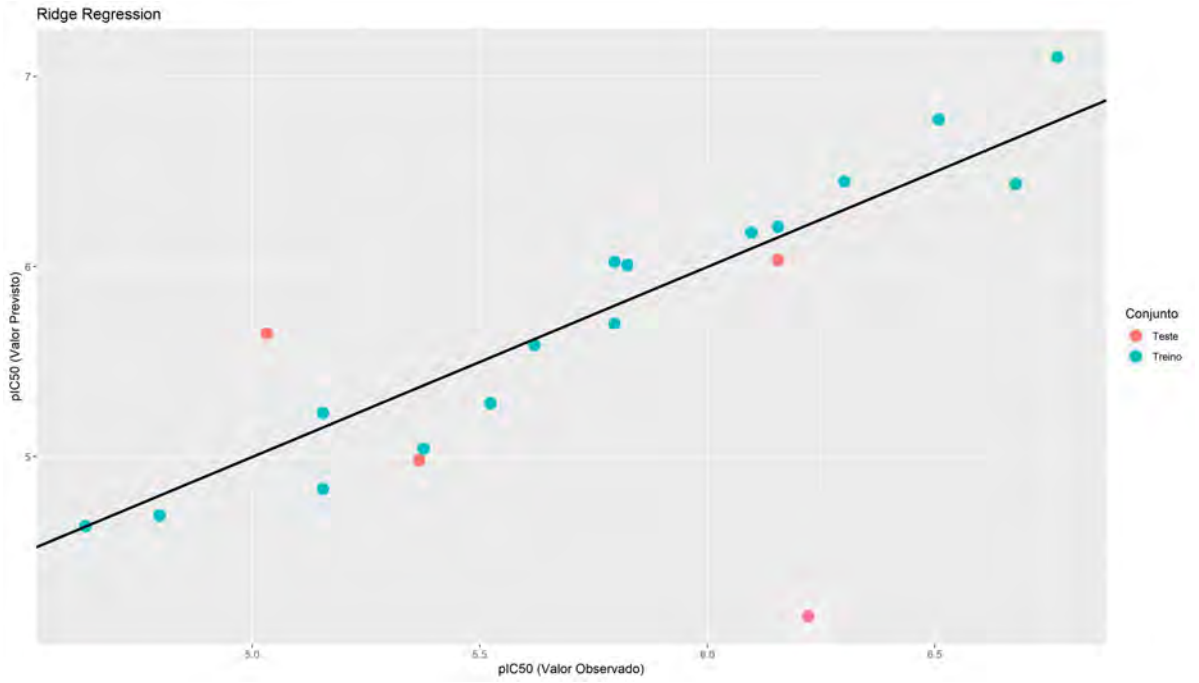


Figura 5.24 – Treino e teste, Ridge Regression

Fonte: A autora, 2023.

Tabela 5.22 – Figuras de mérito para o segundo modelo RR

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,9383	r^2	0,1310	r^2	0,0729
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
RMSE	0,2034	RMSE	0,8028	RMSE	1,0951
MAE	0,1720	MAE	0,6310	MAE	0,7963
Bias	2,22E-16	Bias	-0,1753	Bias	0,4870
- - -	- - -	Q^2	-0,6776	- - -	- - -
- - -	- - -	PRESS_{cv}	10,3112	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	-0,1235	- - -	- - -
- - -	- - -	r^2_p	0,1163	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,8783	- - -	- - -

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r^2_p = 0,1163$ e o coeficiente de determinação permutado médio corrigido, $cr^2_p = 0,8783$. De acordo com [76], os modelos com $cr^2_p > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso. Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_c = 0,9383 > Q^2 = -0,6776$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,2034 < RMSEP = 1,0951. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 1,6160$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobreajuste [111]. O modelo proposto não passou em um requisito específico da área de QSAR $\left(Q^2 > 0,5\right)$. Desta forma, foi investigado outro método de regularização (no caso, a regressão Rede Elástica).

5.3.4 Regressão Rede Elástica

A rede elástica é uma técnica de regularização que combina a penalidade de L_1 (utilizada no modelo LASSO, *least absolute shrinkage and selection operator*) e a penalidade de L_2 (utilizada na construção do modelo de regressão Ridge) para realizar seleção de variáveis e redução da dimensionalidade de maneira mais eficaz. Com o pacote *enet*, dois argumentos foram otimizados através da validação cruzada: *fraction*, que controla a intensidade relativa dessas duas penalidades, e *lambda*, que controla a intensidade da penalidade aplicada ao ajustar o modelo. Neste modelo, fraction = 0,3989796 e lambda = 0,0. A figura 5.25 mostra os valores observados e preditos de pIC_{50}, nos conjuntos treino (círculo azul) e teste (círculo vermelho).

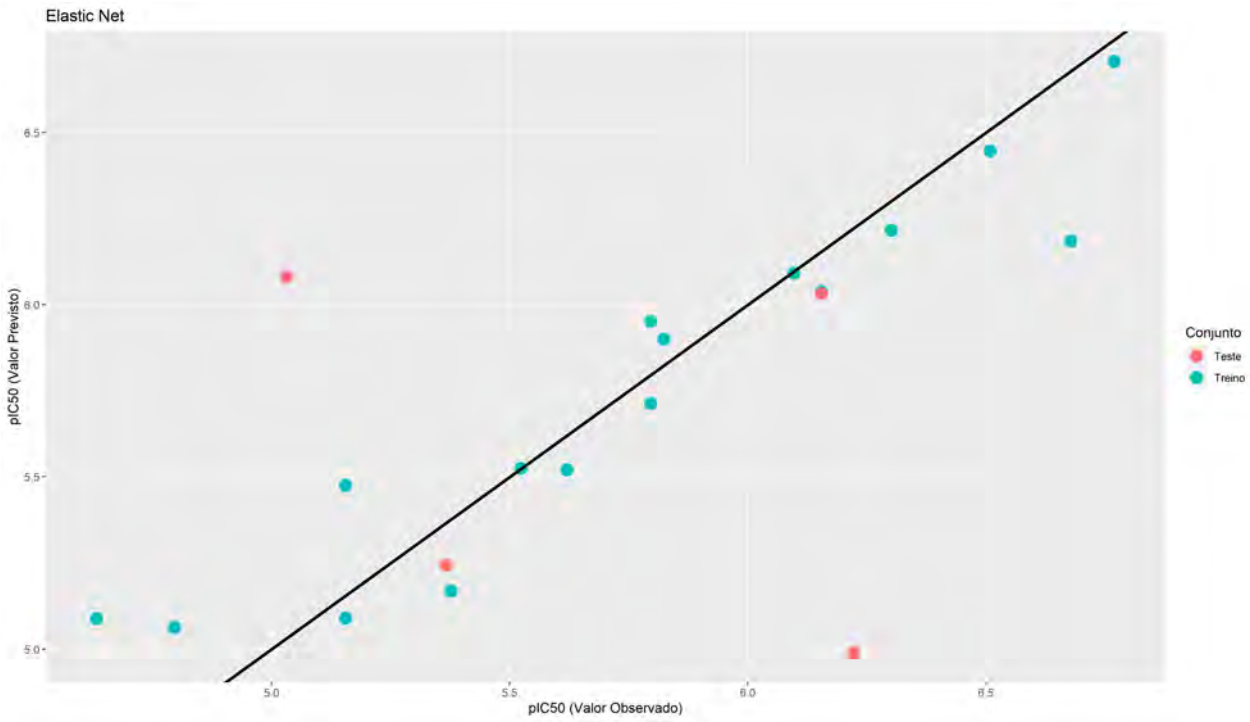


Figura 5.25 – Treino e teste, regressão Rede Elástica

Fonte: A autora, 2023.

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r_p^2 = 0,1247$ e o coeficiente de determinação permutado médio corrigido, $cr_p^2 = 0,8357$. De acordo com [76], os modelos com $cr_p^2 > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso. Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r_c^2 = 0,9004 > Q^2 = 0,2733$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,2163 < RMSEP = 0,8138. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,6271$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobreajuste [111].

Desta forma, serão investigados modelos não lineares.

Tabela 5.23 – Figuras de mérito para o segundo modelo Regressão Rede Elástica

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,9004	r^2	0,3001	r^2	0,1110
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
RMSE	0,2163	RMSE	0,5283	RMSE	0,8138
MAE	0,1598	MAE	0,4236	MAE	0,6317
Bias	5,00E-16	Bias	-0,0480	Bias	0,1069
- - -	- - -	Q^2	0,2733	- - -	- - -
- - -	- - -	PRESS_{cv}	4,4662	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	-0,0813	- - -	- - -
- - -	- - -	r^2_p	0,1247	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,8357	- - -	- - -

5.3.5 Regressão por máquina de vetor suporte

5.3.5.1 kernel RBF

No modelo construído número de vetores de suporte foi 15; o valor da função objetivo foi -10,2195; o erro de treinamento foi 0,153033; a função custo foi C = 2,11; e os parâmetros sigma = 0,01 e epsilon = 0,1. Chama a atenção o número de vetores de suporte utilizado, pois é quase o mesmo que o tamanho do conjunto treino, o que sugere um sibreajuste. A tabela 5.24 mostra as figuras de mérito do modelo obtido. A figura 5.26 mostra os valores observados e preditos de pIC_{50}, dos conjuntos treino (círculo azul) e teste (círculo vermelho).

