UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

ANTONIA REGINA DOS SANTOS GOIS

QUIMIOMETRIA APLICADA À AVALIAÇÃO CLÍNICA DE LESÃO RENAL AGUDA, NEFRITE LÚPICA E ESQUISTOSSOMOSE

Recife
2025

ANTONIA REGINA DOS SANTOS GOIS

QUIMIOMETRIA APLICADA À AVALIAÇÃO CLÍNICA DE LESÃO RENAL AGUDA, NEFRITE LÚPICA E ESQUISTOSSOMOSE

Tese de doutorado apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, como pré-requisito para obtenção do título de doutor em Química. Área de concentração: Analítica.

Orientador (a): Prof. Dr. Ricardo Oliveira da Silva

Recife

2025

ANTONIA REGINA DOS SANTOS GOIS

"QUIMIOMETRIA APLICADA À AVALIAÇÃO CLÍNICA DE LESÃO RENAL AGUDA, NEFRITE LÚPICA E ESQUISTOSSOMOSE"

Tese apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, como requisito para a obtenção do título de Doutor em Química. Área de concentração: Química Analítica.

Aprovada em: 12/09/2025

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
RICARDO OLIVEIRA DA SILVA
Data: 17/09/2025 19:13:02-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ricardo Oliveira da Silva (Orientador)

Universidade Federal de Pernambuco - UFPE

gov.br Documento assinado digitalmente
WELLINGTON PINHEIRO DOS SANTOS
Data: 15/09/2025 15:43:49-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Wellington Pinheiro dos Santos (Examinador Externo)

Universidade Federal de Pernambuco - UFPE

gov.br Documento assinado digitalmente
MARIO RIBEIRO DE MELO JUNIOR
Data: 17/09/2025 18:59:33-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Mário Ribeiro de Melo-Júnior (Examinador Externo)

Universidade Federal de Pernambuco - UFPE

gov.br Documento assinado digitalmente
ANDREA DORIA BATISTA
Data: 17/09/2025 07:28:27-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Andrea Dória Batista (Examinadora Externa)

Universidade Federal de Pernambuco - UFPE

gov.br Documento assinado digitalmente
JOSE LICARION PINTO SEGUNDO NETO
Data: 15/09/2025 12:54:17-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. José Licarion Pinto Segundo Neto (Examinador Externo)

Universidade do Estado do Rio de Janeiro - UERJ

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Gois, Antonia Regina Dos Santos.
Quimiometria aplicada à avaliação clínica de Lesão Renal Aguda, Nefrite Lúpica e esquistossomose / Antonia Regina Dos Santos Gois. - Recife, 2025.
159f.: il.

Tese (Doutorado)- Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-Graduação em Química, 2025.
Orientação: Prof. Dr. Ricardo Oliveira da Silva Recife.

1. Aprendizado de máquina; 2. Lesão Renal Aguda; 3. Nefrite Lúpica; 4. Esquistossomose mansoni. I. Recife, Ricardo Oliveira da Silva. II. Título.

UFPE-Biblioteca Central

A Jesus Misericordioso por tanto amor. A mainha e a painho por me guiarem até aqui.

AGRADECIMENTOS

A Deus, por tanto amor, pelo dom da vida e por ter nos concedido uma natureza tão perfeita.

Aos amores da minha vida: meus pais, Rosa e Domingos, por todo o amor e esforço para que eu pudesse estudar; meus irmãos, Mariana, Antonio e Rafael, por estarem sempre ao meu lado; e meus sobrinhos Davi e Marco, por iluminarem minha vida a cada sorriso banguelo. Sem vocês, eu nada seria.

Ao amor da minha vida, meu esposo Marcelo Hora, por tornar meus dias mais leves, pela paciência e compreensão, e por ser meu apoio constante, minha calmaria em meio ao caos.

Ao meu orientador, Prof. Dr. Ricardo Oliveira da Silva, pela confiança em me aceitar em seu grupo de pesquisa e pelo conhecimento compartilhado, que levarei sempre comigo.

Aos integrantes do LabMeQ, pelas boas conversas, companheirismo, trabalho em grupo e disponibilidade em ajudar.

Aos amigos do pingo de grupo, que são como uma segunda família e estão sempre comigo: Nayara, Guilherme, Diogo, Leo, Marcelinho, Helder, Talita e Hemanoel.

Aos amigos que a vida acadêmica me presenteou, como Camilla, Wenes, Brunna, Thiago, Lary, Leslie, Raissa, Fernando, Carla e Leo.

Ao meu psicólogo e a minha psiquiatra por todo suporte, principalmente nos últimos meses.

Ao grupo de pesquisa em Patologia Molecular e Medicina Genômica, em nome da Professora Paula Sandrin, Gisele, Camila e Braziliano, pela parceria e disponibilidade do material para o estudo.

Ao grupo de pesquisa do Professor Edmundo Lopes, pela parceria e disponibilidade do material para o estudo.

À secretária do PPGQ em nome de Patrícia, por ser sempre tão solícita e não medir esforços para auxiliar os alunos.

À central analítica do DQF em nome de Eliete, por toda disponibilidade e ajuda, além da companhia nos momentos de ansiedade aguardando a instalação do RMN.

Aos professores da banca pelo aceite ao convite.

Aos pacientes dos estudos, que aceitaram participar da pesquisa e doaram seu sangue em prol da pesquisa.

À Universidade Federal de Pernambuco, ao programa de Pós-Graduação em Química, ao CNPq e à Capes, pelo apoio estrutural e financeiro.

Enfim, a todos que, de alguma forma direta ou indireta, contribuíram para o trabalho.

Meu muito obrigada!

RESUMO

Rins e fígado desempenham papel central na homeostase e na eliminação de substâncias tóxicas, sendo frequentemente acometidos por doenças de alta prevalência. Considerando os riscos e as elevadas taxas de mortalidade e morbidade em estágios avançados, o presente trabalho aplicou abordagens metabonômicas e quimiométricas para o diagnóstico de Lesão Renal Aguda (LRA) e o estadiamento de Nefrite Lúpica (NL) e de Fibrose Periportal (FPP) associada à esquistossomose. No estudo da LRA, espectros de RMN de ^{1}H de urina de neonatos prematuros obtidos da literatura foram analisados por algoritmos de aprendizado de máquina, incluindo Regressão Logística (LR), Análise Discriminante Linear (LDA) e Máquina de Vetores de Suporte (SVM). O modelo SVM apresentou melhor desempenho (VPP 100%, sensibilidade 71,4%, especificidade 100% e exatidão 85%), destacando seis metabólitos, valina, lactato, lisina, creatinina, taurina e creatina, relacionados a alterações no metabolismo da taurina. No estudo da NL, amostras séricas foram analisadas por RMN de ^{1}H e os espectros submetidos a combinação dos algoritmos de aprendizado de máquina (LR, LDA e SVM) com técnicas de seleção de variáveis. O modelo LR com seletor SFS-LR apresentou os melhores resultados (exatidão 92,3%), sendo o ácido pirúvico e o lactato os principais discriminantes entre os subtipos de NL, associados a aumento da glicólise e processos inflamatórios. Para a FPP, cinco biomarcadores séricos foram analisados por diferentes algoritmos (LR, LDA, SVM e Árvore de Decisão), com a Árvore de Decisão obtendo o melhor desempenho (exatidão 86%), destacando o número de plaquetas como variável mais relevante. Os resultados reforçam o potencial da metabonômica integrada à quimiometria e ao aprendizado de máquina como ferramenta não invasiva e acessível para diagnóstico e estadiamento de doenças renais e hepáticas.

Palavras-chave: Aprendizado de Máquina; Lesão Renal Aguda; Nefrite Lúpica, Esquistossomose mansoni.

ABSTRACT

The kidneys and liver play a central role in homeostasis and the elimination of toxic substances, and are often affected by highly prevalent diseases. Considering the risks and high mortality and morbidity rates in advanced stages, this study applied metabolomic and chemometric approaches for the diagnosis of Acute Kidney Injury (AKI) and the staging of Lupus Nephritis (LN) and Periportal Fibrosis (PPF) associated with schistosomiasis. In the ARF study, ^{1}H NMR spectra of urine from premature newborns obtained from the literature were analyzed by machine learning algorithms, including Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM). The SVM model performed best (PPV 100%, sensitivity 71.4%, specificity 100%, and accuracy 85%), highlighting six metabolites, valine, lactate, lysine, creatinine, taurine, and creatine, related to changes in taurine metabolism. In the NL study, serum samples were analyzed by ^{1}H NMR, and the spectra were subjected to a combination of machine learning algorithms (LR, LDA, and SVM) with variable selection techniques. The LR model with SFS-LR selector presented the best results (92.3% accuracy), with pyruvic acid and lactate being the main discriminants between NL subtypes, associated with increased glycolysis and inflammatory processes. For FPP, five serum biomarkers were analyzed by different algorithms (LR, LDA, SVM, and Decision Tree), with the Decision Tree obtaining the best performance (86% accuracy), highlighting platelet count as the most relevant variable. The results reinforce the potential of metabolomics integrated with chemometrics and machine learning as a non-invasive and accessible tool for the diagnosis and staging of kidney and liver diseases.

Keywords: Machine Learning; Acute Kidney Injury; Lupus Nephritis; Schistosomiasis mansoni.

LISTA DE ILUSTRAÇÕES

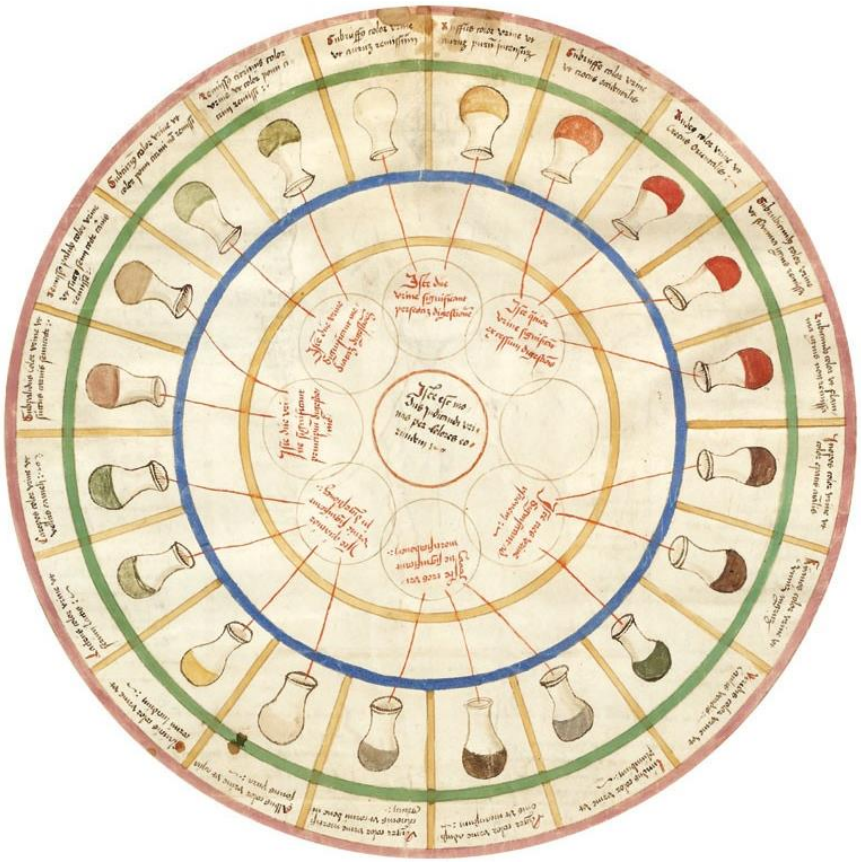
LISTA DE TABELAS

SUMÁRIO

I. PRÓLOGO

A resposta de organismos frente a estímulos ou a ações externas causadoras de doenças ou lesões faz parte de reações complexas e, muitas vezes, imprevisíveis, no qual o objetivo é retornar ao estado de equilíbrio, princípio fundamental da homeostase. Pode-se dizer que essa relação entre doença e perturbação biológica vem sendo investigada desde a Antiguidade. Na Grécia Antiga, já se discutia a importância de modificações em tecidos e fluidos como indicativos de condições clínicas. Na Idade Média, por exemplo, gráficos de urina foram utilizados para associar alterações de cor e odor a diferentes enfermidades (NICHOLSON; LINDON, 2008).

Figura 1 - Roda de urina que descreve possíveis cores, cheiros e sabores da urina e era utilizado para diagnosticar doenças.



Fonte: NICHOLSON; LINDON, 2008

Nesse âmbito, a metabonômica surgiu com o propósito de analisar amostras biológicas de forma global, permitindo a avaliação da resposta metabólica de um organismo frente a agentes externos, e se apresenta como uma abordagem cada vez mais empregada na literatura para auxiliar no entendimento de processos biológicos

complexos. Sua aplicação estende-se a seres humanos, animais, plantas e também à análise da qualidade de alimentos e frutos que possivelmente tenham sofrido alterações externas (CANUTO et al., 2018).

No que diz respeito ao emprego da metabonômica na investigação do perfil metabólico de biofluidos de indivíduos, destacam-se benefícios como a avaliação da suscetibilidade a doenças, a compreensão da resposta e da adaptação a tratamentos, bem como a busca por novos fármacos a partir de descobertas bioquímicas. Para que seja possível a investigação, o fluxo de trabalho de um estudo metabonômico utiliza de técnicas analíticas para análise das amostras e de diferentes métodos quimiométricos e estatísticos para extrair informações dos conjuntos de dados extensos e complexos que são obtidos.

As técnicas quimiométricas são indispensáveis no manuseio de dados durante o estudo metabonômico, em particular, ferramentas de análise multivariadas são altamente empregadas para identificar tendências e padrões. Além disso, com o surgimento e crescente aplicação de algoritmos de aprendizado de máquina, o tratamento de dados complexos alcançou avanços em diversas áreas, inclusive em problemas médicos. Esses algoritmos podem ainda ser empregados em dados de exames de rotina para previsão e compreensão de um conjunto de parâmetros bioquímicos séricos no desenvolvimento de doenças (LIU et al., 2024).

Com o desenvolvimento das tecnologias analíticas e computacionais, novas perspectivas foram abertas para a investigação de inúmeras doenças. Sendo assim, a metabonômica, aliada ao aprendizado de máquina, tem se mostrado uma ferramenta poderosa para identificar alteração metabólicas associados a condições renais (ANEKTHANAKUL et al., 2021). De forma paralela, a análise de biomarcadores séricos provenientes de exames laboratoriais complementa o estudo de doenças hepáticas, fornecendo indicadores clínicos essenciais para diagnóstico e estadiamento (LIU et al., 2024).

Diante deste panorama, o presente trabalho de tese foi motivado pelo interesse em integrar essas abordagens multidisciplinares para avançar na compreensão de três patologias. Examinando aspectos metabonômicos e clínicos, no contexto da Lesão Renal Aguda, Nefrite Lúpica e Esquistossomose Mansoni, por meio da aplicação de técnicas de aprendizado de máquina. Buscou-se extrair informações relevantes tanto dos perfis metabólicos quanto dos biomarcadores séricos.

II. SINOPSE

O presente trabalho está estruturado em três capítulos. O primeiro capítulo apresenta a fundamentação teórica, abordando a estratégia metabonômica, as técnicas analíticas e quimiométricas empregadas durante o desenvolvimento do estudo. O segundo capítulo descreve o desenvolvimento do estudo metabonômico baseado em RMN de ^{1}H, voltado para doenças renais, no qual são utilizados algoritmos de aprendizado de máquina para construir modelos de classificação. O terceiro capítulo traz o desenvolvimento de modelos de classificação utilizando algoritmos de aprendizado de máquina aplicados a biomarcadores séricos para o estadiamento de Fibrose Periportal em pacientes diagnosticados com Esquistossomose Mansoni.

No **capítulo 2**, são descritos modelos de classificação baseados em aprendizado de máquina voltados para o diagnóstico de Lesão Renal Aguda e para o estadiamento de Nefrite Lúpica. Ambos os conjuntos de dados foram obtidos por meio da análise de RMN de ^{1}H. O conjunto referente à Lesão Renal Aguda foi extraído da base de dados *Metabolomics Workbench*, constituído por espectros de amostras de urina. Essa matriz de dados passou por etapas prévias de pré-processamento, incluindo SMOTE; os modelos foram treinados, validados, e as variáveis importantes para a discriminação foram investigadas. Já o conjunto de Nefrite Lúpica foi obtido a partir da análise de amostras de soro realizadas no Departamento de Química Fundamental da UFPE, e passou pelas mesmas etapas do primeiro conjunto, incluindo o uso de algoritmos de seleção de variáveis.