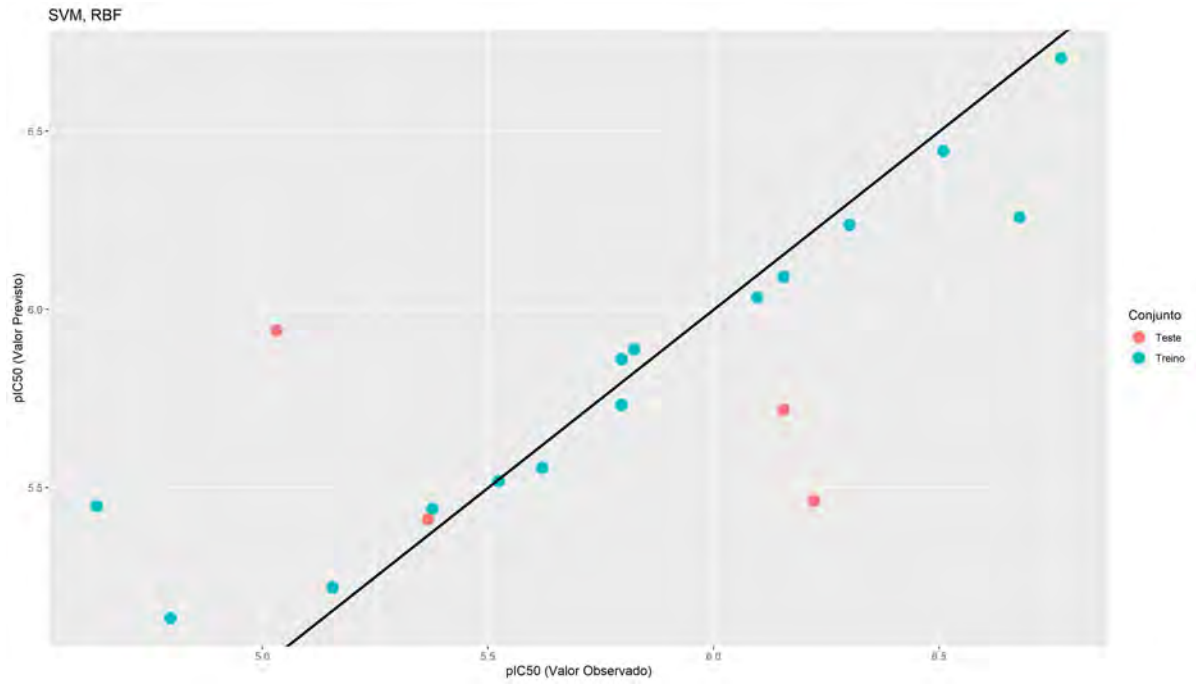


Figura 5.26 – Treino e teste, SVR, kernel RBF

Fonte: A autora, 2023.

Tabela 5.24 – Figuras de mérito para o segundo modelo SVR, com kernel RBF

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,8832	r^2	0,3495	r^2	0,1771
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
RMSE	0,2504	RMSE	0,5142	RMSE	0,6318
MAE	0,1465	MAE	0,3922	MAE	0,5375
Bias	-0,0376	Bias	-0,0615	Bias	0,0601
- - -	- - -	Q^2	0,3118	- - -	- - -
- - -	- - -	PRESS_{cv}	4,2297	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	-0,0791	- - -	- - -
- - -	- - -	r^2_p	0,1364	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,8121	- - -	- - -

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r^2_p = 0,1364$ e o coeficiente de determinação permutado médio corrigido, $cr^2_p = 0,8121$. De acordo com [76], os modelos com $cr^2_p > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso. Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_c = 0,8832 > Q^2 = 0,3118$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,2504 < RMSEP = 0,6318. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,5713$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobreajuste [111]. Além disso, o modelo também reprovou em um dos critérios QSAR $\left(Q^2 > 0,5\right)$. Outro kernel foi investigado.

5.3.5.2 kernel Linear

No modelo construído o número de vetores de suporte foi 14; o valor da função objetivo foi -0,1304; o erro de treinamento foi 0,289517; a função custo foi C = 0,02; e o parâmetro epsilon = 0,1. Chama a atenção o número de vetores de suporte utilizado, pois é quase o mesmo que o tamanho do conjunto treino, o que sugere um sobreajuste. A tabela 5.25 mostra as figuras de mérito do modelo obtido. A figura 5.27 mostra os valores observados e preditos de pIC_{50}, dos conjuntos treino (círculo azul) e teste (círculo vermelho).

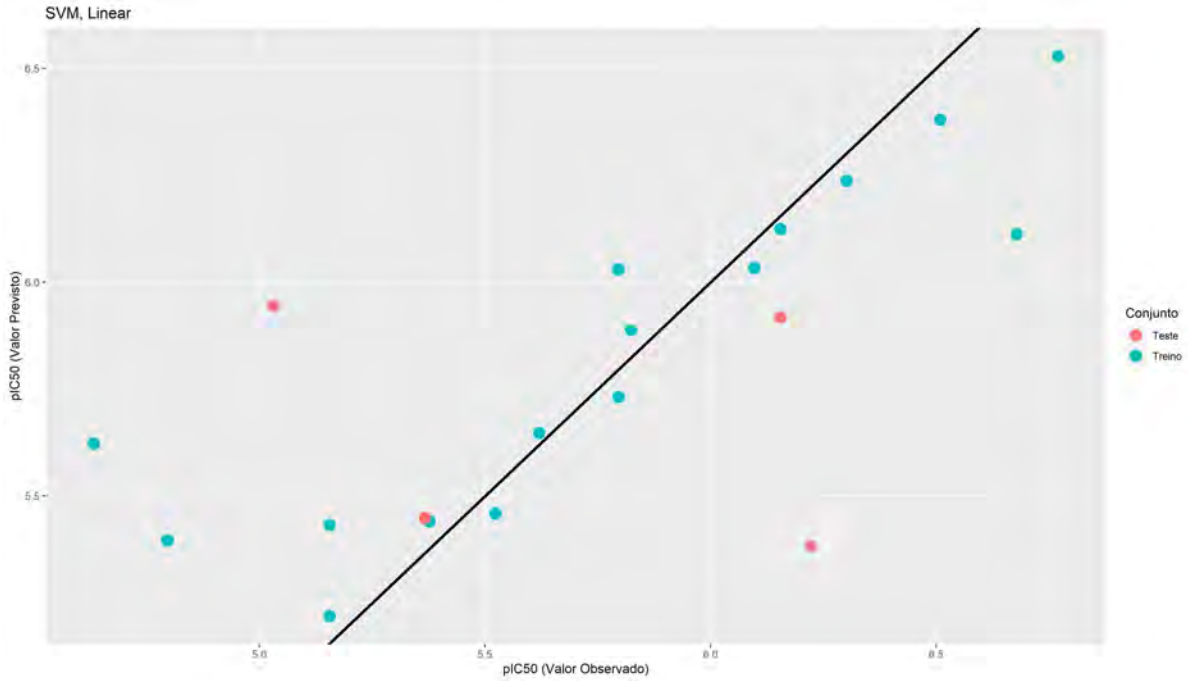


Figura 5.27 – Treino e teste, SVR, kernel linear

Fonte: A autora, 2023.

Tabela 5.25 – Figuras de mérito para o segundo modelo SVR, com kernel Linear

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,7757	r^2	0,1695	r^2	0,0781
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
RMSE	0,3444	RMSE	0,5763	RMSE	0,6329
MAE	0,2215	MAE	0,4682	MAE	0,5181
Bias	-0,0687	Bias	-0,0867	Bias	0,0203
- - -	- - -	Q^2	0,1353	- - -	- - -
- - -	- - -	PRESS_{cv}	5,3147	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	-0,0887	- - -	- - -
- - -	- - -	r^2_p	0,1333	- - -	- - -
- - -	- - -	${}^c r^2_p$	0,7059	- - -	- - -

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r^2_p = 0,1333$ e o coeficiente de determinação permutado médio corrigido, $cr^2_p = 0,7059$. De acordo com [76], os modelos com $cr^2_p > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso. Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_c = 0,7022 > Q^2 = 0,1353$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,3444 < RMSEP = 0,6329. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,6404$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobreajuste [111]. Além de não passar neste critério da área QSAR, também não passou no critério $Q^2 > 0,5$. Desta forma, outro método não linear será investigado.

5.3.6 Regressão por floresta aleatória

5.3.6.1 Sem seleção de variáveis

Na construção do modelo de regressão de floresta aleatória, os dados foram padronizados por escore-z. Foi utilizada a base de dados sem remoção de variáveis com colinearidade. Através do procedimento de validação cruzada *leave-one-out*, o parâmetro *mtry*, que define o número de parâmetros em cada divisão, foi mantido em 6. A tabela 5.26 mostra as figuras de mérito deste modelo preditivo, enquanto que a figura 5.28 mostra os valores preditos e observados de pIC_{50}, nos conjuntos treino (círculo azul) e teste (círculo vermelho).

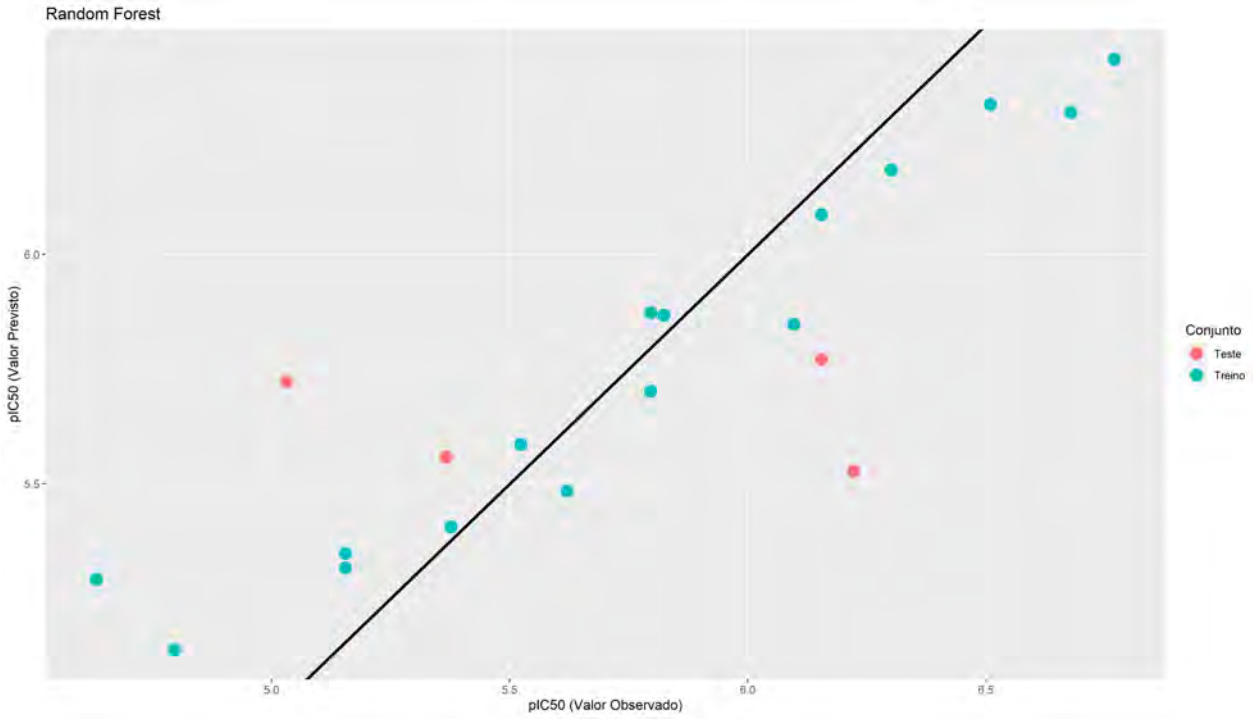


Figura 5.28 – Treino e teste, modelo de regressão random forest

Fonte: A autora, 2023.

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r_p^2 = 0,1650$ e o coeficiente de determinação permutado médio corrigido, $cr_p^2 = 0,8518$. De acordo com [76], os modelos com $cr_p^2 > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso.

Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r_c^2 = 0,9343 > Q^2 = 0,0994$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC = 0,2528 < RMSEP = 0,5346. Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,8349$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobreajuste [111]. Além de não passar neste critério da área QSAR, também não passou no critério $Q^2 > 0,5$.

Desta forma, será investigado o uso de seleção de variáveis na construção de um modelo de regressão random forest.

Tabela 5.26 – Figuras de mérito para o modelo RF

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,9343	r^2	0,0996	r^2	0,0157
RMSE	0,2528	RMSE	0,5882	RMSE	0,5346
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
MAE	0,1955	MAE	0,4730	MAE	0,4900
Bias	-0,0104	Bias	-0,0021	Bias	0,0490
- - -	- - -	Q^2	0,0994	- - -	- - -
- - -	- - -	PRESS_{cv}	5,5352	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	-0,0905	- - -	- - -
- - -	- - -	r^2_p	0,1650	0,1636	- - -
- - -	- - -	${}^c r^2_p$	0,8518	0,8485	- - -

5.3.6.2 Com seleção de variáveis

A partir do modelo gerado anterior, foi extraída a importância de cada variável. A tabela 5.27 mostra a importância de vinte variáveis, em ordem decrescente.

Tabela 5.27 – Contribução das variáveis, Random Forest.

Var.	Importância	Var.	Importância
X21	100	X11	40,97
A2	86,91	C4	36,87
X9	83,39	C5	36,2
A1	74,09	X20	28,97
A3	72,04	D5	28,21
X1	68,55	X22	27,68
A6	66,87	X15	25,03
X2	51,12	A5	25,01
B4	45,02	B1	24,37
X7	42,82	B3	22,14

Na construção do modelo, foram utilizadas as variáveis **A2**, **X9** e **X21**. O parâmetro mtry = 1 foi escolhido através de validação cruzada. A tabela 5.28 mostra as figuras de mérito para o modelo construído. A figura 5.29 mostra os valores preditos e observados, de pIC_{50}, dos conjuntos treino (círculo azul) e teste (círculo vermelho).

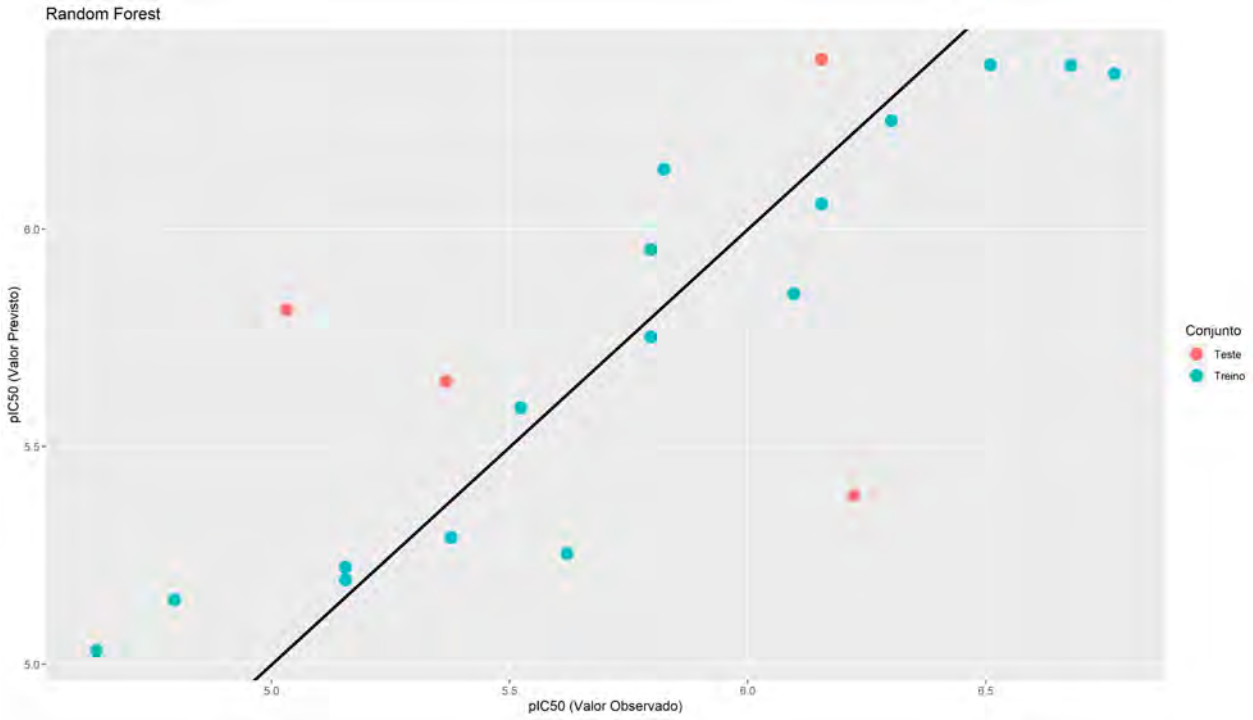


Figura 5.29 – Treino e teste, modelo de regressão

Fonte: A autora, 2023.

Tabela 5.28 – Figuras de mérito para o modelo RF, com seleção de variáveis.

Calibração		Validação Cruzada		Validação	
Métrica	Resultado	Métrica	Resultado	Métrica	Resultado
r^2	0,8815	r^2	0,5569	r^2	0,0159
RMSE	0,2386	RMSE	0,4190	RMSE	0,6008
r^2_{aj}	- - -	- - -	- - -	- - -	- - -
MAE	0,1957	MAE	0,3599	MAE	0,5338
Bias	0,0212	Bias	0,0137	Bias	-0,1169
- - -	- - -	Q^2	0,5429	- - -	- - -
- - -	- - -	PRESS_{cv}	2,8096	- - -	- - -
- - -	- - -	$s_{\text{PRESS}_{cv}}$	0,1397	- - -	- - -
- - -	- - -	r^2_p	0,1528	0,1470	- - -
- - -	- - -	${}^c r^2_p$	0,7190	0,8047	- - -

Após a aleatorização em Y, foram obtidos o coeficiente de determinação permutado médio, com valor $r^2_p = 0,1528$ e o coeficiente de determinação permutado médio corrigido, $cr^2_p = 0,7190$. De acordo com [76], os modelos com $cr^2_p > 0,5$ são considerados estatisticamente robustos, indicando que o modelo desenvolvido não foi obtido meramente ao acaso.

Além das métricas avaliadas, em específico da área de QSAR, o coeficiente de determinação, da etapa de calibração, é maior que o coeficiente de determinação de validação cruzada $r^2_c = 0,8815 > Q^2 = 0,5429$; bem como a raiz quadrada do erro médio de calibração é menor que a da etapa de validação, RMSEC $= 0,2386 <$ RMSEP $= 0,6008$.

Por fim, a diferença entre o r^2 e o Q^2 ($r^2 - Q^2 = 0,3386$) foi maior que o intervalo de 0,2-0,3, o que demonstra que o modelo tem sobre ajuste [111].

6 Considerações

A partir de alguns descritores obtidos dos programas GOLD e Spartan, foi aplicada uma metodologia simples, porém eficaz, de modelagem preditiva em aprendizado de máquina, com o auxílio de metodologias econométricas e quimiométricas.