No **capítulo 3**, são descritos modelos de classificação baseados em aprendizado de máquina voltados para o estadiamento da Fibrose Periportal leve e avançada em pacientes infectados com *Schistosoma mansoni*. Biomarcadores séricos obtidos por exames laboratoriais foram utilizados na construção dos modelos, que foram posteriormente validados. A robustez desses modelos foi comparada ao índice Coutinho, já descrito e utilizado na literatura para discriminar os graus dessa fibrose. Uma revisão da literatura sobre a investigação metabolômica e metabonômica da Fibrose Periportal foi publicada e encontra-se disponível no ANEXO 1, servindo como material complementar que amplia o contexto científico deste trabalho.

III. HIPÓTESE GERAL

Modelos quimiométricos são capazes de diagnosticar Lesão Renal Aguda e Nefrite Lúpica e estadiar a Fibrose Periportal na Esquistossomose com precisão.

IV. Objetivo Geral

Desenvolver, a partir de ensaios metabonômicos e quimiométricos, modelos para o diagnóstico de Lesão Renal Aguda e Nefrite Lúpica e estadiamento da Fibrose Periportal na Esquistossomose.

IV.i. Objetivos específicos

- Investigar e otimizar algoritmos de aprendizado de máquina supervisionado em um conjunto de dados disponíveis em um banco de dados de Lesão Renal Aguda em prematuros.
- Avaliar o desempenho dos modelos com base nas figuras de mérito e os deslocamentos químicos importantes para as discriminações de pacientes com e sem Lesão Renal Aguda.
- Obter espectros de amostras de soro de pacientes com Nefrite Lúpica por Ressonância Magnética Nuclear de ^{1}H.
- Investigar e otimizar algoritmos de aprendizado de máquina supervisionado, combinados com técnicas de seleção de variáveis, para o conjunto de dados de Nefrite Lúpica.
- Avaliar o desempenho dos modelos para o diagnóstico de Nefrite Lúpica e suas combinações com os métodos de seleção de variáveis com base nas figuras de mérito.
- Identificar os metabólitos referentes aos sinais de maior importância para construção do melhor modelo para o diagnóstico de Nefrite Lúpica.
- Desenvolver modelos quimiométricos baseados em algoritmos de aprendizado de máquina para classificação das formas leves e avançadas da Fibrose Periportal, por esquistossomose, em pacientes com Esquistossomose Mansoni utilizando biomarcadores séricos (AST, ALT, FAL, GGT e PLT).
- Avaliar o desempenho dos modelos com base nas figuras de mérito e a importância de cada biomarcador sérico na discriminação das classes.

CAPÍTULO 1

1. FUNDAMENTAÇÃO TEÓRICA

1.1. Metabonômica

As ciências ômicas são um conjunto de abordagens que, através da coleta e análise de um grande número de dados, estudam o funcionamento e as alterações biológicas de células, tecidos e organismos inteiros. As ômicas descrevem sistemas biológicos como um todo e são divididas em: genômica, transcriptômica, proteômica, lipidômica, metabolômica/metabonômica, entre outras. Para lidar com a quantidade de dados que são gerados e sua complexidade, ferramentas estatísticas multivariadas e, na maioria das vezes, aprendizado de máquina são empregados (XU et al., 2023; SHAO et al., 2023).

Dentre as abordagens citadas, a metabonômica é responsável por investigar alterações metabólicas em resposta a estímulos fisiopatológicos e intervenções externas que causaram perturbação na homeostase (NICHOLSON et al., 1999). São então os metabólitos, moléculas de baixo massa molar, intermediários ou produtos de reações químicas catalisadas por diferentes enzimas nos sistemas vivos (KISELEVA et al., 2022).

Na literatura, podemos encontrar trabalhos relacionados a estudos metabonômicos e metabolômicos (SU et al., 2021; HANG et al., 2022; MARINO et al., 2021). É frequente o uso dos termos metabolômico e metabonômico como sinônimos. No entanto, há uma diferença filosófica importante, que tem impacto na estratégia metodológica adotada em cada abordagem. Segundo Nicholson e Lindon (2008), a metabonômica visa medir de maneira ampla a resposta metabólica global a estímulos biológicos ou manipulação genética dos sistemas vivos, com objetivo de entender a mudança através do tempo em sistemas complexos. Enquanto a metabolômica, está voltada para discriminação analítica de amostras biológicas complexas através da caracterização e quantificação de todas as espécies químicas presentes na amostra, concentrando-se em um conjunto específico de metabólitos. A necessidade de identificação e quantificação absoluta dos metabólitos exige a utilização de uma ferramenta de separação de misturas (cromatografia, eletroforese) antes da técnica espectrométrica (normalmente, espectrometria de massas). No caso da abordagem metabonômica, como não há necessidade de quantificação absoluta, a técnica espectrométrica (normalmente, espectroscopia de ressonância magnética nuclear) pode ser utilizada na amostra praticamente *in natura* (DUNN; ELIS, 2005). Neste texto,

será utilizado o termo metabonômica, pois nenhuma técnica de separação de misturas será empregada e os dados espectrais serão obtidos usando RMN.

A possibilidade de caracterizar as alterações metabólicas utilizando da metabonômica, pode fornecer informações sobre mecanismos e rotas capazes de auxiliar no diagnóstico e evolução de doenças, bem como a resposta de organismos a tratamentos (BJERRUM et al., 2021; SU et al., 2021). Desse modo, trabalhos na literatura vêm empregando a metabonômica em diversos estudos envolvendo doenças, entre os quais, pode-se apontar aqueles voltados para o diagnóstico e tratamento de diferentes tipos de câncer (CHEN et al., 2022; QI et al., 2022; OLIVEIRA et al., 2024), HIV (GABAZANA; SITOLE, 2021), lesões hepáticas (HU; SHEN; CHEN, 2023; RODRIGUES et al., 2022), doenças renais (LAI et al., 2023), entre tantas outras. Vale ressaltar que, apesar de a maioria dos estudos ser voltada para doenças, a metabonômica também pode ser aplicada em outras investigações, como por exemplo, estudos envolvendo extrato de plantas (DAI et al., 2021; LI et al., 2022).

1.1.1. Biofluidos

A caracterização dos perfis metabólicos é realizada através da análise de fluidos corporais e tecidos, entre eles: derivados sanguíneos (soro e plasma), urina, fezes, fluido seminal, saliva, suor e até lágrimas (GRASSO et al., 2022). A principal vantagem no uso desses biofluidos está associada a métodos minimamente invasivos utilizados no processo de coleta das amostras (GARZARELLI et al., 2022). Os metabólitos e as lipoproteínas, encontrados nos biofluidos, são secretados por diferentes tecidos em resposta a estímulos fisiológicos ou estressores, tornando-os sensíveis não só às condições de saúde, mas também às variações genéticas, fatores ambientais, estilo de vida, hábitos nutricionais e medicamentos, fornecendo informações importantes em termos sistêmicos (VIGNOLI et al., 2022).

Para fim de análise metabonômica, os biofluidos podem ser utilizados isoladamente ou em conjunto. Yanlan e colaboradores (2023), por exemplo, buscaram mecanismos envolvidos nos sintomas intestinais agudos induzidos por radiação em pacientes com câncer cervical e, para alcançar tal objetivo, utilizaram três biofluidos diferentes: urina, plasma e fezes. Dong e colaboradores (2023), que com o intuito de descobrir biomarcadores para a predição de infarto agudo do miocárdio em pacientes com doença arterial coronariana, também analisaram amostras de soro, urina e fezes, onde os metabólitos sanguíneos funcionaram melhor para a predição.

O soro, frequentemente, é um dos biofluidos mais utilizados para representação fenotípica em diagnósticos e tratamentos (GRASSO et al., 2022). Cerca de 60% do corpo humano é constituído por fluido intracelular e extra-celular, como plasma, linfa, líquor, humor aquoso, fluido pleural, sinovial. Na verdade, o sangue, composto por plasma e células (hemacias, plaquetas e leucócitos) circula pelo sistema circulatório, percorrendo todos os tecidos e órgãos (GUYTON; HALL, 2011).

O soro é o líquido obtido quando o sangue total coagula e, em laboratório, a separação é realizada com centrifugação. Sua composição consiste em água, várias proteínas, peptídeos, aminoácidos, hormônios, compostos de nitrogênio, vários íons e sais, vestígios de ácidos nucléicos, metabólitos e lipídios. Por outro lado, o plasma é obtido utilizando um agente anticoagulante, que é adicionado ao sangue total antes de ser centrifugado para remoção das células sanguíneas (KISELEVA et al., 2022). Entretanto, o anticoagulante pode causar perdas nas informações e interferir nas análises. Além disso, o soro costuma apresentar uma concentração maior de metabólitos, como aminoácidos e derivados (SOTELO-OROZCO et al., 2021). Sendo assim, esse será o material biológico estudado no presente trabalho.

1.1.2. Técnicas Analíticas

Os metabólitos que compõem o sangue podem ser divididos em grupos de acordo com suas propriedades físico-químicas. A diversidade das espécies reflete diretamente nas dificuldades de investigação do metaboloma frente às técnicas de análise disponíveis (KISELEVA et al., 2022). Como resultado, diversos estudos metabonômicos nos últimos anos têm explorado uma variedade de técnicas analíticas. As mais comuns incluem a espectroscopia de ressonância magnética nuclear (RMN) e a espectrometria de massas (MS, do inglês, Mass Spectrometry); a segunda é frequentemente usada com técnicas cromatográficas como LC-MS (do inglês, *Liquid Chromatography-Mass Spectrometry*) e GC-MS (do inglês, *Gas Chromatography-Mass Spectrometry*). A escolha da técnica deve ser feita considerando as características das espécies químicas que estão presentes nas amostras.

No que diz respeito ao uso do GC-MS e LC-MS, as quais são técnicas de separação que apresentam alta eficiência, resolução, repetibilidade e, em conjunto com a detecção por MS, uma alta sensibilidade, capaz de detectar compostos com concentrações da ordem de pM (REY-STOLLE et al., 2022; BJERRUM et al., 2021). Diversos estudos metabonômicos, dedicados à progressão e diagnóstico de doenças

por meio dessas técnicas, podem ser encontrados na literatura. No que se refere as análises por cromatografia líquida acoplada a espectrometria de massas, trabalhos incluem a investigação de diferentes tipos de câncer (YANG et al., 2022; YUAN et al., 2023), diabetes (HAO et al., 2023), insuficiência cardíaca (ZHANG et al; 2022), doença hepática (SHEN et al., 2023), doença renal (LEE et al., 2022; JACOB et al., 2022), entre outras condições. O mesmo pode ser observado nos estudos utilizando a cromatografia gasosa, com trabalhos envolvendo diferentes tipos de câncer (WANG et al., 2023; DI GIOVANNI et al., 2023; EROGLU et al., 2022), tuberculose pulmonar (WANG et al., 2022), doença hepática (HUANG et al., 2021), doença renal (FRANIEK et al., 2022), entre outras.

É comum combinar o uso do GC-MS e do LC-MS em um mesmo trabalho como técnicas complementares uma à outra (KOU et al., 2022; WANG et al., 2022). Por exemplo, Wang e colaboradores (2022), utilizando a abordagem metabolômica, empregaram ambas as técnicas cromatográficas para investigar metabólitos presentes em pacientes que apresentaram diabetes mellitus pós-transplante de fígado. A análise multivariada, realizada pelos autores, demonstrou uma alta qualidade do ajuste para os modelos gerados e permitiu selecionar 30 metabólitos diferenciais (15 de LC-MS, 15 de GC-MS) associados às possíveis vias metabólicas envolvidas.

Entretanto, o estudo metabolômico utilizando MS enfrenta dificuldades, principalmente, na etapa de preparo da amostra, que frequentemente é laboriosa, o que pode ser um entrave ao uso da técnica. Por outro lado, a etapa de identificação dos metabólitos é facilitada, visto que costuma ser feita a partir da comparação dos espectros com bibliotecas espectrais comerciais, com padrão de fragmentação que constam em bancos de dados de metabólitos e, se o equipamento permitir, através da elucidação estrutural de espectros obtidos de fragmentação MS/MS (REY-STOLLE et al., 2022; BJERRUM et al., 2021).

A espectroscopia de ressonância magnética nuclear (RMN) surge então como uma alternativa complementar as demais técnicas citadas. Apesar de apresentar uma menor sensibilidade comparada ao MS, os resultados nesse quesito podem ser melhorados por meio do aumento do número de varreduras, da intensidade de campo, do resfriamento criogênico, do uso de microssondas e de métodos de hiperpolarização (SAIGUSA et al., 2021; RAJA et al., 2020). A técnica ainda apresenta outras vantagens, como robustez, reprodutibilidade e requer mínimo preparo de amostra (GHINI et al., 2023). Além disso, o tempo de análise é relativamente baixo, todos os

metabólitos são medidos simultaneamente e, caso seja o foco do trabalho, compostos de interesse podem ser quantificados (RAJA et al., 2020).

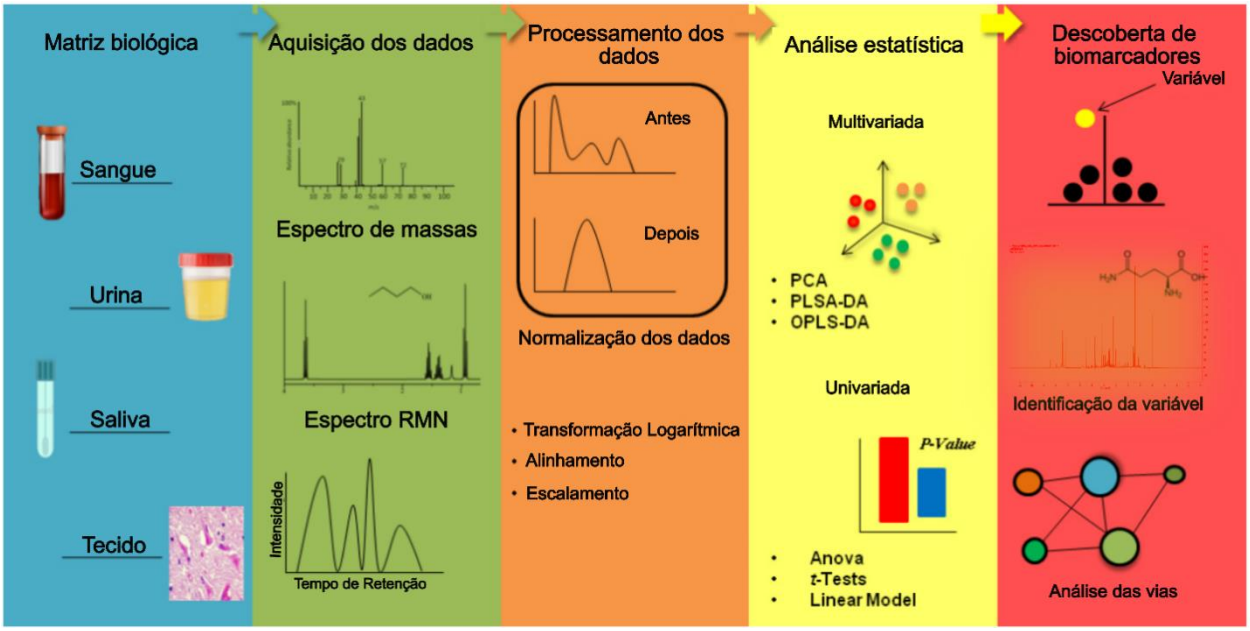
Uma série de trabalhos na literatura está voltada para estudos metabonômicos que utilizam a RMN como técnica de análise (ZHONG et al., 2019; JIN et al., 2022, HANG et al., 2022; YANG et al., 2021). Com relação à investigação de doenças, é possível encontrar trabalhos envolvendo diferentes tipos de câncer (HASUBEK et al., 2023; OLIVEIRA et al., 2024; CARDOSO et al., 2022; RAZAVI et al., 2024), diabetes (BRAGG et al., 2022), doenças cardíacas (DONG et al., 2023; HASSELBALCH et al., 2023), depressão (HUNG et al., 2021), meningite tuberculosa (PARIHAR et al., 2022), infertilidade masculina (NETO et al., 2022), doenças renais (FONSECA et al., 2023; KIM et al., 2023), osteoporose (PONTES et al., 2019), entre outras condições.