Apenas quarenta e sete descritores foram disponibilizados para estudo. A literatura indica que estes podem chegar à casa dos milhares, o que permitiria, talvez, a melhoria do modelo obtido, com variáveis preditoras mais adequadas, além, claro, do emprego de outros métodos para seleção de variáveis (tais como algoritmo genético) e/ou redução de dimensionalidade (como o escalonamento multidimensional).

Diferentes métodos de aprendizado de máquina foram empregados, na predição de atividade biológica pIC_{50} da enzima N-miristoiltransferase, seguindo o protocolo de construção de modelos preditivos propostos na literatura, na estratégia de avaliar somente moléculas com funções químicas semelhantes.

Entre os métodos avaliados (MLR, BS-MLR, PLS, VIP-PLS, UVE-PLS, RR, Enet, SVR com *kernels* RBF e Linear, e RF), quem apresentou o melhor desempenho, atendendo a diferentes critérios, tanto estatísticos quanto específicos da área de QSAR e com menor RMSE, foi a regressão nos melhores subconjuntos (BS-MLR). Todos os demais modelos reprovaram em uma ou duas métricas da área de QSAR.

Quanto ao campo da estatística, o método que teve melhor resultado em validação cruzada foi BS-MLR, já na calibração o método que teve o menor RMSE foi regressão Ridge.

A despeito dos baixos valores dos coeficientes de determinação do conjunto teste, cuja representação gráfica pode ser vista através dos gráficos de valores preditos versus observados, ficou claro que o problema estava no conjunto teste, com resíduos elevados sendo apresentados.

O presente trabalho deve ser expandido para uma base de dados com mais amostras de grupo funcional semelhante, variando tanto as posições quanto os tipos de substituintes. Infelizmente, isto foge ao escopo desta dissertação, pois envolveria as etapas de aquisição de reagentes e meios reacionais; síntese orgânica; purificação através de métodos cromatográficos; caracterização através de espectroscopias de infravermelho médio e de ressonância magnética nuclear de 1H, além da espectrometria de massas; e por fim, avaliação biológica em ensaios de toxicidade. A literatura apontou programas que geram de dezenas até milhares de descritores químicos, o que pode influenciar os resultados.

7

Referências bibliográficas

1 SANTOS, S. S.; ARAÚJO, R. V.; GIAROLLA, J.; SEOUD, O. E.; Ferreira, E. I. *Searching for drugs for Chagas disease, leishmaniasis and schistosomiasis*, *International Journal of Antimicrobial Agents*, V. 55, N. 4, p. 2 e 18, 2020.

2 THAKUR, S.; JOSHI, J.; KAUR, S. *Leishmaniasis diagnosis: an update on the use of parasitological, immunological and molecular methods*. Journal of Parasitic Diseases, v. 44, p. 253–272, 2020.

3 GARCIA, L.S. Estudos computacionais de potenciais inibidores da enzima N-Miristoiltransferase de *Plasmodium falciparum* e *Leishmania donovani*. 2017. 106 f. Tese (Doutorado em Agroquímica). Universidade Federal de Lavras, Minas Gerais.

4 ANDRADE, C. H.; PASQUALOTO, K. F. M.; FERREIRA, E. I.; HOPFINGER, A. J. *4D-QSAR: Perspectives in Drug Design in Drug Design*. Molecules, v. 15, n. 5, p. 3281-3294, 2010.

5 MARTINS, J. P. A.; FERREIRA, M. M. C. *QSAR modeling: um novo pacote computacional open source para gerar e validar modelos QSAR*. Química Nova, V. 36, N. 4, p. 554 - 560, 2013.

6 WHO - *World Health Organization*;.*Leishmaniasis* Disponível em:<https://www.who.int/news-room/fact-sheets/detail/leishmaniasis> . Acesso em 25 de fev, 2024.

7 GARCIA, L. S.; SILVA, D. R.; ASSIS, L. A.; ASSIS, T. M.; GAJO, G. C.; FERNANDES, I. A.; RAMALHO, T. C.; CUNHA, E. F. F.*Design of Novel N-Myristoyltransferase Inhibitors of Leishmania donovani Using Four-Dimensional Quantitative Structure-Activity Relationship Analysis*. Journal of the Brazilian Chemical Society, V. 29, N. 7, p. 1440 - 1454, 2018.

8 GIL, Z.; MARTINEZ-SOTILLO, N.; PINTO-MARTINEZ, A.; MEJIAS, F.; Martinez, J. C.; GALINDO, I.; Oldfield, E.; BENAIM, G. SQ109 inhibits proliferation of Leishmania donovani by disruption of intracellular Ca^{2+} homeostasis, collapsing the mitochondrial electrochemical potential ($\Delta\Psi_m$) and affecting acidocalcisomes. *Parasitology Research*, V. 119, N. 2, p. 649 - 657, 2020.

9 MCCALL, L. I., ZHANG, W. W., MATLASHEWSKI, G. *Determinants for the development of visceral leishmaniasis disease*. PLoS Pathogens, V. 9, N. 1, p. e1003053, 2013

10 SINGH, P.; KUMAR, A. *Deciphering the function of unknown Leishmania donovani cytosolic proteins using hyperparameter-tuned random forest. Network Modeling Analysis in Health Informatics and Bioinformatics*, V. 9, N. 2, p. 1-9, 2020.

11 WALL, R. J.; CARVALHO, S.; MILNE, R.; BUEREN-CALABUIG, J. A.; MONIZ, S.; CANTIZANI-PEREZ, J.; MACLEAN, L.; KESSLER, A.; COTILLO, I.; SASTRY, L.; MANTHRI, S.; PATTERSON, S.; ZUCCOTTO, F.; THOMPSON, S.; MARTÍN, J. J.;

MARCO, M.; MILES, T. J.; RYCKER, M.; Thomas, M.G. ; FAIRLAMB, A. H.; GILBERT, I. H.; WYLLIE, S. *The Q_i site of cytochrome b is a promiscuous drug target in Trypanosoma cruzi and Leishmania donovani*. ACS Infectious Diseases, V. 6, N. 3, p. 515 - 528, 2020.

12 WHO. World Health Organization. *Neglected tropical diseases*. Disponível em: <https://www.who.int/data/gho/data/themes/topics/gho-ntd-leishmaniasis>. Acesso em: 31 jan, 2022.

13 MENEZES, M. Pesquisa mostra a evolução da leishmania brasileira. FIOCRUZ, 04/02/2021. Disponível em: <https://portal.fiocruz.br/noticia/pesquisa-mostra-evolucao-da-leishmania-brasileira>. Acesso em: 12 fev. 2022.

14 PARÁ Pará está entre os estados que mais registram mortes por leishmaniose no Brasil. **Correio Paraense**, 23/08/2021. Disponível em: <https://correioparaense.com.br/2021/08/23/para-esta-entre-os-estados-que-mais-registram-mortes-por-leishmaniose-no-brasil/>. Acesso em: 23 ago. 2021.

15 SACKS, D. L. *Metacyclogenesis in Leishmania promastigotes*. Exp Parasitol, v. 69, p. 100 - 103, 1989.

16 JACOBSON, R. L.; SCHELEIN, Y.; EISENBERG, C. L.. *The biological function of sand fly and glycosidades*. Med Microbiol Immnunol, v. 190, p. 51 - 55, 2001.

17 PÊSSOA, S. B.; MARTINS, A. V. *Tripanosomatidae* - Gênero *Leishmania*. *In* Parasitologia Médica. Rio de Janeiro: Editora Guanabara Koogan, p.77 - 118, 1977.

18 GHORBANI, M.; FARHOUDI, R. *Leishmaniasis in humans: drug orvaccine therapy?*. Drug Des Devel Ther., V. 12: p. 25 – 40, 2018.

19 BRASIL. *Controle e Combate à Leishmaniose*. Biblioteca Virtual em Saúde. Disponível em: <https://bvsms.saude.gov.br/10-a-17-8-semana-nacional-de-controle-e-combate-a-leishmaniose>. Acesso em: 31 jan. 2022.

20 ELMAHALLAWY, E. K.; AGIL, A. *Treatment of leishmaniasis: a review and assessment of recent research*, Curr Pharm Des, V. 21, N. 17, p. 2259 - 2275, 2015.

21 FREITAS-JUNIOR, L. H.; CHATELAIN, E.; KIM, H. A., SIQUEIRA-NETO, J. L. *Visceral leishmaniasis treatment: what do we have, what do we need and how to deliver it?*. International Journal for Parasitology: Drugs and Drug Resistance, V. 2, p. 11 - 19, 2012.

22 AKBARI, M.; ORYAN, A.; HATAM, G. *Application of nanotechnology in treatment of leishmaniasis: a review*. Acta Tropica, V. 172, p. 86 – 90, 2017.

23 Brereton R.G.. *Chemometrics Data Driven Extraction for Science*.Editora Wiley 2ª edição, 2018.

24 SUNDAR, S.; GUPTA L. B.; RASTOGI, V.; AGRAWAL, G.; MURRAY, H. W.*Short-course, cost-effective treatment with amphotericin B-fat emulsion cures visceral leishmaniasis.* Trans R Soc Trop Med Hyg, V. 94, p. 200 – 204, 2000.

25 SUNDAR, S.; SINGH, A. *Recent developments and future prospects in the treatment of visceral leishmaniasis.* Therapeutic Advances in Infectious Disease, V. 3, N. 3–4, p. 98 – 109, 2016.

26 MUSA, A.; KHALIL, E.; HAILU, A.; OLOBO, J.; BALASEGARAM, M.; OMOLLO, R. et alli. *Sodium stibogluconate (SSG) and paromomycin combination compared to SSG for visceral leishmaniasis in East Africa: a randomised controlled trial.* PLoS Negl Tropical Diseases, V. 6, N. 6, p. e1674, 2012.

27 MARTIN, D. D.; BEAUCHAMP, E.; BERTHIAUME, L. G.; *Post-translational myristoylation: Fat matters in cellular life and death*. Biochimie, V. 93, N. 1, p. 18 - 31, 2011.

28 RODRIGUES, R. P.; MANTOANI, S. P.; ALMEIDA, J. R.; PINSETTA, F. R.; SEMIGHINI, E. P.; SILVA, V. B.; SILVA, C. H. P. *Estratégias de Triagem Virtual no Planejamento de Fármacos*. Revista Virtual Quimíca, V. 4, N. 6, p. 739 - 776, 2012.

29 SANT'ANNA, C. M. R. Métodos de modelagem molecular para estudo e planejamento de compostos bioativos: Uma introdução. Rev. Virtual de Química, V. 1, N. 1, p. 49 - 57, 2009.

30 OLIVEIRA, J.; S., Análise via aprendizado de máquinas do controle de reações químicas. Monografia (Bacharelado em Química). 72 f. Diadema: Universidade Federal de São Paulo, 2023.

31 BARBOSA, L. C. B.; SILVA, G. M.; FORTINI, R. M. Aprendizado de máquina e big data na descoberta de novos medicamentos: um mapeamento sistemático. Brazilian Journal of Development, V. 9, N. 8, p. 24562 - 24581, 2023.

32 MARTINS, J. P. A.; FERREIRA, M. M. C. *QSAR modeling: um novo pacote computacional open source para gerar e validar modelos QSAR*. Química Nova, V. 36, N. 4, p. 554 - 560, 2013.

33 PEREDA, P. C.; ALVES, D. Econometria Aplicada. Rio de Janeiro: Elsevier, pág. 207, 2018.

34 JONES, G.; WILLETT, P.; GLEN, R. C.; LEACH, A. R.; TAYLOR, R. Development and Validation of a Genetic Algorithm for Flexible docking. Journal of Molecular Biology, v. 267, n. 3, p. 727-748, 1997.

35 FERREIRA, R. S.; OLIVA, G.; ANDRICOPULO, A. D. Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala: oportunidades e desafios em PD de fármacos. Química Nova, V. 34, N. 10, p. 1770 - 1778, 2011.

36 BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T. N.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P. E. The Protein Data Bank. Nucleic Acids Research, V. 28, N. 1, p. 235-242, 2000.

37 BRANNIGAN, J.A., SMITH, B.A., YU, Z., HODGKINSON, M.R., LEATHERBARROW, R.J., TATE, E.W., BRZOZOWSKI, A.M., SMITH, D.F., WILKINSON, A.J.. *Structure of N-myristoyltransferase from L. donovani*. *Protein Data Bank* (PDB), 2009. PDB DOI: https://doi.org/10.2210/pdb2wuu/pdb

38 HOPFINGER, A. J.; WANG, S.; TOKARSKI, J. S.; JIN, B.; Albuquerque, M. G.; MADHAV, P. J.; DURAISWAMI, C. *Construction of 3D-QSAR models using the 4D-QSAR analysis formalism*. Journal of the American Chemical Society, V. 119, N. 43, p. 10509 - 10524, 1997.

39 KIRALJ, R.; FERREIRA, M. M. C. *Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application*. Journal of the Brazilian Chemical Society, V. 20, N. 4, p. 770 - 787, 2009.

40 BRUCE, P.; BRUCE, A. Prática para Cientistas de dados - 50 Conceitos Essenciais . Editora Alta Books, 1ª edição, 2019.

41 FERREIRA, M. M. C. *Multivariate QSAR*. Journal of the Brazilian Chemical Society, v. 13, n. 6, p. 742-753, 2002.

42 GAUDIO, A. C.; ZANDONADE, E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. *Química Nova*, v. 24, n. 5, p. 658-671, 2001.

43 LEATHERBARROW, Robin; TATE, Edwards; YU, Zhiyong; RACKHAM, Mark. *Nobel Compounds and their use in therapy*. WO 2013083991 Al. Reino Unido: World Intellectual Property Organization, 2013.

44 ROCHA, S. F. L. S.; OLANDA, C. G.; FOKOUE, H. H.; SANT'ANNA, C. M. R. *Virtual screening techniques in drug discovery: review and recent applications*. Current Topics in Medicinal Chemistry, v. 19, n. 19, p. 1751 - 1767, 2019.

45 ADENIJI, S.E.; UBA, S.; UZAIRU, A. QSAR modeling and molecular docking analysis of some active compounds against Mycobacterium tuberculosis Receptor (Mtb CYP121). Journal of Patogens, V. 2018.

46 Appendix A. A.1.6 pg 330, YOUNG, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. John Wiley Sons, 2001.

47 Cambridge Crystallographic Data Center (CCDC). Software GOLD,criado em 2022. Disponível em: <https://https://www.ccdc.cam.ac.uk/solutions/software/gold/>. Acesso em: 26 jul. 2023.

48 Wavefunction, Software SPARTAN, criado em 2020. Disponível em: <https://https://https://www.wavefun.com/>. Acesso em: 26 jul. 2023.

49 MORGANTINI, F. M.. Análise quimiométrica de seletividade nos sítios ativos de cisteíno proteases da família da papaína para o estudo de substâncias antiparasitárias e antineoplásicas. Instituto de Química de São Carlos da Universidade de São Paulo, 2021.

50 LEACH, A. R..*Molecular Modelling: Principles and Applications*, 2^{nd} ed. Englewood Cliffs: Prentice Hall, 2001.

51 VERLI, H.; BARREIRO, E. J. Um paradigma da química medicinal: a flexibilidade dos ligantes e receptores. Química Nova, v. 28, n. 1, p. 95 - 102, 2005.

52 Software GOLD - Genetic Optmization for Ligand Docagem (2022.3.0) disponibilizado pelo Cambridge Crystallographic Data Centre – CCDC https://www.ccdc.cam.ac.uk

53 ELDRIDGE, M. D.; Murray, C. W.; AUTON, T. R.; PAOLINI, G. V.; MEE, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. Journal of Computer-Aided Molecular Design, v. 11, n. 5, p. 425–445, 1997.

54 MOOJI, W. T. M.; VERDONK, M. L. General and Targeted Statistical Potentials for Protein Ligand Interactions. PROTEINS: Structure, Function, and Bioinformatics, v. 61, n. 2, p. 272 - 287, 2005.

55 KORB, O.; STÜTZLE, T.; EXNER, T. E. Empirical Scoring Functions for Advanced Protein Ligand Docking with PLANTS. Journal of chemical information and modeling, v. 49, n. 1, p. 84–96, 2009.

56 R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

57 Posit team. RStudio: Integrated Development Environment for R. Posit Software, 2023. Disponível em: <http://www.posit.co/>. Acesso: 01 jan. 2022.

58 VENABLES, W. N.; SMITH, D. M.; R Core Team. An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics Version 4.0.5 (2021-03-31)

59 WICKHAM, H.; BRYAN, J. (2023). *readxl: Read Excel Files*. R package version 1.4.2, https://CRAN.R-project.org/package=readxl.

60 WICKHAM, H. (2023). *pryr: Tools for Computing on the Language*. R package version 0.1.6, <https://CRAN.R-project.org/package=pryr>.

61 MEYER, D.; DIMITRIADOU, E.; HORNIK, K.; WEINGESSEL, A.; LEISCH, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-13, <https://CRAN.R-project.org/package=e1071>.

62 SCHLOERKE, B.; COOK, D.; LARMARANGE, J.; BRIATTE, F.; MARBACH, M.; THOEN, E.; ELBERG, A.; CROWLEY, J. (2021). *GGally: Extension to 'ggplot2'*. R package version 2.1.2, https://CRAN.R-project.org/package=GGally.

63 URBANE, S.; JOHNSON, K. (2022). *tiff: Read and Write TIFF Images*. R package version 0.1-11, https://CRAN.R-project.org/package=tiff.

64 KUHN, M. (2008). *Building Predictive Models in R Using the caret Package*. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05.

65 WEI, T.; SIMKO, V. (2021). *R package corrplot: Visualization of a Correlation Matrix (Version 0.92)*. https://github.com/taiyun/corrplot

66 ZEILEIS, A.; HOTHORN, T. (2002). *Diagnostic Checking in Regression Relationships*. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

67 HEBBALI, A. (2020). *olsrr: Tools for Building OLS Regression Models*. R package version 0.5.3, https://CRAN.R-project.org/package=olsrr.

68 WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

69 WICKHAM, H.; VAUGHAN, D.; GIRLICH, M. (2023). *tidyr: Tidy Messy Data*. R package version 1.3.0, <https://CRAN.R-project.org/package=tidyr> .

70 AUGUIE, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3, <https://CRAN.R-project.org/package=gridExtra>.

71 FILZMOSER, P.; VARMUZA, K. (2017). *chemometrics: Multivariate Statistical Analysis in Chemometrics*. R package version 1.4.2, <https://CRAN.R-project.org/package=chemometrics>.

72 TODOROV, V; FILZMOSER, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software, 32(3), 1-47. URL <https://www.jstatsoft.org/article/view/v032i03/>.