O emprego do RMN permite a análise de amostras sem preparos laboriosos, o que facilita a investigação de diferentes materiais biológicos. Trabalhos como os citados acima, mostram a infinidade de possibilidades quando se trabalha com metabonômica, isso sem citar a diversidade de abordagens que podem ser empregadas durante o tratamento dos dados.

1.1.3. Fluxograma do estudo metabonômico

Para um estudo metabonômico, é necessário definir o que será investigado e a população de estudo, bem como o material a ser analisado, o preparo da amostra, a aquisição de dados e outras etapas que compõem o fluxo de trabalho. Todo o processo, desde a coleta da amostra até a identificação dos metabólitos, consiste em uma série de etapas que devem ser seguidas para garantir o sucesso durante a execução do estudo. Alguns trabalhos na literatura, como o de DE SAN-MARTIN e colaboradores (2021) e de Bjerrum e colaboradores (2021), discutem o fluxo de trabalho comumente aplicado, apresentado na Figura 2.

Figura 2 – Etapas no desenvolvimento de um estudo metabonômico.



Fonte: Adaptado de DE SAN-MARTIN e colaboradores (2021).

A seleção da população de estudo deve considerar características como o estágio e grau da doença, intervenções cirúrgicas, entre outras. Os dados demográficos e clínicos precisam ser devidamente registrados, visto que os metabólitos podem ser alterados devido à idade, sexo, estilo de vida, condições fisiopatológicas e intervenções cirúrgicas. A coleta e o preparo das amostras são processos importantes para garantir a integridade das mesmas durante todo o período até a análise. Antes de preparar a amostra, todas devem ser randomizadas para evitar o viés analítico. Na maioria das vezes, apenas a diluição da amostra em D_2O é suficiente para a análise no RMN. A aquisição dos espectros é realizada através de experimentos de RMN de ^1H e, quando necessário, experimentos bidimensionais podem ser aplicados. O processamento dos espectros é feito para gerar uma matriz numérica e envolve a redução de ruído e correção de linha de base. O pré-processamento da matriz inclui a aplicação de normalização das amostras, escalamento ou transformações das variáveis. Na análise multivariada, busca-se por padrões ou tendências de agrupamento das amostras e constrói-se modelos de predição com valores significativos de validação para que possam ser usados na identificação dos metabólitos. A identificação dos metabólitos é realizada atribuindo sinais indicados pelo modelo como importantes.

1.2. Espectroscopia de RMN

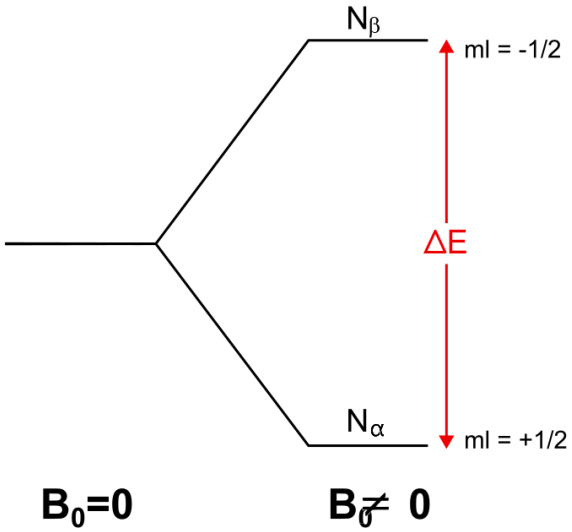
A espectroscopia ressonância magnética nuclear (RMN) apresenta vantagens frente as técnicas cromatográficas, principalmente, no que diz respeito a facilidade do preparo de amostra (EMWAS et al., 2019). É considerada a técnica mais poderosa para a análise química (NEULING et al., 2023) e uma das predominantes, além do MS, na metabonômica e na análise estrutural (MIELKO et al., 2021).

1.3. Fundamentos

A espectroscopia de Ressonância Magnética Nuclear (RMN) envolve a interação entre a radiação eletromagnética, na faixa da radiofrequência, e a matéria, mais especificamente, os núcleos atômicos. Os núcleos atômicos possuem quatro propriedades físicas fundamentais: massa, carga elétrica, momento magnético e momento angular. No contexto da RMN, é crucial compreender a interação entre o momento magnético nuclear e o campo magnético aplicado. Os núcleos magnéticos se comportam como pequenas barras magnéticas frente a um campo magnético, e seu número de spin (*I*) é uma propriedade intrínseca associada às partículas subatômicas e uma forma de momento angular (LEVITT, 2008).

Partículas com spin *I* possuem (*I*) níveis degenerados, mas na presença de um campo magnético externo a degenerescência é quebrada (LEVITT, 2008). Por exemplo, núcleos com *I* = 1/2, como é o caso do núcleo de hidrogênio-1, podem assumir dois estados quânticos. Na ausência do campo magnético externo (B_0), ambos os estados quânticos são equivalentes e nenhuma excitação é possível. Na presença de B_0 há quebra da degenerescência, o dipolo magnético (μ) do núcleo se alinha na direção do campo aplicado e o núcleo tem duas possibilidades de estados quânticos, de menor e de maior energia (DIEHL, 2008). A diferença de energia entre os dois níveis gerados (α e β) é dada por $\Delta E = hv$. A Figura 3 representa os dois níveis de energia de núcleos com spin 1/2.

Figura 3. Geração de dois níveis de energia em núcleos com spin $\frac{1}{2}$ frente a um campo magnético externo.



Fonte: Adaptado de DIEHL (2008).

A ocupação nos dois estados de energia depende da energia térmica e magnética. Quando em temperatura ambiente e o campo aplicado é fraco, as populações em ambos os estados são praticamente iguais, de modo que os spins em direções opostas quase se cancelam (SILVERSTEIN et al., 2019). O sinal detectado é resultado do excesso de spin que ocupa o estado de menor energia, pois esses spins podem ser excitados pelo pulso de radiofrequência e, ao retornarem ao equilíbrio, liberam a energia que gera o sinal registrado.

O excesso de spin pode ser aumentado diminuindo a energia térmica, reduzindo a temperatura. Por exemplo, considerando apenas o efeito da temperatura, no zero absoluto todos os spins estariam no estado de menor energia, consequentemente, o excesso de spin seria máximo, apesar de não ser algo prático de se desenvolver em um laboratório. Ao invés disso, a intensidade do campo pode ser aumentada, logo, maior será a energia entre os estados, menor a excitação dos spins para N_β e maior o excesso de spins em N_α (SILVERSTEIN et al., 2019).

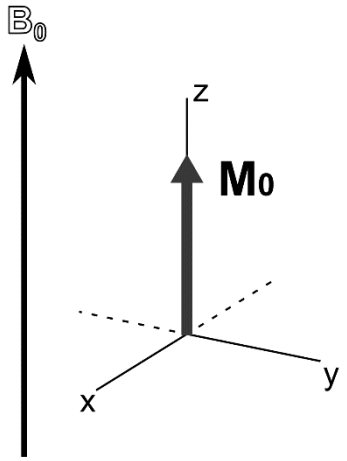
A diferença da população entre os dois estados pode ser calculada através da distribuição de Boltzmann (Eq. 1):

$$\frac{N_\alpha}{N_\beta} = e^{\Delta E / k_B T} \quad (1)$$

A distribuição de Boltzmann mostra a forte dependência entre as populações e a intensidade do campo magnético. Em um campo magnético forte, como em 600 MHz, a relação N_α/N_β é apenas 0,999904. Mesmo em uma pequena fração de excesso de spin, a quantidade do núcleo presente na amostra é vasta permitindo observar um número maior de spins que serão responsáveis pelo sinal no espectro (DIEHL, 2008).

Para auxiliar no entendimento do experimento dentro do RMN, pode-se usar o sistema de coordenadas. Considerando o sistema em equilíbrio térmico, a aplicação do campo estático B_0 faz com que o momento magnético nuclear precessione na direção do eixo Z, resultando em uma magnetização (M_0), que corresponde ao somatório dos momentos magnéticos no eixo z (Eq. 2).

Figura 4 – Sistema de coordenadas representando o vetor de magnetização resultante do efeito de B_0.



Fonte: A autora (2025).

$$M_0 = \Sigma\mu \tag{2}$$

O movimento de precessão possui uma frequência angular chamada de frequência de Larmor (v). A equação fundamental da RMN (Eq. 3) correlaciona a frequência de Larmor com a intensidade do campo magnético. A constante magnetogírica (γ) é específica para cada núcleo e indica a proporcionalidade entre momento magnético e número de spin, como demostrado na Eq. 4 (SILVERSTEIN et al., 2019).

$$v = \left(\frac{\gamma}{2\pi}\right) B_0 \tag{3}$$

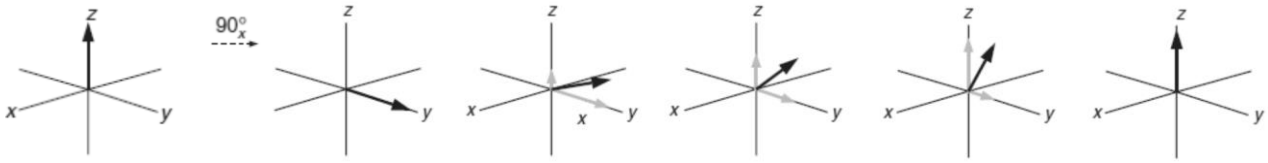
$$\gamma = \frac{2\pi\mu}{hI} \quad (4)$$

Através da equação fundamental da RMN, é possível descrever um instrumento em termos da intensidade do campo magnético ou em relação a frequência de Larmor em unidades tesla (T) ou megahertz (MHz), respectivamente. Por exemplo, um equipamento com campo de 9,39 T é chamado de espectrômetro de RMN de 400 MHz, frequência de ressonância do ^{1}H frente ao campo.

Um experimento simples de RMN pulsado é realizado com um campo magnético oscilante (B_1), de frequência ν, aplicado na forma de pulso perpendicular ao campo magnético estático (B_0). Esse pulso de radiofrequência rotaciona o vetor de magnetização (M_0) do eixo z para o plano xy, resultando na nova componente de magnetização M_{xy}. Durante esse processo, ocorre a coerência de spin entre os spins nucleares, garantindo uma relação de fase entre eles. A voltagem induzida pela precessão dos spins no plano xy é recebida na bobina da sonda, e o sinal resultante é conhecido como decaimento livre da indução (FID). Em seguida, a magnetização retorna ao eixo z, restabelecendo o equilíbrio, processo conhecido como relaxação (COLNAGO; ANDRADE, 2017). O sinal é adquirido como uma função do tempo e, por meio da transformada de Fourier, é convertido para o domínio da frequência, permitindo a observação do espectro de RMN.

Após o pulso, dois processos de relaxação estão envolvidos no FID, longitudinal e transversal. A relaxação longitudinal refere-se ao processo pelo qual a magnetização retorna ao equilíbrio ao longo do eixo do campo magnético B_0 e é caracterizada pelo tempo T1, que indica quanto tempo Mz leva para retornar à condição inicial. A relaxação transversal está relacionada ao decaimento da magnetização no plano xy devido à perda de coerência de fase entre os spins, sendo representada pelo tempo T2, que indica quanto tempo Mxy leva para retornar à condição inicial (COLNAGO; ANDRADE, 2017). Os tempos de relaxação são variáveis extremamente importantes na RMN e podem ser empregados em diferentes experimentos como uma forma de filtrar espécies durante a aquisição do espectro. Embora T1 e T2 geralmente tenham a mesma ordem de grandeza, em moléculas maiores e/ou sistemas mais viscosos, T2 tende a ser menor que T1. Nas seções dedicadas aos experimentos aplicados, no presente trabalho, isso será abordado novamente.

Figura 5 – Vetor de magnetização após o pulso perpendicular ao campo magnético B_0. Componentes no plano xy (que determina T2) e no eixo z (que determina T1).



Fonte: Adaptado de SILVA (2017)

É relevante destacar que a frequência de absorção do spin nuclear pode ser afetada e, consequentemente, os tempos de relaxação sofrerem alterações. Por exemplo, a intensidade do campo magnético no núcleo pode ser alterada pela densidade eletrônica que o envolve. Essa proteção, resultante da precessão dos elétrons sob a influência de B_0, gera um campo magnético adicional que pode se opor ao campo magnético aplicado. Com o aumento da densidade eletrônica ao redor do núcleo, o campo efetivo diminui, levando a frequências de ressonância mais baixas. Em outras palavras, além da dependência do campo B_0, a frequência pode variar a depender do ambiente químico em que o spin nuclear se encontra (RULE; HITCHENS, 2006).

A escala de deslocamento químico (δ) é a forma utilizada para remover a dependência da frequência de ressonância do campo B_0, onde todas as frequências são convertidas para uma escala adimensional. Desse modo, é possível comparar espectros obtidos em campos magnéticos com diferentes intensidades (RULE; HITCHENS, 2006). O deslocamento é calculado a partir da Eq. 5, sendo expresso em ppm:

$$\delta = \frac{(v_A - v_R)}{v_R} \times 10^6 \ (ppm) \qquad (5)$$

onde, v_R é a frequência do núcleo de referência e v_A a frequência do núcleo investigado. Com as variações nos deslocamentos químicos ao longo do espectro é possível atribuir os sinais com base no ambiente químico e, consequentemente, elucidar estruturas dos compostos investigados.

1.3.1. Sequência de pulsos Pré-saturação

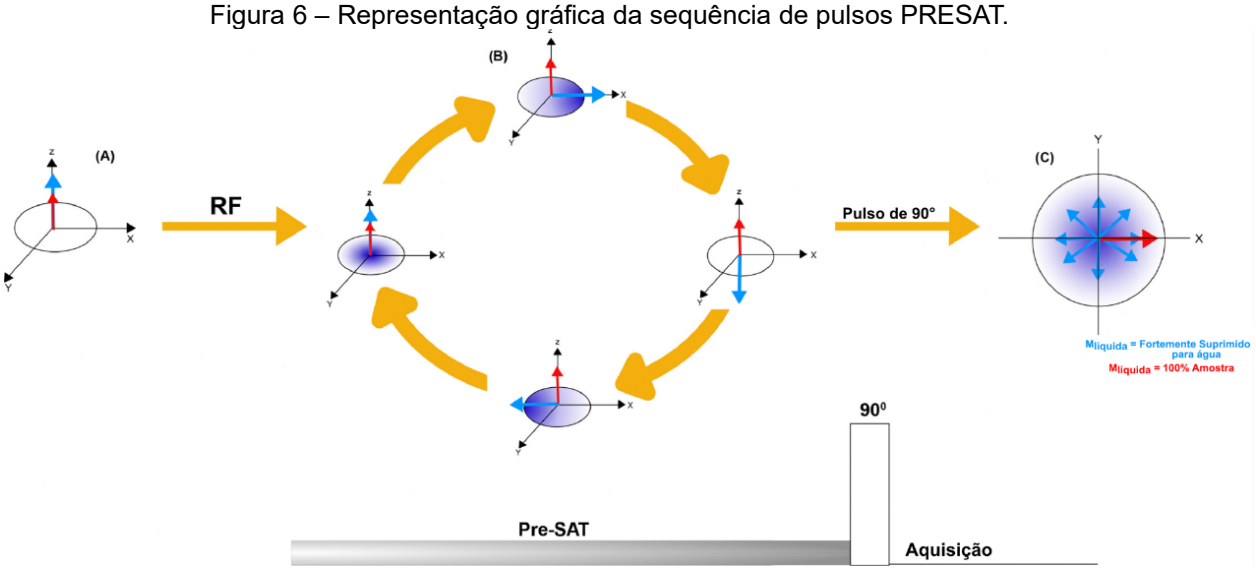
Amostras de biofluidos (soro, plasma ou urina) apresentam uma grande quantidade de água em sua composição. Quando submetida à análise por

espectroscopia de RMN de ^{1}H, a água apresenta um sinal intenso no espectro que pode sobrepor sinais de analitos importantes para o estudo, além de saturar o detector e causar distorções de linha de base (ZHENG; PRICE, 2010). Algumas alternativas podem ser aplicadas para minimizar ou reduzir ao máximo o sinal de próton do solvente, entre elas, o uso de solventes deuterados e os métodos de supressão são os mais comuns (MO; RAFTERY, 2008).