73 LUMLEY, T.; Miller, A. (2020). *leaps: Regression Subset Selection*. R package version 3.1, <https://CRAN.R-project.org/package=leaps>.

74 FERNANDES, J. P. S.. Planejamento e síntese de compostos potencialmente ligantes dos receptores 5-HT_2C e H_4. Universidade de São Paulo - USP. São Paulo, 30 de novembro de 2013.

75 FOX, J.; WEISBERG, S. (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>

76 MITRA, I.; SAHA, A.; ROY, K. Chemometric QSAR modeling and in silico design of antioxidant NO donor phenols. Sci Pharm. V. 79, N. 1, p. 31 – 57, 2011.

77 REVELLE, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. R package version 2.3.6, <https://CRAN.R-project.org/package=psych>

78 LE, S.; JOSSE, J.; HUSSON, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

79 LILAND, K.; MEVIK, B.; WEHRENS, R. (2023). pls: Partial Least Squares and Principal Component Regression. R package version 2.8-2, <https://CRAN.R-project.org/package=pls>.

80 MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SÆBØ, S. (2012). A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 118, pp. 62-69.

81 KARATZOGLOU, A.; SMOLA, A.; HORNIK, K.; ZEILEIS, A. (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software, 11 (9), 1-20. doi:10.18637/jss.v011.i09 <https://doi.org/10.18637/jss.v011.i09>.

82 ZOU, H.; HaASTIE, T. (2020). elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. R package version 1.3, <https://CRAN.R-project.org/package=elasticnet>.

83 FRITSCH, S.; GUENTHER, F.; WRIGHT, M. (2019). neuralnet: Training of Neural Networks. R package version 1.44.2, <https://CRAN.R-project.org/package=neuralnet>.

84 BERGMEIR, C.; BENITEZ, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46(7), 1-26.

85 LIAW, A.; WIENER, M. (2002). Classification and Regression by randomForest. R News 2 (3), 18–22.

86 BROWN, S. *Measures of Shape: Skewness and Kurtosis*. BrownMath.com, Atualização: 17/05/2022. Disponível em: https://brownmath.com/stat/shape.htm. Acesso: 15 dez. 2023.

87 KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. New York: Springer, p. 31 - 33, 2013.

88 LEVINE, D. M.; STEPHAN, D. F.; SZABAT, K. A. Estatística: Teoria e aplicações usando o Microsoft Excel em Português, 7ª ed. Rio de Janeiro: LTC, p. 1318, 2016.

89 WEHRENS, R. Chemometrics with R: Multivariate data analysis in the natural sciences and life Sciences. Heidelberg: Springer, 2011.

90 VARMUZA, K.; FILZMOSER, P. *Introduction to multivariate statistical analysis in chemometrics*. Boca Raton: CRC Press, 2009.

91 WICKLIN, R. *Geometry of multivariate versus univariate outliers*. SAS Blogs *The DO Loop*, 25 mar 2019. Disponível em: <https://blogs.sas.com/content/iml/2019/03/25/geometry-multivariate-univariate-outliers.html>

92 VAPNIK, V.N.. Statistical Learning Theory. John Wiley Sons, New York, 1998.

93 DARNAG R., MINAOUI B., FAKIR M.*QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression*. Jornal Árabe de Química Volume 10, Suplemento 1, fevereiro de 2017 , páginas S600-S608.

94 BREIMAN L.; *Random Forest*Kluwer Academic Publishers. Manufactured in The Netherlands Statistics Department, University of California, Berkeley, CA 94720, 2001.

95 UFF - Universidade Federal Fluminense - Material didático do curso de extensão "Introdução ao Machine Learning - oferecido pelo Laboratório de Estatística1 do Departamento de Estatística da Universidade Federal Fluminense, 2023.

96 WALPOLE, R. E.; MYERS, R. H.; MYERS, S. L.; YE, K. Probability Statistics for Engineers Scientists, 9^{th} ed. Boston: Pearson Education, 2012.

97 MONTGOMERY, D. C.; Peck, E. A.; VINING, G. G. Introduction to linear regression analysis, 5^{th} ed. Hoboken: John Wiley Sons,p.168, 2012.

98 GREENE, W. H. Econometric analysis, 6th ed. New Jersey: Pearson education, 2008.

99 FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A.; CARVALHO, A. C. P. L. F. Inteligência artificial: Uma abordagem de aprendizado de máquina, 2ª ed. Rio de Janeiro: LTC, p.151, 2023.

100 IZBICKI, R.; SANTOS, T. M. Aprendizado de máquina: Uma abordagem estatística. São paulo, Editora UICLAP, 1 edição, pág. 14, 2020.

101 ASSUNÇÃO, J. V. C.. Uma breve introdução à mineração de dados: Bases para a ciência de dados, com exemplos em R. São Paulo: Novatec, pág. 185-18,2021.

102 POMERANTSEV, A. L.. Chemometrics in Excel. Hoboken: John Wiley Sons, pág. 161, 2014.

103 ROY, K.; KAR, S.; DAS, R. N. A primer on QSAR/QSPR modeling: Fundamental concepts. Springer, 2015

104 GUJARATI, D.N.; PORTER, D.C. Econometria Básica, 5a ed. Porto Alegre: AMGH, 2011.

105 MALBOUISSON, C.; TAVARES, G. F. Modelo de regressão linear clássico. Econometria na prática. Rio de Janeiro: Altabooks, p. 71, 2017.

106 RAMANATHAN, R.. Statistical methods in econometrics. San Diego: Academic Press, p. 281, 1993.

107 MONTGOMERY, D. C.; RUNGER, G. C. Estatística aplicada e probabilidade para engenheiros. 2ª ed. Rio de Janeiro: LTC, p. 460, 2003.

108 ZAIONTZ, C. Shapiro-Wilk Expanded Test. Real Statistics with Excel. Disponível em: <https://real-statistics.com/tests- normality-and-symmetry/statistical-tests-normality-symmetry/shapiro-wilk-expanded-test/>. Acesso: 25 dez. 2023

109 PEDACE, R. Econometrics for dummies. Hoboken: John Wiley Sons, 2013.

110 TODESCHINI, R.; CONSONNI, V.; GRAMATICA, P. *Chemometrics in QSAR*. In: BROWN, S. D. (editor) et al. Compreensive chemometrics: chemical and biochemical data analysis. Elsevier, 2009.

111 FERREIRA, M . M .c .; KIRALJ,R.. Métodos quimiométricos em relações quantitativas estrutura - atividade (QSAR), p. 387 - 454. *In:* Montanari C.A..Química Medicinal - Métodos e Fundamentos em Planejamento de Fármacos.Editora Universidade de São Paulo, 2019.

112 MONTGOMERY, D. C.; RUNGER, G. C. Estatística Aplicada e Probabilidade para Engenheiros, 7ª ed. Rio de Janeiro: LTC, 2021.

113 CLEAR, J. All models are wrong, some are useful. James Clear. Disponível em: <https://jamesclear.com/all-models-are-wrong>. Acesso: 18 dez. 2022.

114 ROYSTON, J. P. An extension of shapiro and wilk's w test for normality to large samples. Journal of the Royal Statistical Society - Series C (Applied Statistics), v. 31, N. 2, p. 115 – 224, 1982.

115 LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Análise de dados multivariados*. São Paulo: Cengage Learning, 2011.

116 Google. Colaboratory. Disponível em: <https://colab.research.google.com/notebooks/welcome.ipynb>. Acesso em: 02 mar., 2022.

117 StackOverflow. Developer Survey Results 2018. Disponível em: <https://insights.stackoverflow.com/survey/2018technology-most-loved-dreaded-and-wanted-languages>. Acesso em: 02 mar. 2022.

118 SINGH, P. quantitative structure-activity relationship study of substituted-[1,2,4] oxadiazoles as $S1P_1$ agonists. Journal of Current Chemical Pharmaceutical Sciences, V. 3, N. 1, p. 64 - 79, 2013.

119 FRIEDMAN, J. H. Multivariate adaptive regression splines. The Annals of Statistics, V. 19, N. 1, p. 1 – 67, 1991.

120 LEE, C.; YANG, W., Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Physical Review B, V. 37, N. 2, p. 785 – 789, 1988.

121 Secretaria da saúde. *Leishmaniose Visceral*. Disponível em: <https://saude.rs.gov.br/leishmanione-viscera>. Acesso em: 30 fev., 2022.

122 Ministério da saúde. Situação epidemiológica da Leishmaniose Visceral. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/l/leishmaniose-visceral/situacao-epidemiologica-da-leishmaniose-visceral. Acesso em: 28 fev., 2024.

123 SILVA, A. C. L; CARVALHO, B. F. C. O impacto da malária causada pelo Plasmodium vivax em crianças no Brasil. Research, Society and Development, v. 12, n. 11, p. 1 - 9, 2023.

124 Bressan C.; Brasil P.. Malária. Ciência Portal Fiocruz - Agência Fiocruz de Notícias do Instituto Nacional de Infectologia (INI/Fiocruz). Disponível em: < https://portal.fiocruz.br/doenca/malaria>. Acesso em: 04 nov. 2023.

125 PRICE, R. N.; DOUGLAS, N. M.; ANSTEY, N. M. New developments in Plasmodium vivax malaria: severe disease and the rise of chloroquine resistance. Current Opinion in Infectious Diseases, v. 22, N. 5, p. 430 - 435, 2009.

126 OBOH, M. A.; OYEBOLA, K. M.; IDOWUB, E. T.; BADIANE, A. S., OTUBANJO, O. A.; NDIAYE, D. Rising report of *Plasmodium vivax* in sub-Saharan Africa: Implications for malaria elimination agenda. Scientific African, V. 10, e00596, 2020.