A sequência de pulsos de pré-saturação do sinal do solvente, mais conhecida como PRESAT, é um método comumente aplicado na supressão do sinal de água. Vantagens como a facilidade de uso e a simples implementação são responsáveis pela sua popularidade. O experimento consiste em um pulso constante de baixa intensidade na frequência de ressonância do solvente que, com o tempo, resulta na quebra de coerência dos spins ao redor do plano xy durante a aquisição, evitando a aquisição do sinal na bobina (SOONG et al., 2024).

No equilíbrio, o vetor de magnetização da água e do soluto estão alinhados em torno do eixo z, o pulso é então aplicado (entre 1 e 10 s) ao longo do eixo y na frequência de ressonância do solvente (água, no presente trabalho) fazendo com que o vetor gire em torno do eixo zx, perdendo a coerência dos spins. A moléculas em investigação não sofrem perturbação, então um pulso de 90º é aplicado invertendo sua magnetização e da água para plano xy para obtenção do sinal. Os spins da água permanecem decoerentes e se cancelam, o resultado é um sinal do solvente bastante reduzido, mantendo o sinal dos analitos de interesse (SOONG et al., 2024). A Figura 6 ilustra esse processo.

Figura 6 – Representação gráfica da sequência de pulsos PRESAT.



Fonte: SOONG e colaboradores (2024)

1.3.2. Sequência de pulsos NOESY

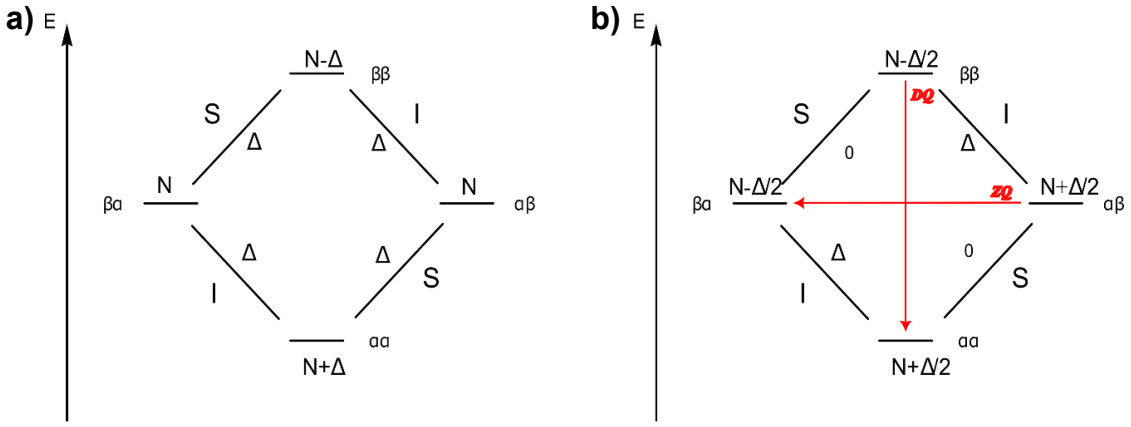
A depender do tipo de biofluido a ser analisado e dos analitos de interesse, diferentes sequências de pulsos podem ser empregadas com o intuito de obter melhores resultados. Para amostras de soro ou plasma, em geral, são aplicados experimentos de CPMG (Carr-Purcell-Meiboom-Gill) e de difusão. Para urina, experimentos de espectroscopia do efeito nuclear Overhauser (NOESY) são suficientes, visto que os sinais do espectro de RMN desse material são constituídos praticamente de metabólitos de baix(GHINI et al., 2023).

Os núcleos podem ter dois tipos de interações ou acoplamentos magnéticos: dipolo-dipolo e spin-spin, sendo que o segundo é conhecido como acoplamento escalar. A sequência de pulsos NOESY está fundamentada na teoria do efeito nuclear Overhauser (NOE). O NOE se baseia no acoplamento dipolo-dipolo entre dois spins: um spin de interesse (*I*), que terá sua intensidade de sinal influenciada pelo relaxamento de um spin perturbado (*S*).

Durante o experimento, o spin S é saturado ($N_\alpha = N_\beta$) com um pulso longo de baixa frequência. O spin I, espacialmente próximo, ainda apresenta diferença de população (Δ) entre os estados, mas quando o sistema busca reestabelecer o equilíbrio e igualar as populações do spin S, suas transições interferem no equilíbrio das populações do spin I, afetando a intensidade do sinal de forma positiva ou negativa (DOOST et al., 2019; KUMAR; GRACE, 2017).

O efeito pode ser observado por meio de diagramas de energia, Figura 7. O equilíbrio pode ser reestabelecido por meio dois tipos de transições de energia, zero-quanta (ZQ) ou duplo-quanta (DQ). Transições proibidas pela mecânica quântica, mas que podem ocorrer devido as interações spin-rede (SILVA, 2010).

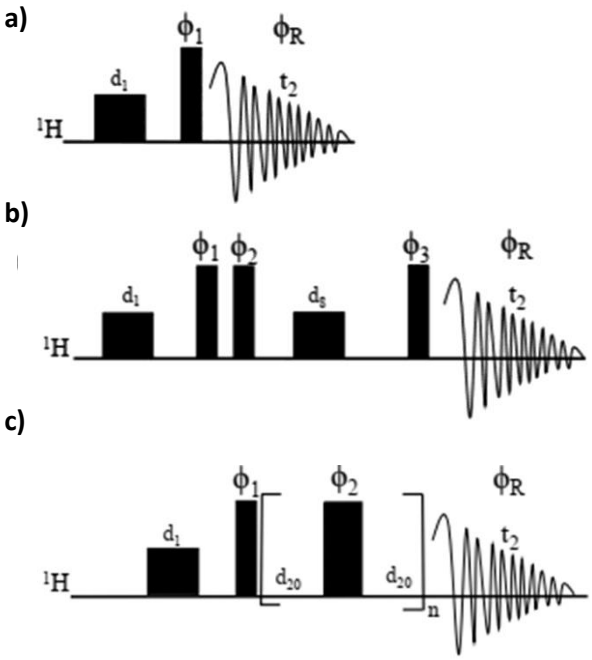
Figura 7 – Diagrama de energia dos spins (I e S). a) Distribuição populacional na presença de B_0. b) Populações do spin S igualadas por meio das transições proibidas.



Fonte: Adaptado de SILVA (2010).

A combinação do PRESAT e NOESY, NOESY 1D PRESAT, é considerada uma das sequências de pulsos preferidas em estudos metabolômicos/metabonômicos, devido à sua alta capacidade de supressão do sinal da água (SINGH et al., 2023). As medições de NOE utilizam três pulsos de 90°: o primeiro alinha a magnetização dos spins para o plano transversal, seguido de um período de evolução t1; o segundo pulso alinha a magnetização na direção longitudinal. Como resultado dessa magnetização fora do equilíbrio, o NOE é produzido durante um período de mistura (τ_{mix}). No terceiro pulso, a magnetização encontra-se novamente no plano transversal e os dados são adquiridos (t2) (KUMAR; GRACE, 2017; HUSSAIN et al., 2022). A Figura 8 apresenta graficamente as sequências de pulsos citadas e a CPMG, sequência que será abordada na próxima seção.

Figura 8 – Sequências de pulsos. (a) PRESAT, (b) NOESY 1D com PRESAT, e (c) CPMG com PRESAT. Os termos d_1, d_8, d_{20} e t_2 são atraso de relaxamento, tempo de mistura, tempo de meio-spin-eco e tempo de aquisição, respectivamente; e ϕ_1, ϕ_2, ϕ_3, ϕ_4 e ϕ_5 são fases de pulsos, enquanto ϕ_R é a fase do receptor



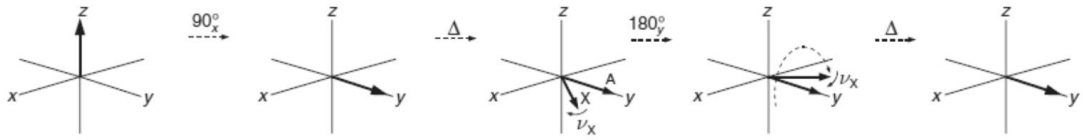
Fonte: Adaptado de SINGH e colaboradores (2023).

1.3.3. Sequência de pulsos CPMG

Além da quantidade de água, amostras de biofluidos, principalmente soro, possuem compostos de alta massa molar em sua composição, geralmente proteínas, que apresentam sinal alargado no espectro e podem sobrepor os sinais de metabólitos de interesse. Apesar das proteínas poderem ser removidas nas etapas de preparo de amostra, há preocupações a respeito da integridade da amostra durante o processo (MA et al., 2023). Como alternativa, a sequência de pulsos conhecida como filtro de T_2 ou CPMG (Carr-Purcell-Meiboom-Gill) pode ser empregada com o objetivo de filtrar os compostos de alta massa molar, sem a necessidade de remoção física prévia.

O CPMG foi baseado na técnica *spin echo* de HAHN (1950), publicada em seu artigo em 1950s, com sequência de pulsos proposta por Carr e Purcell (1954) e com sucessivos pulsos coerentes propostos por Meiboom e Gill (1958). O experimento consiste na aplicação de um pulso de 90º, levando a magnetização para o plano xy, seguido de uma série de pulsos de refocagem de 180º no eixo y em intervalos de tempo (τ).

Figura 9 – Sequência de pulsos CPMG



Fonte: SILVA (2017)

Inicialmente, o vetor de magnetização está na direção do campo magnético externo, em seguida o pulso de 90° é aplicado levando a magnetização para o plano xy. Logo depois, a magnetização transversal e o sinal medido decaem no intervalo de tempo τ, consequência da defasagem dos spins. A defasagem é o fenômeno onde vários spins apresentam frequências de Larmor diferentes da frequência base de Larmor (ω_0). Com o pulso de 180º aplicado em y, as componentes no plano são giradas, invertendo a direção dos vetores que são refocalizados no tempo τ gerando o sinal de *echo* (COLNAGO; ANDRADE, 2017; JUNG; WEIGEL, 2013).

Há duas situações nas quais os efeitos de defasagem levam ao relaxamento transversal e que devem ser consideradas na sequência de *spin echo*: quando o efeito de defasagem é causado apenas pelas heterogeneidades do campo magnético e quando é causado por flutuações adicionais do campo que variam com o tempo. Nesse segundo caso, as flutuações no campo ocorrem devido as interações aleatórias de *spin-spin*, resultado do movimento browniano das moléculas. Desse modo, os *spins* encontram-se ambientes e frequências diferentes antes e depois do pulso medido, ocasionando uma defasagem estocástica dos spins e levando ao decaimento irreversível da magnetização transversal e do sinal medido (JUNG; WEIGEL, 2013).

Compostos de alta massa molar, como as proteínas, apresentam valores menores de T2 em comparação com moléculas de baixa massa molar e relaxam mais rapidamente. Portanto, é crucial que o tempo entre o pulso de 180° e o sinal de echo seja suficientemente longo para permitir que a magnetização transversal dos spins das espécies a serem filtradas decaia, mas ainda menor que o T2 das espécies de interesse. No presente trabalho, a sequência de pulso CPMG será aplicada após a supressão da água na análise de amostras de soro.

1.3.4. Processamento dos espectros

O processamento dos espectros de RMN é de fundamental importância antes de seguir para o tratamento estatístico e visa remover variações indesejadas causadas

pelo estado da amostra, diluição, modo de medição, entre outros fatores. As etapas do processamento dos espectros envolvem a correção da linha de base e da fase, alinhamento dos sinais, *binning* espectral e construção da matriz de dados.

Desajustes de fase entre o campo RF e o receptor do sinal podem resultar em espectros com inversão de fase, situação na qual há uma mistura de linhas espectrais absortivas desejáveis e linhas dispersivas, resultando em picos que não têm forma simétrica e podem apresentar partes positivas e negativas (WORLEY; POWERS, 2014). A correção da fase pode ser realizada em softwares, como o MestReNova, de forma manual ou automática. Variações de cunho experimental e instrumental, assim como distorções e/ou ruídos de baixas frequências, na linha de base, podem ser removidas ou minimizadas realizando a correção da linha de base. Esse ajuste pode ser realizado por meio de ajustes polinomiais (SUN; XIA, 2024).

Deslocamentos espectrais são comuns em espectros de RMN, sendo o alinhamento dos picos fundamental para a modelagem (MISHRA et al., 2020). Esse alinhamento pode ser realizado de maneira simples, utilizando um sinal de referência de um padrão adicionado à amostra ou um sinal conhecido presente na mesma. Alternativamente, existem métodos mais robustos que utilizam espectros de referência para garantir um alinhamento preciso (SUN; XIA, 2024). O método de *binning* minimiza o efeito das variações de deslocamento espectrais entre as amostras e ainda reduz o número de variáveis a serem analisadas. Ele divide o espectro em regiões (*bins*) de largura uniforme, que são integradas para compor a matriz de dados (CHAI et al., 2023). Geralmente, a largura do bin varia entre 0,01 e 0,05 ppm (SAVORANI et al., 2010).

Após as etapas de processamento dos espectros, os dados são organizados em uma matriz, na qual as variáveis estão nas colunas (deslocamento químico) e as amostras nas linhas. Nesse formato, a matriz de dados está pronta para as etapas de pré-processamento e pré-tratamento.

1.4. Quimiometria

O desenvolvimento tecnológico ao longo dos anos e a introdução de instrumentos de análises capazes de gerar uma grande quantidade de dados, tornou a extração de informações dos experimentos um desafio. A Quimiometria surgiu da necessidade de traduzir esses grandes conjuntos de dados, por meio de ferramentas matemáticas e estatísticas, em informações importantes (FERREIRA, 2015). A IUPAC

define Quimiometria como a aplicação da estatística para a análise de dados químicos (da química orgânica, analítica ou medicinal) e o planejamento de experimentos e simulações químicas (IUPAC, 2019). A análise de dados químicos na Quimiometria diz respeito à aplicação de algoritmos para o processamento de dados multivariados, permitindo manusear e melhorar respostas analíticas de modo que etapas de preparo de amostras podem ser poupadas (SAVELIEV et al., 2024). Esses algoritmos são formalismos matemáticos que podem extrair informações atribuídas ao conjunto de dados. Eles são métodos de reconhecimento de padrões, através dos quais podemos identificar tendências e agrupamentos, e serão empregados neste estudo.

A estrutura de um estudo quimiométrico pode ser dividida em pré-processamento e pré-tratamento dos dados, redução da dimensionalidade, emprego de métodos de reconhecimento de padrões (análise exploratória e classificação) e a tomada de decisão. Os métodos de pré-tratamento visam remover variações que podem levar a viés dentro do conjunto de dados. Os métodos de reconhecimento de padrão podem ser divididos em supervisionados e não supervisionados. A principal diferença entre eles está no conhecimento dos *outputs* alvos (valores de saída, classe das amostras) durante o aprendizado, que é necessária na construção dos modelos supervisionados, mas que não é utilizada nos modelos não supervisionados (GALVAN et al., 2023).

Os modelos não supervisionados podem ser divididos em agrupamento e redução de dimensão. Entre os principais métodos empregados na clusterização, podemos destacar o agrupamento hierárquico (HCA, do inglês, *Hierarchical Clustering Analysis*) e o *K-means*. Na redução de dimensão, temos como destaque a análise de componentes principais (PCA, do inglês, *Principal Component Analysis*) (CHI et al., 2024).

No que diz respeito aos modelos supervisionados, estes podem ser divididos em classificadores lineares e não lineares. Entre os principais algoritmos lineares de aprendizado de máquina estão a análise discriminante linear (LDA, do inglês, *Linear* Discriminant Analysis), regressão logística (LR, do inglês, *Logistic Regression*), a regressão por mínimos quadrados parciais (PLS, do inglês, *Partial Least Squares*), a análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês, *Partial Least Squares Discriminant Analysis*). Entre os classificadores não lineares, temos o K-vizinhos mais próximos (KNN, do inglês, *K-Nearest Neighbors*), árvores de decisão (DT, do inglês, *Decision Tree*), floresta aleatória (RF, do inglês, *Random Forest*),

máquina de vetores de suporte (SVM, do inglês, *Support Vector Machine*), entre outros (CHI et al., 2024).

Os algoritmos de aprendizado de máquina escolhidos em um estudo quimiométrico dependem das características dos dados experimentais, da composição da matriz de dados (número de amostras e variáveis), da demanda computacional exigida e do resultado que os pesquisadores visam obter. É comum ver trabalhos que utilizam mais de um algoritmo (HAYATI et al., 2024; OLIVEIRA et al., 2024) para encontrar aquele que melhor se adequa e explica seus dados. O presente trabalho selecionou a PCA como modelo não supervisionado para a análise exploratória, e a LR, a LDA, o SVM e a DT como modelos de classificação a serem investigados. Esses algoritmos serão discutidos nas próximas seções.