127 ROSENBERG, R. Plasmodium vivax in Africa: hidden in plain sight? Trends in Parasitology, V. 23, N. 5, p. 193 - 196, 2007.

128 SILVA, I. N., Spatti, D. H., Flauzino, R. A, Liboni L. H. B. e Alves, S. F. R.. Artificial Neural Networks, *A Practical Course* ,Springer International Publishing Switzerland, 2017. DOI: 10.1007/978-3-319-43162-8.

129 SIMON S. HAYKIN, Redes Neurais, 2 edição, Bookman, 2001.

130 SOUZA, M., SILVA, A. T. R. e CARVALHO, R.L..Redes Neurais MLP Aplicadas na Previsão de Casos Confirmados de Covid-19 no Brasil,International Journal of Development Research, 2020. DOI: 10.37118/ijdr.20278.11.2020.

131 MCKINNEY W.; the Pandas Development TeamPANDAS: powerful Python data analysis, 2022.

132 Teetor, P.. R CookBook. Boca Raton: O'Reilly Media, 2011.

133 AZEVEDO, N. **Desenvolvimento de modelos de predição de atividade inibitória sobre a DNA girase de micobactérias baseados em estudos de modelagem molecular**. Orientador: Carlos Mauricio Sant'Anna. Dissertação (Mestrado em Modelagem Matemática e Computacional), Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, 2021.

134 GALVÃO, R. K. H.; ARAÚJO, M. C. U. Variable selection. *In:* Brown S. D., Tauler R., Walczak B.. Comprehensive chemometrics: chemical and biochemical data analysis. Elsevier, v. 3, p. 233 - 283, 2009.

135 KMENTA, J. Elementos de econometria. São Paulo: Atlas, 1978.

136 VENNERS, B. *The making of Python: A conversation with Guido van Rossum, part I*. ARTIMA, 13/01/2003. Disponível em: <https://www.artima.com/articles/the-making-of-python>. Acesso em: 02 mar., 2022.

137 LUNA, A. S. ; LIMA, I. C. A. ; HENRIQUES, C. A. ; ARAUJO, L. R. R. ; ROCHA, W. F. C. ; VERDAN, J. Prediction of fatty methyl esters and physical properties of soybean oil/biodiesel blends by near and mid-infrared spectra using data fusion strategy. Analytical Methods, p. 4808-4818, 2017.

138 WOLD, S.; JOHANSSON, E.; COCCHI, M. in: PLS: Partial Least Squares Projections to Latent Structures, 3D QSAR in drug design, 1, 1993, pp. 523–550.

139 CENTNER, V.; MASSART, D.; DE NOORD, O.; DE JONG, S.; VANDEGINSTE, B.; STERNA, C. Elimination of uninformative variables for multivariate calibration, Analytical Chemistry, v. 68, p. 3851 - 3858, 1996.

140 WANG, Z. X.; HE, Q. P.; WANG, J. Comparison of variable selection methods for PLS-based soft sensor modeling. Journal of Process Control, v. 26, p. 56 - 72, 2015

141 SOUZA, A. M.; BREITKREITZ, M. C.; FILGUEIRAS, P. R.; ROHWEDDER, J. J. R.; POPPI, R. J. Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: um tutorial, parte II. Química Nova, V. 36, N. 7, p. 1057-1065, 2013.

142 HOERL, A. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, v. 12, n. 1, p. 55 – 67, 1970.

143 Tibshirani, R. Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological), v. 58, n. 1, p. 267 – 288, 1996.

144 JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning with applications in R. New York: Springer, 2013.

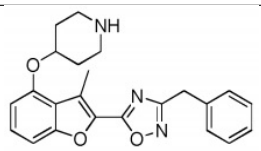
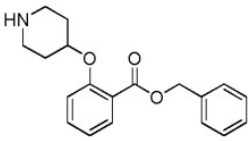
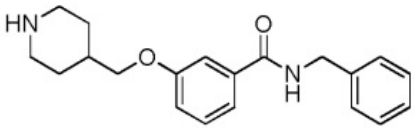
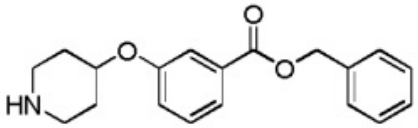
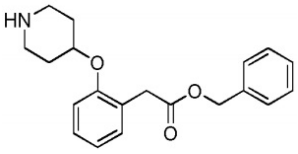
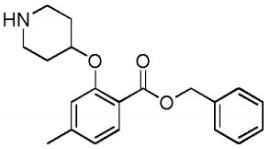
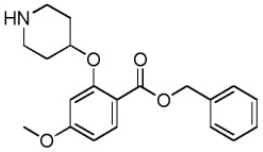
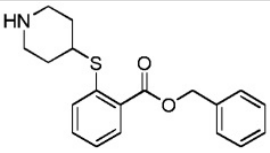
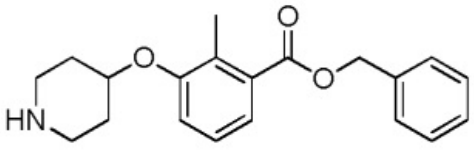
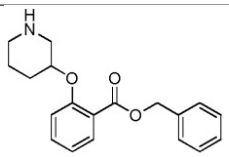
145 MONTGOMERY, D. C.; RUNGER, G. C. Applied statistics and probability for engineers, 3rd ed. New York: John Wiley & Sons, 2003.

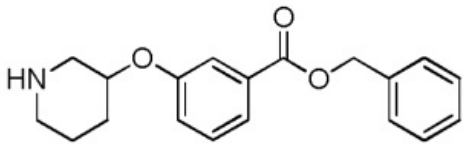
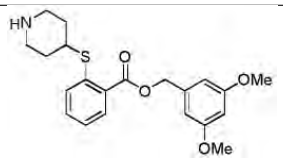
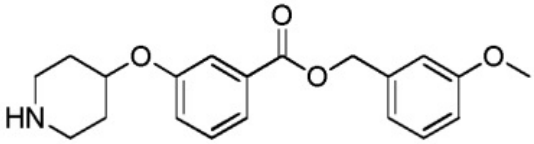
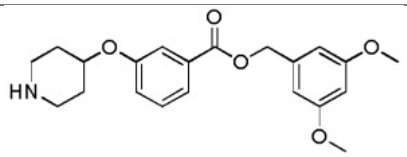
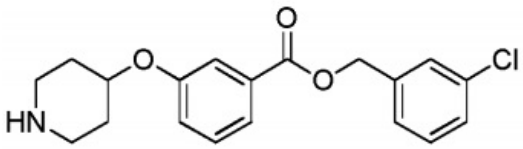
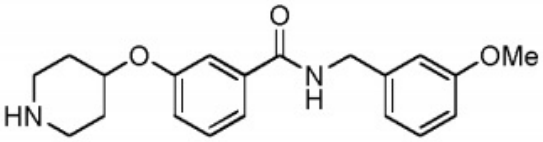
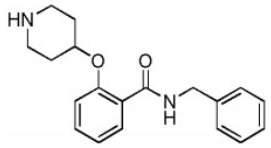
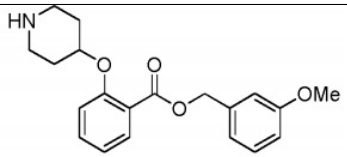
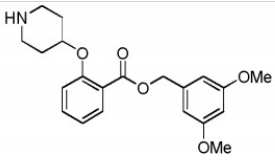
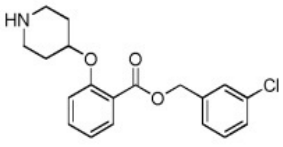
146 MATSUO, T. O uso da regressão de cumeeira em experimentos agronômicos. Dissertação de mestrado. Piracicaba: ESALQ-USP, 1986. 105 f.

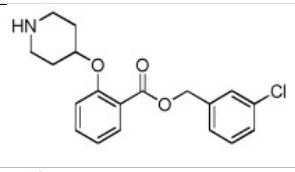
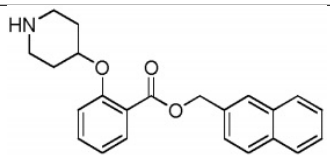
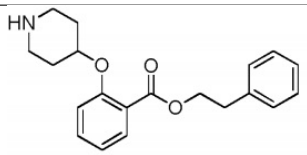
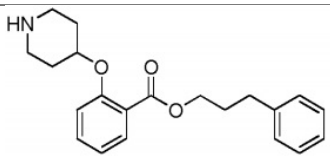
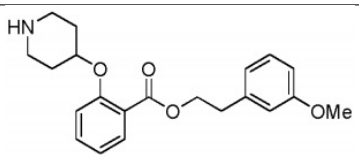
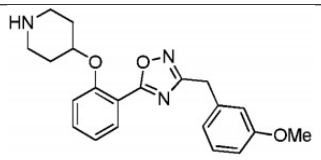
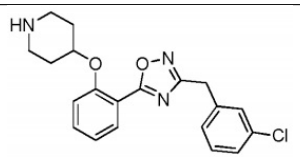
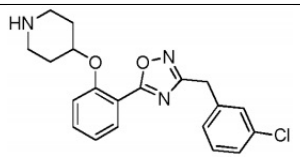
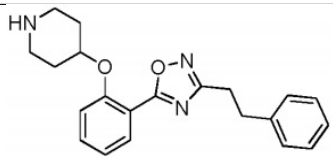
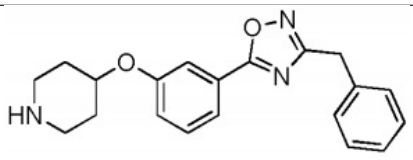
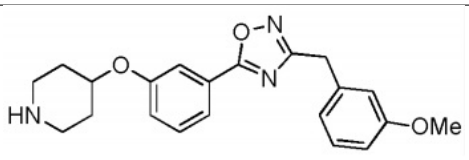
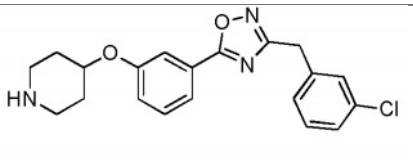
147 PIMENTEL, E. C. G.; QUEIROZ, S. A.; CARVALHEIRO, R.; FRIES, L. A. Estimativas de efeitos genéticos em bezerros cruzados por diferentes modelos e métodos de estimação. Revista Brasileira de Zootecnia, v. 35, n. 3, p. 1020-1027, 2006.