1.4.1. Pré-processamento e Pré-tratamento de Dados

Processamentos podem ser realizados tanto nas amostras (linhas) quanto nas variáveis (colunas) em uma matriz de dados. O objetivo, como descrito por Ferreira (2015), é minimizar variações indesejadas que não foram removidas durante a aquisição de dados e não serão eliminadas naturalmente durante a análise, influenciando negativamente os resultados. Desse modo, espera-se que as informações de interesse possam se destacar e resultar em uma melhor modelagem. Vale ressaltar que os tratamentos também podem levar à perda de informações importantes, distorcer e comprometer os resultados (MISHRA et al., 2020). Por isso, deve-se dar tanta atenção quanto nas demais etapas. Esta seção será dedicada aos métodos de pré-processamento e pré-tratamento aplicados à matriz de dados, uma vez que a etapa de processamento dos espectros foi abordada anteriormente.

A melhoria na relação sinal/ruído (S/R) é um dos resultados obtidos na etapa de pré-processamento, na qual métodos de suavização aplicados às amostras, como a média móvel ou o filtro de Savitzky-Golay, são responsáveis por tal feito (FERREIRA, 2015). No entanto, as alterações nos dados podem ser feitas de forma distorcida ou eliminar variáveis importantes. Portanto, os métodos de suavização devem ser empregados com cuidado e descartados quando observadas perdas de informações.

Fatores como falta de homogeneidade da amostra, diferenças na preparação e outras fontes de variação experimental estão atrelados a erros sistemáticos indesejados responsáveis por diferenças de concentração entre amostras (SUN; XIA, 2024). A normalização surge como método aplicado para remover variação indesejada

de amostra para amostra, com o objetivo de torná-las comparáveis entre si. Nesse processo, cada uma das variáveis da amostra é dividida por um fator de normalização.

Na normalização pelo comprimento do vetor, a norma de valor absoluto e a euclidiana, comumente usadas, são fatores de normalização que podem ser empregados. A norma de valor absoluto é obtida pela soma dos valores absolutos de todas as variáveis da amostra. Enquanto a norma euclidiana é dada pela raiz quadrada da soma do valor quadrado de todas as variáveis, e é a mais utilizada (FERREIRA, 2015).

A equação 6 expressa a normalização feita em cada linha da matriz, onde $\|x_i\|$ pode ser de norma 1 (Eq. 7) ou norma 2 (Eq. 8):

$$x_{ij} = \frac{x_{ij}}{\|x_i\|} \tag{6}$$

$$\|x_i\|_1 = \sum_{j=1}^{J} |x_{ij}| \tag{7}$$

$$\|x_i\|_2 = \sqrt{\sum_{j=1}^{J} x^2{}_{ij}} \tag{8}$$

Além dos problemas que podem interferir na análise de dados mencionados até agora, é comum haver variáveis que apresentam ordens de grandeza muito diferentes. A diferença entre as dimensões (escalas) dos dados em cada variável afeta diretamente o modelo. Dessa forma, as variáveis mais intensas sobressaem sobre as demais. Métodos de escalamento podem ser empregados para minimizar essas diferenças e tornar as variáveis comparáveis. O escalamento de variância unitária (Eq. 9), ou auto-escalamento, é frequentemente aplicado. Trata-se de um método de escalamento baseado na dispersão dos dados, em que as variáveis centradas na média são divididas por seus respectivos desvios-padrão (Eq. 10). Como resultado, os dados se tornam adimensionais, não variando mais com respeito à unidade original dos dados (CAMPOS; REIS, 2020; FERREIRA, 2015).

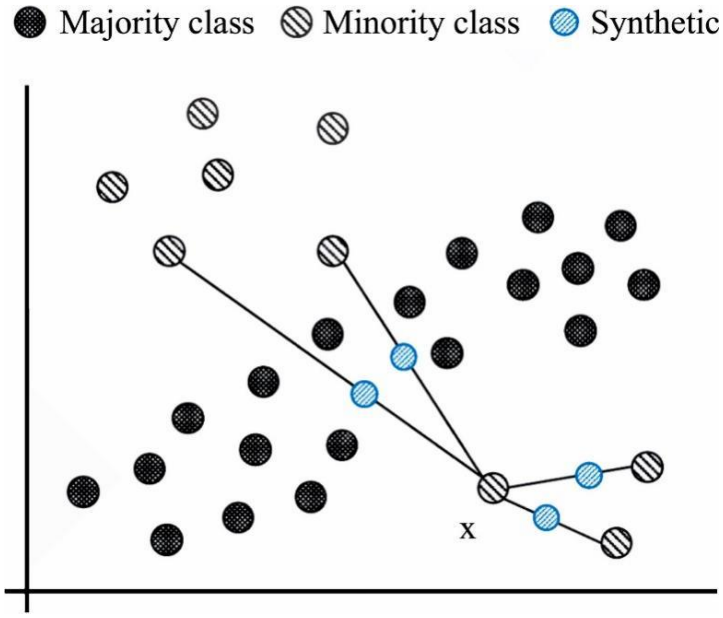
$$x_{ij} = \frac{x_{ij} - x_j}{s_j} \tag{9}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (10)$$

Apesar de as etapas de pré-processamento corrigirem ruídos e possíveis erros aleatórios inerentes às amostras e variáveis durante o processo de aquisição dos dados, existe outro desafio a ser enfrentado durante a modelagem: o desbalanceamento entre as classes de um conjunto de dados, que pode resultar em um modelo enviesado, com desempenho superior para as amostras da classe majoritária. Aumentar a representatividade da classe minoritária pode contribuir para a capacidade preditiva e a robustez do modelo, sendo considerado, nesses casos, um tipo de pré-processamento. Assim, o uso de técnicas de reamostragem pode equilibrar os conjuntos de treinamento, reduzindo o viés e balanceando a influência das classes.

Entre os algoritmos existentes para realizar a amostragem em um conjunto de dados, destaca-se a técnica de Sobreamostragem de Minoria Sintética (SMOTE, do inglês *Synthetic Minority Over-sampling Technique*). Em vez de simplesmente replicar as amostras da classe minoritária, o SMOTE gera novas amostras sintéticas a partir das existentes. Para isso, seleciona amostras que estão próximas umas das outras em um espaço de características e cria novas amostras ao longo do segmento que conecta pares de amostras reais. Durante o processo, o SMOTE seleciona uma amostra aleatória da classe minoritária, encontra seus *k* vizinhos mais próximos da mesma classe e gera uma amostra sintética em um ponto aleatório no segmento de reta que conecta a amostra original ao vizinho selecionado (Figura 10) (RODRIGUES; LUNA; PINTO, 2023).

Figura 10 – Geração de amostra sintética utilizando a técnica SMOTE.



Fonte: MORENO-BAREA et al., 2022.

As amostras sintéticas geradas pelo SMOTE tendem a ser menos dispersas do que aquelas pertencentes à distribuição original. Isso ocorre porque o processo de geração por interpolação linear posiciona essas amostras mais próximas do centro da distribuição da classe minoritária. A diferença entre as distribuições antes e depois da aplicação do SMOTE pode ser significativa, especialmente quando há um número reduzido de amostras minoritárias (ELREEDY; ATIYA, 2024). Por isso, o uso do SMOTE deve ser feito com cautela.

Por fim, o conjunto de dados pode apresentar uma alta dimensão, onde, o número de variáveis pode exceder muito o número de amostras. Esse perfil de dados pode afetar significativamente a precisão e estabilidade do modelo, resultando em dificuldades durante a modelagem. A redução da dimensionalidade pode ser utilizada como alternativa para melhorar o desempenho da modelagem nesses casos, isso pode ser feito por meio de algoritmos de seleção de variáveis ou até mesmo utilizando a PCA, como já relatado na literatura (HAYATI et al., 2024).

1.4.2. Redução da Dimensionalidade

1.4.2.1. Técnicas de Seleção de Variáveis

O desequilíbrio em conjuntos de dados, no que diz respeito a relação do número de variáveis e observações, é um dos grandes problemas enfrentados durante a modelagem por aprendizado de máquina. A presença de variáveis altamente

correlacionadas e que não apresentam relação alguma com a condição a ser investigada, além de formarem um espaço de alta dimensão, pode enganar o treinamento do algoritmo. A seleção de variáveis é uma etapa importante, e frequentemente empregada em aprendizado de máquina, que tem como objetivo principal selecionar um subconjunto de variáveis relevantes com base nos critérios de avaliação de suas importâncias. Como resultado, espera-se um melhor desempenho nos processos de aprendizado e classificação, além da redução no custo computacional e melhor interpretabilidade do modelo (LABORY; NJOMGUE-FOTSO; BOTTINI, 2024).

Os métodos de seleção de variáveis podem ser divididos em três categorias principais: métodos de filtro (do inglês, *filter methods*), métodos de empacotamento (do inglês*, wrapper methods*) e os métodos integrados (do inglês, *embedded methods*). Os métodos de filtro avaliam o desempenho estatístico dos dados de forma independente do modelo de aprendizado, combinando técnicas de ranqueamento com critérios estatísticos e utilizando ordenações para selecionar as variáveis. Todo o processo de filtragem das variáveis irrelevantes ocorre antes do início do processo de classificação e não envolve o algoritmo de aprendizado, diferente dos métodos de empacotamento (JIA et al., 2022; HULJANAH et al., 2019).

Os métodos de empacotamento avaliam subconjuntos de variáveis medindo sua influência sobre a exatidão do modelo durante o treinamento até encontrar as combinações mais adequadas, resultando em uma grande quantidade de cálculos e maior demanda computacional do que métodos de filtro, principalmente quando conjuntos de dados volumosos são utilizados (JIA et al., 2022; HULJANAH et al., 2019). Eles podem ser divididos entre determinísticos e meta-heurísticos. As duas últimas classes serão empregadas neste presente trabalho por meio dos algoritmos SFS (do inglês, *Sequential Forward Selection*) e GA (do inglês, *Genetic Algorithm*), respectivamente. Apesar do gasto computacional utilizando os métodos de empacotamento, eles apresentam melhor desempenho do que os de filtro (JIA et al., 2022).

Os métodos integrados avaliam as variáveis durante cada iteração do processo de treinamento do modelo, extraindo gradualmente as que mais contribuem em determinadas iterações. Assim como nos métodos empacotas, os métodos integrados buscam por um subconjunto ótimo de variáveis que estão incorporadas na construção do classificador e que tem grande impacto na exatidão durante o treinamento. O

método mais utilizado é o de regularização, que penaliza as variáveis menos relevantes atribuindo limites aos coeficientes. Os algoritmos SFM (do inglês, *Select From Model*) e LASSO (do inglês, *Least Absolute Shrinkage and Selection Operator*) fazem parte desse tipo de métodos (ALMARWI; AL-GAPHARI, 2024; HULJANAH et al., 2019).

O desempenho do algoritmo para seleção de variáveis depende da natureza dos dados, dos objetivos da aplicação e dos algoritmos utilizados no processo, este último, no caso dos métodos empacotados e integrados. No presente trabalho, três métodos foram avaliados, sendo eles o SFM, SFS e GA.

O algoritmo SFM realiza a seleção de variáveis com base na importância atribuída a cada uma delas por um estimador. Inicialmente, define-se um valor limite de importância, uma fronteira entre as variáveis que serão selecionadas e as que serão eliminadas. Em seguida, as variáveis com importância abaixo do limite são eliminadas. As demais são, então, utilizadas para o treinamento de um modelo de aprendizado de máquina (ELDAHSHAN; ALHABSHY; MOHAMMED, 2023). Esse método é comumente aplicado com estimadores baseados em árvores (como *Random Forest* ou *Extra Trees*), utilizando a importância baseada no índice de Gini, mas também pode ser usado com modelos lineares (OLIVEIRA et al., 2024).

O algoritmo SFS começa com nenhuma variável e adiciona uma a uma a cada iteração até que um critério ou o número desejado de características seja atendido. Como resultado, tem-se a solução ótima local entre as iterações ou a solução com o número desejado de variáveis no subconjunto. Basicamente, o algoritmo testa todas as variáveis uma por uma e escolhe a melhor; em seguida, combina essa variável com cada uma das outras, testando as combinações em pares e selecionando a melhor. O processo é repetido, adicionando uma nova variável à combinação anterior. Por fim, o resultado final é a melhor solução entre essas iterações: o conjunto de variáveis que, passo a passo, mais contribuiu para o desempenho do modelo (ALMAGHTHAWI; AHMAD; ALSAADI, 2022).

O GA é um algoritmo de busca meta-heurística adaptativa, baseado em mecanismos da evolução natural, usa operações genéticas até atingir uma solução ótima. Ele avalia um conjunto de soluções candidatas, chamadas cromossomos, e simula o princípio da sobrevivência do mais apto, conforme a teoria de Darwin. Os cromossomos são definidos como uma sequência de conjunto de variáveis. (HABIB; VICENTE-PALACIOS; GARCÍA-SÁNCHEZ, 2025).

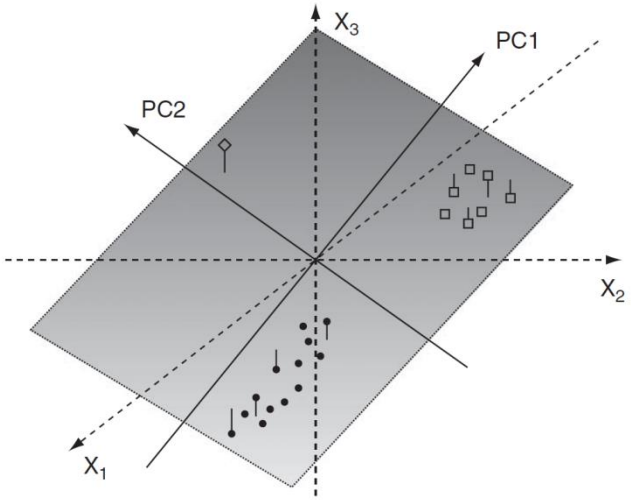
Segundo Chiesa et al. (2020), o Algoritmo Genético é composto por cinco etapas principais: inicia-se com a criação de uma população de soluções aleatórias, seguida pela avaliação de cada uma com base em uma função de aptidão. Em seguida, são selecionadas as melhores soluções, que passam por cruzamentos e mutações aleatórias com baixa probabilidade. Essas três últimas fases compõem o processo evolutivo, por meio do qual se obtém uma nova geração de soluções mais aptas.

1.4.2.2. Análise de Componentes Principais

A análise de componentes principais (PCA) é uma técnica de análise estatística multivariada amplamente utilizada para análise exploratória de dados, detecção de amostras anômalas (*outliers*), redução de dimensionalidade, entre outras aplicações. Classificada como um algoritmo de transformação linear não supervisionado, a PCA projeta os dados de alta dimensionalidade em um novo sistema de coordenadas, onde as novas variáveis, denominadas componentes principais (PCs), são combinações lineares das variáveis originais (ANOWAR et al., 2021). Considere uma matriz X de dimensão $m \times n$, onde as linhas representam amostras ou objetos e as colunas representam variáveis ou características. A PCA decompõe essa matriz por meio da combinação linear dessas variáveis, reduzindo a dimensionalidade dos dados e explicando a variação presente.

As componentes principais (PCs) têm como objetivo explicar a máxima variância dos dados. A primeira componente principal (PC1) explica a maior parte da variação dos dados. A segunda componente principal (PC2) é ortogonal à PC1 e captura a variância que a primeira não explicou. O mesmo ocorre com a terceira componente principal (PC3) em relação à PC2, e assim por diante, até que toda a variação nos dados seja explicada (LEE; JEMAIN, 2021). A Figura 11 traz uma representação gráfica da projeção das amostras (ou objetos) por PC1 e PC2.

Figura 11 – Projeção dos objetos no plano formado pelas duas primeiras PCs.



Fonte: ESBENSEN; GELADI, 2009.

A decomposição de uma matriz X utilizando combinação linear, geralmente centrada na média ou autoescalada, com variância máxima pode ser expressa pela Eq. (11):

$$\mathbf{X = TP^T + E} \tag{11}$$

Na qual, **T** é a matriz de escores, **P** é a matriz ortonormal de pesos e **E** é a matriz residual (LEE; JEMAIN, 2021). Os escores carregam informações sobre as amostras, enquanto os pesos (ou *loadings*) representam as contribuições das variáveis para cada componente principal (FERREIRA, 2015). A matriz residual pode ser definida como as diferenças entre as variáveis originais e as reconstruídas (GARCIA-ALVAREZ et al., 2023).

A decomposição em valores singulares (SVD, do inglês, *Singular Value Decomposition*) e o NIPALS (do inglês, *Nonlinear Iterative Partial Least Squares*) são os algoritmos mais utilizados na PCA. Ambos buscam autovetores e autovalores da matriz de covariância de X ou da matriz de correlação, se a matriz tiver sido autoescalada (LEE; JEMAIN, 2021).

O cálculo das matrizes T e P utilizando o SVD pode ser descrito pela Eq. 12:

$$\mathbf{X=USV^T + E} \tag{12}$$

Onde, **X** e **E** são os mesmos da Eq. (11). O produto **US** representa a matriz de escores **T**, onde **U** é a matriz ortonormal dos escores e **S** é a matriz diagonal com os valores singulares. A matriz **V** é a matriz ortonormal de pesos **P**. A matriz $\mathbf{SS^T}$ é uma

matriz diagonal que contém os valores singulares ao quadrado, que são equivalentes à matriz diagonal de autovalores $\mathbf{\Lambda}$. Os autovalores (λ) estão ordenados em ordem decrescente e correspondem à variância explicada (ESBENSEN; GELADI, 2009; GARCIA-ALVAREZ et al., 2023).

A PCA permite a visualização e interpretação dos dados de maneira mais aprofundada do que quando analisadas apenas variáveis individuais. Portanto, é quase sempre a primeira análise realizada na análise estatística multivariada (ESBENSEN; GELADI, 2009). Em aprendizado de máquina, é frequentemente utilizada como um passo de pré-processamento (HAYATI et al., 2024).

1.4.3. Modelos de Classificação

Os modelos de classificação são compostos por algoritmos de aprendizado supervisionado. O objetivo é treinar os modelos com base nos valores de saída para guiar o processo de aprendizado. Como resultado, o modelo treinado é capaz de prever dados novos e não vistos. Os algoritmos de aprendizado de máquina supervisionados são amplamente utilizados em estudos metabolômicos e metabonômicos (YANG et al., 2017; CHI et al., 2021; SUN et al., 2024). Durante o estudo, uma série de modelos pode ser testada, desde os mais simples até os mais complexos, e a escolha dependerá do seu desempenho.

O presente estudo investigou algoritmos de aprendizado paramétricos e não paramétricos. Modelos paramétricos assumem que os dados seguem uma distribuição de probabilidade específica e que a função que os descreve possui um número fixo de parâmetros, o que proporciona uma interpretação mais direta dos resultados. Por outro lado, modelos não paramétricos não fazem essas suposições, e o número de parâmetros não é fixo, podendo aumentar conforme o volume de dados cresce, tornando o modelo mais flexível para representar dados complexos. Contudo, esses modelos tendem a ser mais difíceis de interpretar e apresentam maior custo computacional e tendem a sobreajustes com maior frequência (IMAM, MUSILEK, REFORMAT, 2024)

1.4.3.1. Regressão Logística

A regressão logística (LR) é uma forma de regressão generalizada que permite modelar a probabilidade de um resultado binário. Ela usa a função logística (ou sigmoide) para transformar a combinação linear das variáveis independentes em uma

probabilidade que varia entre 0 e 1, adequada para dados binários. A LR pode analisar tanto problemas de regressão como de classificação (ZHENG et al., 2024). O modelo padrão de regressão logística, é descrito pela equação 13:

$$\log(odds) = ln\left[\frac{P(Y)}{1-P(Y)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$\frac{P(Y)}{1-P(Y)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}$$

$$P(Y) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} - P(Y)e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}$$

$$P(Y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 - e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \quad (14)$$

Os termos $x_{1,i}$, $x_{2,i}$ ... $x_{p,i}$ são as variáveis preditoras. Y_i é uma variável aleatória de Bernoulli que representa o resultado para o indivíduo *i*. O resultado é geralmente codificado numericamente como 1 (caso) ou 0 (controle). Portanto, $P(Y_i = 1)$ é a probabilidade de o fenótipo ser igual a um caso (DUMANCAS et al., 2015). Assim, a regressão logística relaciona a probabilidade Y às variáveis preditoras (Eq. 14).

Os coeficientes de regressão refletem o efeito de cada variável independente na capacidade preditiva do modelo. O objetivo é estimar os parâmetros β desconhecidos responsáveis por um hiperplano que possa separar e classificar claramente todas as amostras (PAN et al., 2024). A estimativa de máxima verossimilhança é então utilizada para encontrar o conjunto de parâmetros para os quais a probabilidade dos dados observados é maior (BOATENG; ABAYE, 2019).

A LR é considerada uma técnica estatística mais flexível do que outras, porém apresenta problemas quando as variáveis estão altamente correlacionadas, ou seja, quando carregam a mesma informação, resultando em problemas de multicolinearidade (FERNANDES et al., 2019). Para lidar com esse problema comum em conjuntos de dados, pode-se fazer uso de ferramentas para reduzir a dimensionalidade dos dados, por exemplo.

Estudos metabonômicos ou metabolômicos são comumente encontrados na literatura empregando a Regressão Logística (LR) como modelo de diagnóstico. Foi o caso de Yang e colaboradores (2017), que, utilizando alguns algoritmos de aprendizado de máquina, chegaram a um modelo de diagnóstico com a LR baseada em três metabólitos, com precisão de 94,64%. Já no trabalho de Sun e colaboradores

(2024), mesmo selecionando 12 metabólitos para a LR e obtendo uma excelente capacidade preditiva com a validação do conjunto de treinamento, os resultados foram insatisfatórios no conjunto de validação, apresentando uma baixa capacidade preditiva.

Considerando a quantidade de algoritmos de aprendizado de máquina supervisionados que podem ser empregados e as características particulares de cada conjunto de dados, um único estudo pode empregar vários algoritmos.

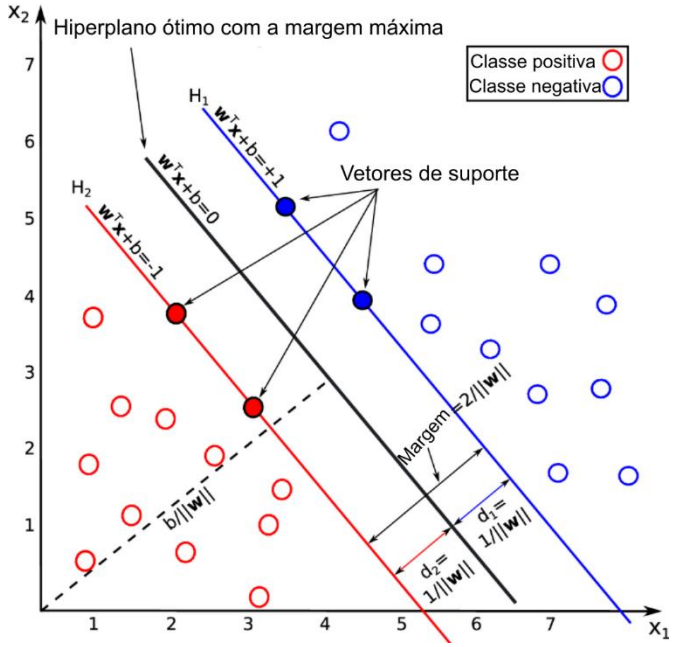
1.4.3.2. Máquina de Vetores de Suporte

A máquina de vetores de suporte (SVM) é um algoritmo que busca encontrar um hiperplano em um espaço n-dimensional, onde n é o número de covariáveis que separa as amostras em suas classes. Por exemplo, dados constituídos de respostas binárias (y = {-1,1}) linearmente separáveis podem ser separados por um vetor de peso (w) encontrado pelo SVM (DEBIK et al., 2021). Nesse caso, o hiperplano que separa as classes é descrito pela equação 15:

$$\mathbf{w^T x + b = 0} \quad (15)$$

Sendo que b é um termo escalar que ajusta a posição do hiperplano, sem alterar sua orientação, para que a margem seja maximizada (DEBIK et al., 2021). Essa margem se refere à distância das amostras de um grupo mais próximas das amostras do outro grupo. As amostras que melhor definem a margem são os vetores de suporte.

Figura 12 – Exemplo de um problema de classificação binária com dados lineares separáveis usando SVM.



Fonte: Adaptado de THARWAT (2019).

O espaço divido pelo hiperplano resulta em dois meios espaços: um positivo, onde as amostras da classe positiva se encontram, e um negativo, onde as amostras da classe negativa estão. Dois planos são construídos para cada classe e podem ser definidos por:

$$\mathbf{w^T x + b \geq 1},\ \text{para a classe positiva;}$$

$$\mathbf{w^T x + b \leq 1},\ \text{para a classe negativa.}$$

A margem pode ser descrita pela distância desses planos até o hiperplano e calculada como $\frac{2}{\|w\|}$ (THARWAT, 2019).

A maximização da margem permite classificar corretamente novos dados que estejam dentro dessa margem, de cada lado do hiperplano de classificação, uma característica única do SVM.

A classificação do SVM pode ser linear ou não linear. Nos casos em que os dados não são separáveis linearmente, as funções de kernel ou truques de kernel podem ser usadas para transformar os dados em espaços de recursos de dimensão maior. O primeiro passo é definir a função de kernel adequada (ROY; CHAKRABORTY, 2023). No caso do SVM com kernel linear, o parâmetro de regularização, denominado C, é um hiperparâmetro de ajuste que permite flexibilizar as classificações erradas feitas

pela margem. Com grandes valores de C, o modelo tenta classificar todos os pontos de treinamento corretamente, mesmo que uma margem pequena seja escolhida para o hiperplano. Já com pequenos valores de C, o modelo será otimizado para uma margem maior separando o hiperplano, mesmo que haja mais classificações incorretas. (MENDEZ; REINKE; BROADHURST, 2019).

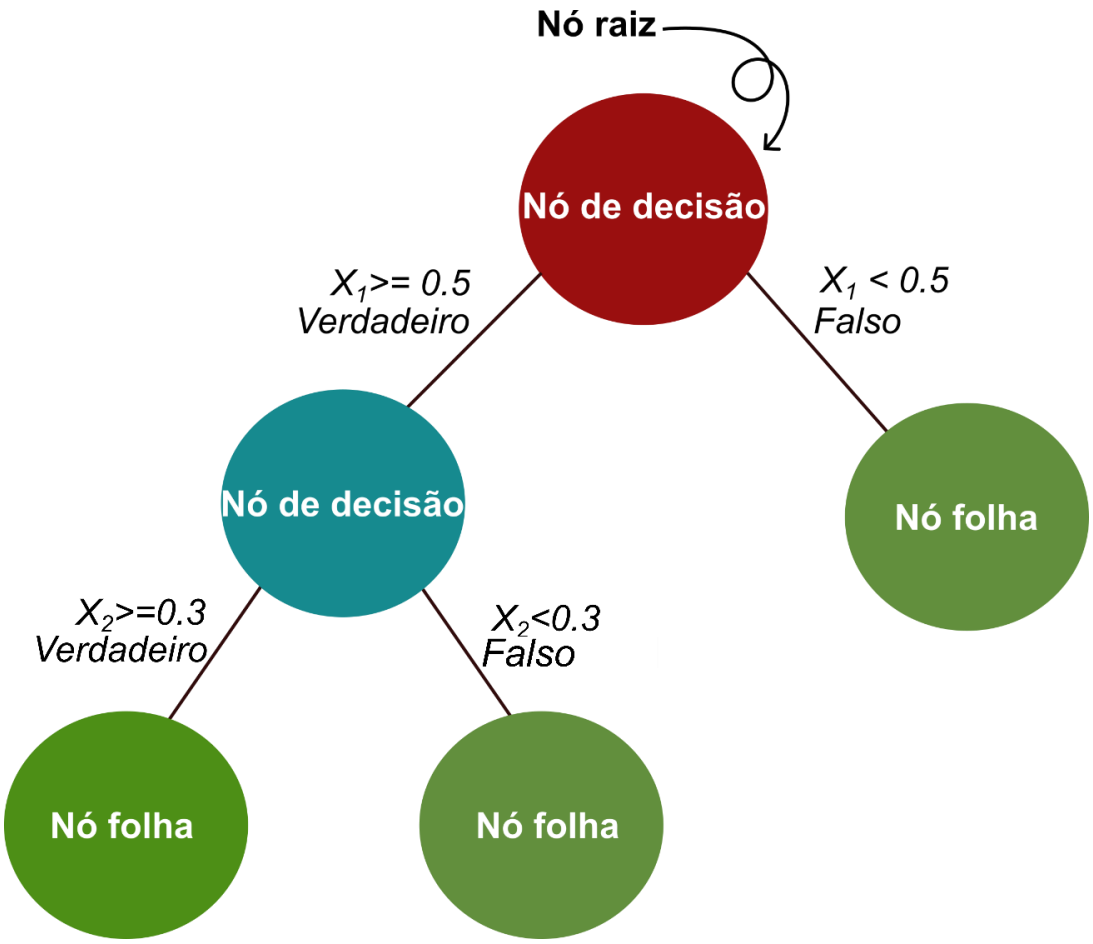
Estudos metabolômicos e metabonômicos descritos na literatura vêm, ao longo dos anos, investigando o potencial do SVM no diagnóstico de doenças, empregando diferentes kernels e obtendo resultados satisfatórios, com modelos de boa capacidade preditiva (ZHENG et al., 2017; WANG et al., 2023). Uma das principais vantagens do SVM como algoritmo de aprendizado de máquina supervisionado na estratégia metabonômica é sua capacidade de lidar de forma robusta com a multicolinearidade entre as variáveis preditoras.

1.4.3.3. Árvore de Decisão

Algoritmos baseados em árvores fazem parte do conjunto de métodos supervisionados, os quais dividem o espaço das variáveis em um conjunto de regras de decisão hierárquicas. Esses modelos podem ser empregados tanto em tarefas de classificação quanto de regressão, sendo mais comuns na resolução de problemas de classificação (LIU; XUAN; XIAO, 2025).

Entre esses algoritmos, destaca-se a Árvore de Decisão (DT, do inglês *Decision Tree*), que recebe esse nome devido à sua estrutura ramificada, semelhante a uma árvore invertida. Essa estrutura é composta por um nó raiz (ponto inicial), nós de decisão (pontos onde são avaliadas variáveis preditoras), galhos (regras ou condições de decisão) e nós folha (resultados ou classes finais). A DT realiza sucessivas divisões com base nas variáveis, a partir do nó raiz, que corresponde à variável mais discriminativa, até que cada caminho leve a um nó folha, onde a classificação é definida com base nas categorias do problema (BANSAL; GOYAL; CHOUDHARY, 2022). A Figura 13 apresenta um modelo de DT binária. O processo inicia no nó raiz, onde a variável X_1 é avaliada e segue a condição $X_1 \geq 0.5$.

Figura 13 - Modelo genérico de uma DT.



Fonte: A autora (2025).

A divisão dos nós e a construção da árvore são realizadas com base no ganho de informação, que mede a quantidade de informação que uma variável fornece sobre a classe. O algoritmo DT escolhe para dividir o nó inicial a variável que maximiza esse ganho (BANSAL; GOYAL; CHOUDHARY, 2022).

O ganho de informação (GI) é baseado na entropia, ou melhor, mede a redução esperada dessa medida, a qual expressa a impureza dos dados. Assim, o GI é definido como a diferença entre a entropia do conjunto original e a entropia condicional após a divisão com base na característica selecionada (TANGIRALA, 2020). A entropia é calculada como a soma, para cada classe, da probabilidade da classe multiplicada pelo seu logaritmo em base 2, com sinal negativo. Considerando um conjunto de dados L e uma variável x com V valores distintos, seja $|L^v|$ o subconjunto de L em que x=v. A entropia de L e o ganho de informação de x são dados por:

$$Entropia\ (L) = -\sum_{i=1}^{j} p_i \log_2(p_i) \tag{16}$$

$$GI(L,x) = Entropia\ (L)\ - \sum_{v=1}^{V} \frac{|L^V|}{|L|} Entropia\ (L^V) \quad (17)$$

O critério de impureza de Gini e a profundidade máxima costumam ser usados para evitar sobreajustes (HAQUE; ISLAM; ERFAN, 2025). O índice de Gini determina a impureza ou pureza de uma classe durante a criação da DT, após a divisão ao longo de uma variável específica (TANGIRALA, 2020). O cálculo do índice de Gini pode ser realizado usando a expressão:

$$GINI(L) = 1 - \sum_{i=1}^{j} \mathrm{p}_i{}^2 \quad (18)$$

Onde, j é o número total de classes, p_i é a proporção de observações no nó que pertencem a classe i.