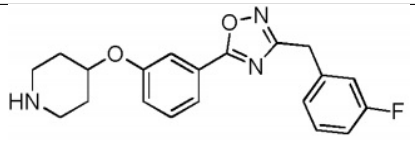
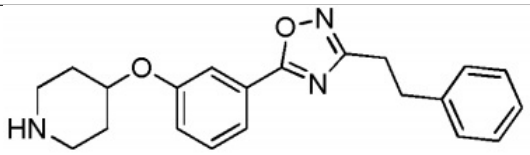
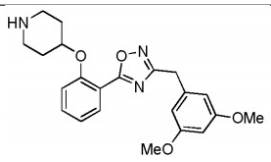
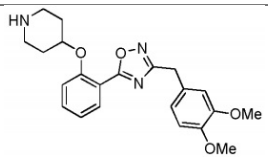
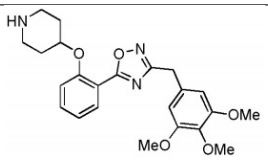
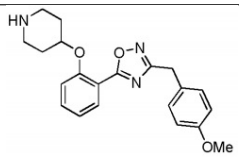
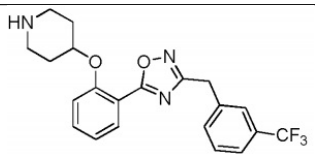
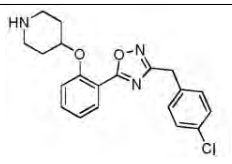
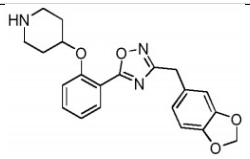
8

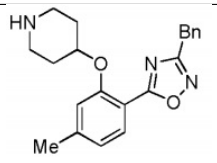
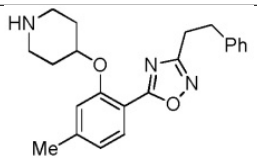
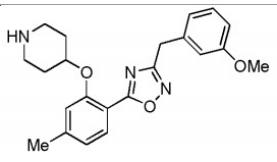
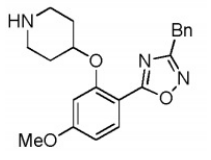
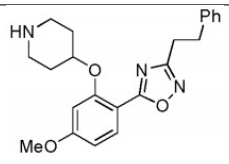
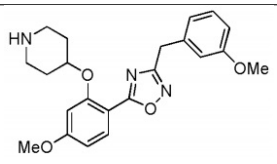
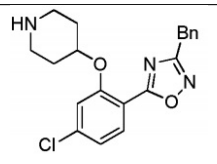
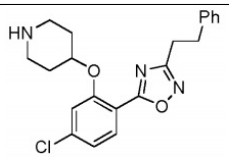
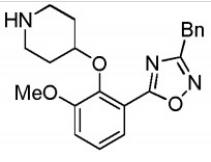
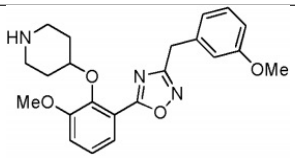
Primeiro apêndice

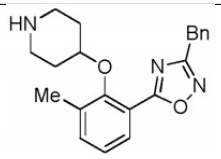
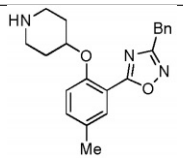
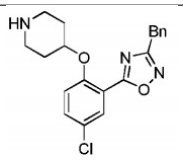
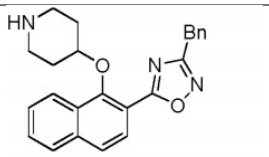
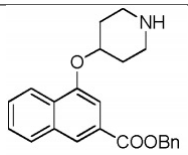
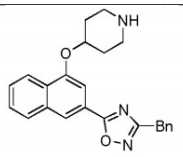
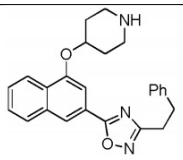
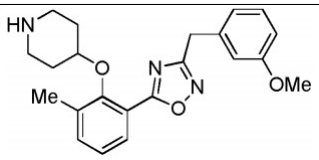
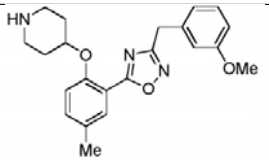
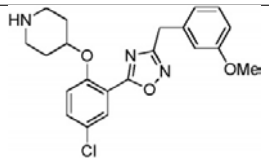
Amostra	Leatherbarrow	Estrutura Molecular	IC_{50} *Ld*
1	23		0.5
2	48		0.6
3	49		3.5
4	50		0.8
5	51		4.2
6	52		9.3
7	53		23.3
8	54		4.3
9	55		1.65
10	56		7

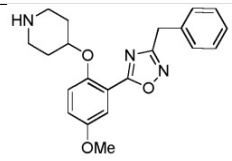
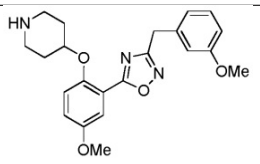
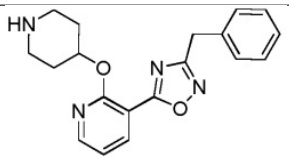
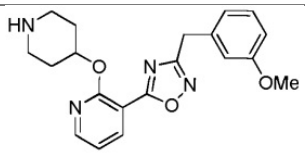
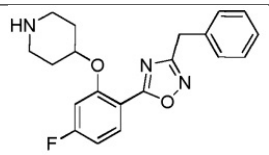
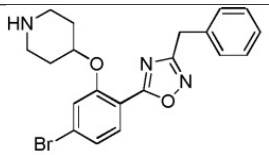
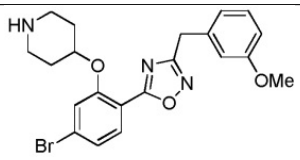
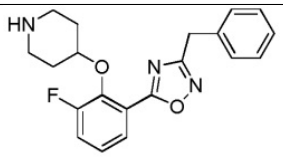
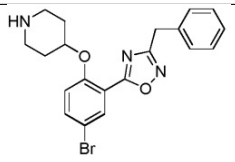
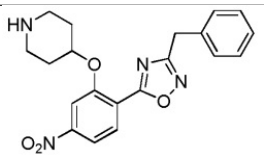
11	57		7
12	58		1
13	59		0.21
14	60		0.17
15	61		0.33
16	62		1.6
17	63		3
18	64		0.31
19	65		0.7
20	67		1.6

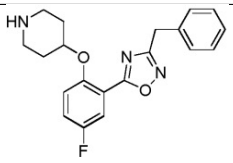
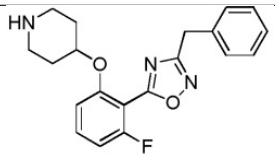
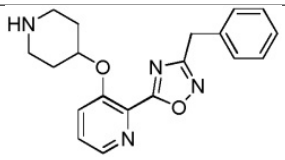
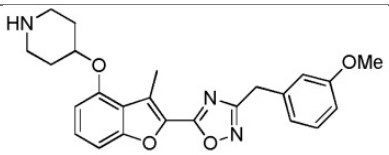
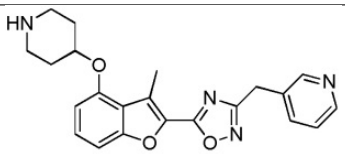
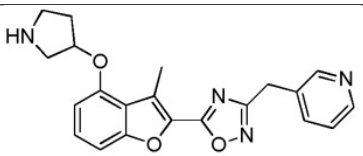
21	68		4
22	69		0.7
23	70		1.5
24	71		0.5
25	73		0.066
26	74		4.39
27	75		0.14
28	76		0.8
29	77		7
30	78		0.15
31	79		0.3
32	80		1

33	81		2.08
34	82		4.59
35	83		3.1
36	84		7.29
37	85		0.39
38	86		8.30
39	87		0.24
40	88		0.28
41	89		34

42	90		135
43	91		5.8
44	92		30.1
45	93		51
46	94		3.4
47	95		5.3
48	96		20.3
49	97		17
50	98		4.5
51	99		7.7

52	100		0.87
53	101		4.29
54	103		38.20
55	108		56
56	110		21.8
57	111		21.8
58	119		1.6
59	120		0.27
60	121		0.02
61	122		8.50

62	123		0.05
63	124		0.39
64	125		0.1
65	126		0.17
66	128		4.8
67	129		0.68
68	130		2.1
69	131		9.9
70	132		3.1
71	134		0.13

72	135		0.65
73	136		4.5
74	138		2.7
75	139		0.01
76	140		7.0
77	24		2.06