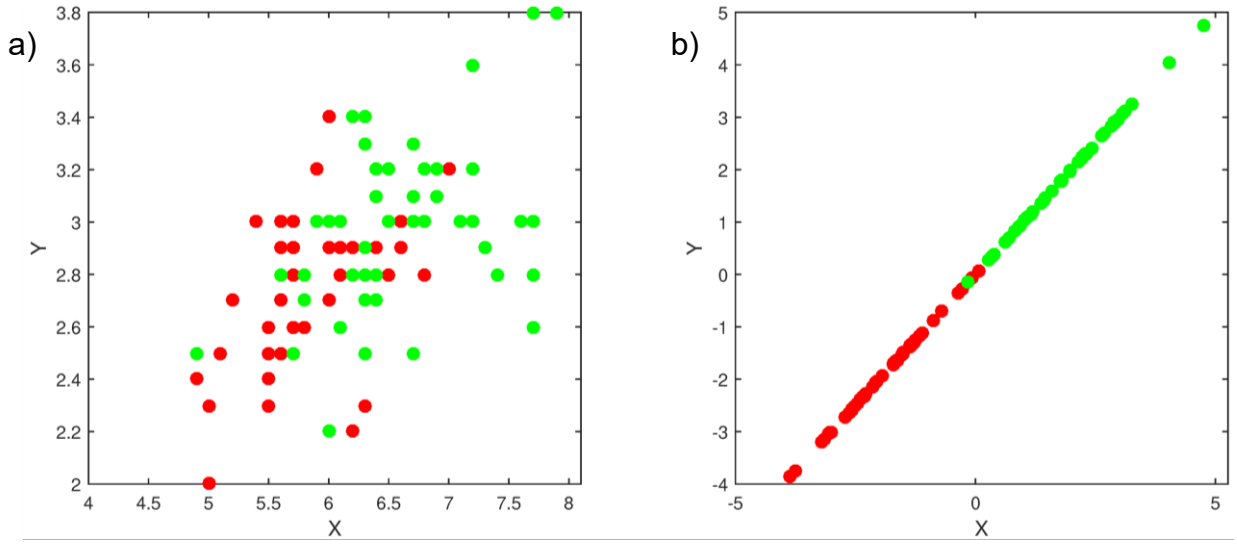
As DT são amplamente utilizadas em tarefas de aprendizado de máquina, isso se deve ao fato de serem intuitivas e interpretáveis (HAQUE; ISLAM; ERFAN, 2025). Por serem algoritmos não paramétricos, funciona bem com dados não linearmente separáveis.

Além disso, há métodos baseados em DT, como a Floresta Aleatória (RF, do inglês *Random Forest*). O algoritmo RF constrói diversas árvores de decisão a partir de diferentes subconjuntos do conjunto de dados e combina seus resultados para melhorar a precisão das previsões. Em problemas de classificação, o resultado final é determinado pela maioria dos votos das árvores, enquanto, em regressão, utiliza-se a média das previsões (AHMAD et al., 2022).

1.4.3.4. Análise Discriminante Linear

A análise discriminante linear (LDA) é um algoritmo de aprendizado de máquina que tem como objetivo encontrar uma combinação linear de variáveis independentes que maximize a separação entre diferentes classes e minimize a variabilidade dentro das classes (QU; PEI, 2024).

Figura 14 – Exemplo da distribuição de um conjunto de dados composto por duas classes. a) Antes da LDA e b) Depois da LDA.



Fonte: (QU; PEI, 2024)

A LDA é baseada no teorema de Bayes, que calcula a probabilidade posterior de uma amostra pertencer a uma determinada classe. Ela foi desenvolvida para variáveis independentes com distribuição normal e parte do pressuposto de que as classes possuem matrizes de covariância comuns, facilitando o cálculo das probabilidades a posteriori e permitindo a implementação do algoritmo de forma eficiente (BOEDEKER; KEARNS, 2019; GRAF; ZELDOVICH; FRIEDRICH, 2024).

A função discriminante linear (Eq. 19) pode ser expressa como:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \tag{19}$$

onde x é o vetor de entrada das variáveis, w é o vetor dos coeficientes lineares e w_0 é um termo de enviesamento (intercepto da função). Esses dois últimos termos são calculados com base na média e na matriz de covariância de cada classe nos dados de treinamento. Embora LDA e Regressão Logística (LR) se assemelhem ao trabalharem com funções de probabilidade, a principal diferença está na estimativa dos coeficientes (GRAF; ZELDOVICH; FRIEDRICH, 2024).

A LDA possui algumas desvantagens, entre elas, a falta de robustez frente a ruídos e *outliers*, e a singularidade da matriz de dispersão dentro da classe, está última pode ser causada por um número insuficiente de amostras e variáveis altamente correlacionadas (QU; PEI, 2024). Apesar dessas limitações, trabalhos como o de Hayati e colaboradores (2024) mostram que, mesmo em dados de alta dimensionalidade, utilizando técnicas de pré-processamento, é possível obter

modelos LDA com até 87% de precisão. Com a aplicação de técnicas para reduzir a dimensionalidade dos dados, como a PCA, o modelo alcançou uma precisão de 100%.

1.5. Validação e figuras de mérito

A etapa de validação é essencial para garantir a robustez e a confiabilidade do modelo, especialmente se a intenção for introduzir essa abordagem na rotina clínica. Neste estudo, abordaremos seis métricas, ou figuras de mérito, que fornecem uma visão geral do desempenho do modelo: AUROC, sensibilidade, especificidade, precisão, exatidão e *F1-Score*.

A curva ROC (do inglês, *Receiver Operating Characteristic*) resume a capacidade do modelo em classificar corretamente as amostras às suas classes. A curva é plotada em função da taxa de verdadeiros positivos (TPR, do inglês, *True Positive Rate*) e da taxa de falsos positivos (FPR, do inglês, *False Positive Rate*). A área sob a curva (AUROC, do inglês*, Area Under Receiver Operating Characteristic*) pode ser vista como a probabilidade de um paciente selecionado aleatoriamente seja classificado realmente como um paciente do que como um controle (STOJANOV et al., 2023). A TPR do modelo é a sua sensibilidade e a FPR é numericamente igual a 1 – especificidade, no qual a especificidade é a taxa de verdadeiros negativos (POLO; MIOT, 2020).

Todas essas figuras de méritos são calculadas com base na matriz de contingência, utilizando os valores de VP, FN, FP e VN (KAWAMURA, 2002). Um exemplo desse tipo de matriz é apresentado na Tabela 1.

Tabela 1 – Matriz de contingência genérica.

		Diagnóstico Padrão	
		Positivo	**Negativo**
Teste	**Positivo**	VP	FP
	Negativo	FN	VN

Fonte: Adaptado de KAWAMURA (2002).

Onde,

Verdadeiros positivos (VP) – número de pacientes previstos corretamente como positivo.

Falsos negativos (FN) – número de pacientes previstos erroneamente como controle.

Falsos positivos (FP) – número de controle previstos erroneamente como pacientes.

Verdadeiros negativos (VN) – número de controle previstos corretamente como controle.

A sensibilidade e a especificidade são calculadas com base as equações 20 e 21:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{20}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{21}$$

O Valor Preditivo Positivo (VPP) é a figura de mérito que fornece uma estimativa da capacidade preditiva do modelo voltada para a classe positiva, ou seja, a probabilidade de classificar um indivíduo doente como positivo (STOJANOV et al., 2023). Desse modo, pode ser calculada da seguinte maneira (Eq. 22):

$$VPP = \frac{VP}{VP + FP} \tag{22}$$

O Valor Preditivo Negativo (VPN) é a probabilidade de classificar um indivíduo do grupo controle como negativo, e pode ser calculado pela Eq. (23).

$$VPN = \frac{VN}{VN + FN} \tag{23}$$

A exatidão fornece uma avaliação mais ampla do modelo, estimando o número de previsões corretas, tanto positivas quanto negativas. Sua fórmula (Eq. 24) é definida como:

$$Exatidão = \frac{VP + VN}{VP + FP + VN + FN} \tag{24}$$

O *F1-Score* é uma figura de mérito que equilibra precisão e sensibilidade (Eq. 25). Isso é importante porque uma alta precisão pode mascarar um desempenho ruim na identificação de casos positivos, que reflete em uma baixa sensibilidade (LABORY; NJOMGUE-FOTSO; BOTTINI, 2024). Portanto, quanto mais próximo de 1 for o valor do *F1-Score*, mais robusto é o modelo.

(25)

$$F1 - score = 2 \times \frac{VPP \times Sensibilidade}{VPP\ + Sensibilidade}$$

Com a disponibilidade dos dados de predição e dos rótulos verdadeiros, as figuras de mérito são calculadas com base nessas predições, fornecendo uma avaliação do desempenho dos modelos nos dados fornecidos.

O coeficiente Kappa de Cohen (1960) é utilizado para avaliar a concordância entre dois diferentes métodos classificadores. O coeficiente é calculado relacionando a proporção de concordância entre os classificadores (p_o) com a proporção de concordância ao acaso entre os classificadores (p_e) (VACH; GERKE, 2023):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{26}$$

Onde,

$p_o = \frac{VP + VN}{N}$

$p_e = \left[\left(\frac{VP+FP}{N}\right) \times \left(\frac{VP+FN}{N}\right)\right] + \left[\left(\frac{VN+FN}{N}\right) \times \left(\frac{VN+FP}{N}\right)\right]$, e N representa o número de observações.

CAPÍTULO 2

2. Ensaios Metabonômicos para o Monitoramento de Doenças Renais

Doenças renais afetam o desempenho e a capacidade do rim de filtrar o excesso de água e resíduos do sangue, resultando em consequências sistêmicas que tornam diagnósticos e tratamentos difíceis e custosos. Entre algumas condições comuns estão a nefropatia diabética, lesão renal aguda e doenças renais policísticas, que podem progredir para a doença renal crônica (ABBISS; MAKER; TRENGOVE, 2019). Diabetes e hipertensão são as principais causas da doença renal crônica, que afeta cerca de 850 milhões de pessoas em todo o mundo (ISN, 2023). Essas condições estão associadas à maioria dos diagnósticos de insuficiência renal, uma condição em que a TFG é menor que 60 ml/min/1,73m² ou inferior a 15% do normal por 3 meses ou mais (AGUIAR et al., 2020). Outras condições, como lúpus e outras doenças imunológicas, cálculos renais, próstata aumentada e infecções repetidas do trato urinário, também podem causar danos aos rins a longo prazo, levando à insuficiência renal e, em estágios mais avançados, à doença renal terminal (ERNSTMEYER; CHRISTMAN, 2024).

Os tratamentos substitutivos, nos casos de doença renal terminal, compreendem diálise ou transplante. Apesar do transplante apresentar melhores resultados, os pacientes ainda possuem expectativa de vida reduzida em comparação com a população em geral, além de complicações relacionadas à infecção, rejeição e malignidade (TESFAYE et al., 2024). A doença renal crônica em estágio avançado pode levar à insuficiência renal e, consequentemente, à doença renal terminal. Trata-se de um problema de saúde pública que impõe um grande fardo aos pacientes, aos sistemas de saúde e à sociedade como um todo.

O desenvolvimento de pesquisas voltadas para diagnóstico precoce, tratamentos, mecanismos envolvendo a doença, entre outras investigações, fazem parte de um conjunto de medidas que podem ser adotadas para contribuir com profissionais de saúde no enfretamento do problema e tratamento. A abordagem metabonômica tem se destacado como uma alternativa promissora no estudo e diagnóstico de diversas doenças, incluindo as renais (WANG et al., 2021). Esta abordagem visa investigar metabólitos e mecanismos bioquímicos associados às alterações nos perfis metabólicos devido à presença da Lesão Renal Aguda e Nefrite Lúpica, podendo assim auxiliar no diagnóstico precoce, estadiamento e prognóstico dessas doenças (SILVA et al., 2018; XU et al., 2023; FONSECA et al., 2023). Estudos

metabonômicos utilizam a análise de biofluidos (sangue, urina, fezes, entre outros) e apresentam a forma de amostragem minimamente invasiva como grande vantagem.

Como discutido no Capítulo 1, a espectroscopia de ressonância magnética nuclear (RMN) está entre as principais técnicas analíticas empregadas em estudos metabonômicos. Sendo assim, considerando as vantagens do uso da espectroscopia de RMN, a estratégia metabonômica e o rápido avanço tecnológico que tende cada vez mais ao uso de técnicas de aprendizado de máquina, o presente capítulo busca demonstrar que o emprego de algoritmos de aprendizado de máquina pode efetivamente extrair e utilizar informações dos espectros de RMN de ^{1}H para classificar com precisão pacientes com doenças renais. Aqui, foram desenvolvidos modelos metabonômicos, em dois estudos: (1) diagnostico de Lesão Renal Aguda em recém-nascidos; e (2) estadiamento de Nefrite Lúpica, proliferativa com e sem lesão membranosa, em adultos.

2.1. Lesão Renal Aguda (LRA)

Classificada como uma síndrome, a Lesão Renal Aguda (LRA) é a perda aguda da função renal que pode apresentar formas leves até as mais graves, como perda completa da função renal, necessitando de TRS. Atrelados a LRA estão o aumento da morbidade, a mortalidade de curto e longo prazo e altos custos com tratamentos (STRAUSS et al., 2024).

No que diz respeito aos fatores de riscos associados ao desenvolvimento da LRA, a DRC é um deles, no qual, adultos com uma Taxa de Filtração Glomerular estimada (TFGe) menor que 60 ml/min/1,73m² estão particularmente em risco. Além disso, a DRC também é reconhecida como sequela da LRA. Outras condições como doença cardíaca, hipertensão, diabetes, idade igual ou superior a 65 anos, malignidade, sepse e pacientes cirúrgicos estão entre os principais fatores de risco (MOHAMED; MARTIN, 2024).

Trata-se de uma síndrome complexa que ocorre em diversos aspectos e manifestações clínicas, na qual, a Kidney Disease: Improving Global Outcomes (KDIGO) orienta que profissionais de saúde usem a definição de LRA da AKI Network (MEHTA et al., 2007) como estratégia de diagnóstico (KHWAJA, 2012). Os critérios adotados pela AKI são:

- Redução abrupta na função renal, atualmente definida como um aumento maior ou igual a 0,3 mg/dl (≥ 26,5 mol/l) da creatinina sérica, em 48 horas.
- Aumento da creatinina sérica para valores maiores ou iguais a 1,5 vezes o valor de referência, nos últimos sete dias.
- Redução na produção de urina (volume ≤ 0,5 ml/kg/h durante 6 h).

A KDIGO ainda sugere que o estadiamento da doença seja realizado de acordo com a gravidade, considerando três estágios. A Tabela 2 traz a definição da LRA em pacientes pediátricos. É importante destacar a necessidade de estadiamento visto que o risco de morte aumenta a cada estágio.

Tabela 2 - Estadiamento de LRA em pediátricos proposto pela KDIGO.

Estágio	Creatinina sérica (CrS)	Produção de urina
1	≥0,3 mg/dL em 48 horas ou aumento de ≥50% em 7 dias	>0,5 e ≤ 1 mL/kg/h
2	Aumento da creatinina ≥100%	>0,3 e ≤ 0,5 mL/kg/h
3	Aumento da creatinina ≥200% ou CrS ≥4 mg/dL ou recebimento de diálise ou eGFR≤35 mL/min/1,73 m² (neonatal cut-off: CrS > 2.5 mg/dL)	≤ 0,3 mL/kg/h

Fonte: (KHWAJA, 2012)

O diagnóstico de LRA, pelos critérios KDIGO, deve ser interpretado em conjunto com a condição clínica do paciente, uma vez que a creatinina sérica e a produção de urina também podem ser afetadas por fatores não renais. A possibilidade de os sintomas da disfunção renal começarem antes mesmo do indivíduo buscar ajuda médica, aliada ao tempo necessário para definir com precisão a presença da síndrome, pode resultar no diagnóstico tardio (PICKERS et al., 2021). Levando em consideração os fatores citados acima e a alta taxa de LRA em pacientes em unidades de terapia intensiva, a busca por métodos de diagnóstico precoce e os mecanismos associados ao tratamento da LRA são desafios atuais (XU et al., 2023).

A metabonômica surge como uma alternativa a ser utilizada na investigação de diversas doenças, incluindo a LRA, seja voltada para o diagnóstico, estadiamento ou prognóstico. Apesar da maior incidência de estudos metabonômicos abordando a DRC, a LRA também vem sendo investigada. São encontrados trabalhos envolvendo

diferentes faixa-etárias, desde adultos (XU et al., 2023), crianças (FRANIEK et al., 2022) e até neonatais prematuros (MERCIER et al., 2017).

Assim como nos estudos encontrados na literatura, o presente trabalho buscou desenvolver uma abordagem que possa auxiliar no diagnóstico e estadiamento voltado para doenças renais utilizando técnicas de aprendizado de máquina. A tecnologia ômica foi empregada à análise global de metabólitos urinários e séricos com o intuito de encontrar ferramentas que possam discriminar pacientes acometidos pela Lesão Renal Aguda e Nefrite Lúpica, e os principais metabólitos responsáveis capazes de auxiliar na tomada de decisão.

2.2. Nefrite Lúpica

O Lúpus Eritematoso Sistêmico (LES) é uma doença autoimune crônica, complexa, que provoca lesão inflamatória e de origem imunológica em diversos sistemas, entre eles o musculoesquelético, nervoso e renal (KHAROUF et al., 2025). Essa condição é caracterizada pela ação de anticorpos autorreativos contra antígenos nucleares, formando complexos imunes que se depositam em diferentes órgãos, desencadeando processos inflamatórios locais (SIMON et al., 2025). Quando a deposição desses complexos ocorre nos glomérulos dos rins, resulta na Nefrite Lúpica (NL), que provoca lesões de diferentes categorias histológicas (IKEUCHI et al., 2025). A NL afeta de 30% a 50% dos pacientes com LES, constitui uma de suas manifestações mais graves, e até 20% dos pacientes com NL podem chegar a desenvolver uma doença renal terminal (DRT) em 10 anos de diagnóstico de LES (SIMON et al., 2025; ANDERS et al., 2020).

A NL é uma forma de glomerulonefrite e suas categorias histológicas são divididas em seis classes distintas, que estão relacionadas ao grau de envolvimento renal, sendo a biópsia seu padrão ouro de diagnóstico (SAMY et al., 2025). A classificação foi atualizada pela Sociedade Internacional de Nefrologia/Sociedade de Patologia Renal (ISN/RPS) (BAJEMA et al., 2018).

Na classe I, a NL apresenta gromérulos normais na microscopia ótica, mas a imunoflorescência mostra pequenos depósitos imunocomplexos no interior dos glomérulos, que são unidades funcionais dos rins. Na classe II, esses depósitos causam aumento do número de células mesangiais (hiperplasia mesangial) e matriz mesangial. Nas classes III e IV, chamadas de NL proliferativa focal e difusa, respectivamente, esses depósitos aparecem em várias regiões do glomérulo, e a

diferença entre elas está na quantidade de glomérulos afetados: menos de 50% na classe III e 50% ou mais na classe IV. Na classe V, os depósitos aparecem no espessamento difuso da membrana basal glomerular, levando a um tipo de lesão chamada membranosa (MUSA; ROUT; QURIE, 2025).

Entre as principais características associadas às classes, as III e IV apresentam as maiores probabilidades de ter lesões ativas e o maior risco de progressão para a DRC. Juntamente com a classe VI, que representa a fase crônica terminal da doença, são consideradas as formas mais graves da NL (MUSA; ROUT; QURIE, 2025). No que diz respeito aos parâmetros clínicos, as classes III e IV, em geral, apresentam proteinúria acima de 500 mg/24 h e, em casos graves, perda da função renal. Além disso, existe a combinação de alterações proliferativas e membranosas, as classes III e IV podem estar associadas à classe V, também chamadas de formas mistas (III/IV+V) (REIS-NETO et al., 2024).

A NL mista possui parâmetros clínicos e histológicos mais graves do que a NL de classe V, como observado no trabalho de Samy e colaboradores (2025), isso porque a classe mista conta com as classes graves da doença (III e IV). Entretanto, quando comparada a NL mista com a NL proliferativa pura (III e IV), achados na literatura sugerem que a NL mista tem um prognóstico pior do que a LN proliferativa pura, podendo resultar em desfechos renais desfavoráveis (IKEUCHI et al., 2016). Ainda, a classe histopatológica pode mudar durante o tratamento da NL, sendo clinicamente difícil de classificar a NL proliferativa da LN membranosa ou da classe mista, por estarem envolvidas com proteinúria grave e diminuição da função renal (ANEKTHANAKUL et al., 2021).

Na literatura, a maioria dos trabalhos utilizam parâmetros clínicos para propor alternativas pouco invasivas na classificação histopatológica da NL (ARAÚJO JÚNIOR et al., 2023; WANG et al., 2025). Em relação à metabonômica ou metabolômica, um dos poucos achados é o estudo de Anekthanakul e colaboradores (2021), que utilizou análise direcionada para encontrar metabólitos capazes de distinguir indivíduos saudáveis de pacientes com NL, bem como a classe III/IV da classe V pura. Unindo a razão entre dois metabólitos com a TFGe e a UPCR, os autores chegaram em resultados promissores.

Dado que a biópsia renal é um procedimento invasivo, que o prognóstico da NL mista ainda ser incerto e que seu desfecho clínico pode diferir da NL proliferativa pura, faz-se necessário a continuidade na busca por alternativas não invasivas que possam,

não só discriminar essas classes e auxiliar na tomada de decisão, como também fornecer informações a respeito da diferença entre elas a nível metabólico. Sendo assim, o presente estudo investigou amostras de pacientes adultos com NL proliferativa e NL mista (proliferativa e membranosa).

Abaixo, seguem os estudos desenvolvidos no âmbito desta tese.

Estudo 1 – Diagnóstico de Lesão Renal Aguda em Recém-Nascidos Prematuros

2.3. Objetivos específicos

- Investigar e otimizar algoritmos de aprendizado de máquina supervisionado em um conjunto de dados de LRA em prematuros publicado – Conjunto de dados de Lesão Renal Aguda.
- Avaliar o desempenho dos modelos com base nas figuras de mérito e os deslocamentos químicos importantes para as discriminações das classes.
- Identificar os principais metabólitos responsáveis pela discriminação entre presença e ausência de LRA nesta população.

2.4. Materiais e Métodos

2.4.1. Conjunto de dados – Lesão Renal Aguda (LRA)

O estudo envolveu um total de 40 bebês prematuros (<31 semanas), que coletaram urina no dia 2 de vida, 20 deles com o diagnóstico de LRA (11 do sexo feminino) e 20 deles sem o diagnóstico de LRA (13 do sexo feminino). Foram incluídos os critérios de inclusão, peso ao nascer (≤ 1200 g) e idade gestacional (≤ 31 semanas), e de exclusão doença renal congênita e não sobrevivência após as primeiras 48 horas de vida (MERCIER et al., 2017).

Incialmente, o conjunto de dados utilizado foi proveniente da literatura, do trabalho de MERCIER e colaboradores (2017), intitulado *Preterm Neonatal Urinary Renal Developmental and Acute Kidney Injury Metabolomic Profiling: An Exploratory Study*. Estes dados estão disponíveis no site do *National Metabolomics Data Repository (NMDR)* do *NIH Common Fund*, o *Metabolomics Workbench*, **https://www.metabolomicsworkbench.org** , onde foi atribuído o ID do Projeto ***PR000048***. Os dados podem ser acessados diretamente por meio do DOI do

Projeto: ***10.21228/M8101K***. Este trabalho é apoiado pelo *Metabolomics Workbench/National Metabolomics Data Repository* (NMDR) (concessão nº U2C-DK119886), *Common Fund Data Ecosystem* (CFDE) (concessão nº 3OT2OD030544) e *Metabolomics Consortium Coordinating Center* (M3C) (concessão nº 1U2C-DK119889).

2.4.2. Espectroscopia de RMN de ^{1}H - LRA

O biofluido investigado foi urina. As amostras foram preparadas com adição de D_2O e uma solução de padrão interno contendo ácido 4,4-dimetil-4-silapentano-1-sulfônico (DSS, referência de deslocamento químico), imidazol (indicador de pH) e NaN_3 (para inibir o crescimento bacteriano) foi adicionada. Os tubos foram misturados, centrifugados e uma alíquota de 200 µL do sobrenadante foi transferida para tubos de RMN de 3 mm. Mais detalhes do preparo de amostra podem ser acessados no *Metabolomics Workbench (Projeto **PR000048**)*.

As 40 amostras de urina foram analisadas por RMN de ^{1}H em um espectrômetro da Bruker de 22,3 T, operando a 950 MHz, localizado no *David H. Murdock Research Institute* em Kannapolis, NC, EUA. Os dados foram adquiridos usando uma sequência de pulsos de presaturação 1D NOESY (noesypr1d, [tempo de reciclagem (TR) de 2 s - 90°-t1-90°-tm-90°-aquisição do decaimento de indução livre]) e os espectros foram processados usando no software Chenomx NMR Suite 7.51 Professional (Chenomx, Edmonyon, Alberta Canada)

2.4.3. Processamento dos dados - LRA

Neste estudo, o tratamento dos espectros e dos dados foi conduzido de maneira diferente do realizado no artigo original (MERCIER et al., 2017) em alguns aspectos. Abaixo estão descritos os tratamentos realizados por eles e, em seguida, os procedimentos adotados no presente estudo.

MERCIER e colaboradores: A fase e a linha de base foram corrigidas manualmente para cada espectro. Os espectros foram referenciados internamente ao sinal do DSS. A divisão dos espectros foi feita com bin de 0,04 ppm entre δ 0,50-9,00 ppm (229 variáveis), excluindo a região do sinal do DSS, água (4,68 - 4,88 ppm), urea (5.60–6.00 ppm) e imidazol (7,20 - 7,28 ppm). A normalização foi feita pela integral total de cada espectro (NCV - norma 1). Em seguida, a matriz formada pelos dados

normalizados e agrupados por bin foi escalada pelo método de Pareto e centralizada antes das análises multivariadas, que foram: PCA e OPLS-DA. A validação cruzada de 7-fold foi utilizada para avaliar o desempenho do modelo com base na sua capacidade de predição (Q^2).

Presente estudo: Utilizando o software MestReNova (versão 12.0.0) a fase foi corrigida manualmente e a linha de base usando polinômios de Bernstein. Os espectros foram referenciados internamente ao sinal do DSS. A divisão dos espectros foi feita com o mesmo valor de bin (0,04 ppm), mas na região entre δ 0,50 – 4,60 ppm (102 variáveis). A matriz de dados foi criada em *.csv* contendo 40 amostras e 102 variáveis.

2.4.4. Análise Quimiométrica

A análise quimiométrica do conjunto de dados, incluindo as etapas de pré-processamento, visualização, análise exploratória e modelos de classificação, foram desenvolvidos em linguagem *Python 3*, utilizando o ambiente interativo de programação *Google Colaboratory (Colab),* uma plataforma gratuita oferecida pelo Google para execução em *Jupyter Notebook.* Para a execução dos algoritmos de aprendizados de máquina e todas as etapas necessária, uma série de bibliotecas foi utilizada:

Scikit-learn (PEDREGOSA et al., 2011), *numpy* (HARRIS et al., 2020), p*andas* (MCKINNEY, 2010), *seaborn* (WASKOM, 2021), *scipy* (VIRTANEN et al., 2020), *matplotlib* (HUNTER, 2007), *statsmodels* (SEABOLD; PERKTOLD, 2010), *imbalanced-learn* (LEMAITRE; NOGUEIRA; ARIDAS, 2017) e *tqdm* (MATIYASEVICH, 2015).

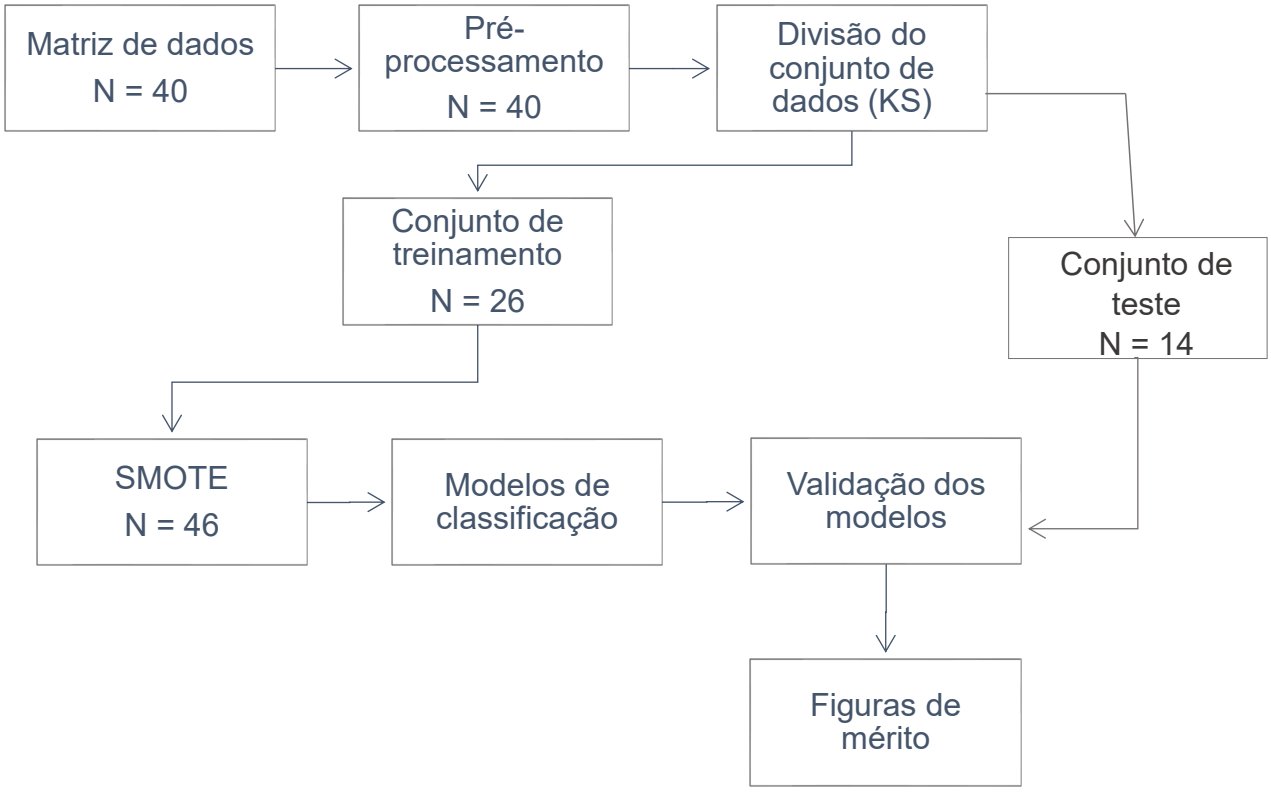
Inicialmente, os dados foram normalizados utilizando a normalização pelo comprimento do vetor (norma 2) e autoescalados para a PCA e métodos de classificação. Durante a PCA, os outliers identificados foram removidos. LR, SVM e LDA foram os algoritmos de classificação escolhidos e tiveram seus parâmetros otimizados utilizando o *GridSearchCV.* O desempenho dos modelos de classificação foi avaliado utilizando 8 figuras de mérito calculadas a partir da matriz de contingência obtida da validação do conjunto de teste. As figuras de mérito foram: exatidão (Eq. 24), VPP (Eq. 22), VPN (Eq. 23), sensibilidade (Eq. 20), especificidade (Eq. 21), *F1-Score* (Eq. 25), Kappa (Eq. 26) e AUROC. O teste de permutação foi empregado em todos os modelos de classificação para avaliar a não aleatoriedade da predição.

Inicialmente, os 40 espectros das amostras de urina de bebês prematuros foram divididos em conjunto de treinamento e teste, 13 e 7 amostras por classe (com e sem LRA), respectivamente, utilizando o algoritmo Kennard-Stone (KENNARD; STONE, 1969), que seleciona as amostras com base nas distâncias entre elas. A reprodutibilidade dos modelos foi garantida com o parâmetro random_state.

Embora as classes de treinamento já estivessem balanceadas (13 amostras cada), a Técnica de Sobre-amostragem de Minorias Sintéticas (SMOTE, do inglês *Synthetic Minority Over-sampling Technique*) foi utilizada para sintetizar e aumentar o número de amostras, devido ao número reduzido de amostras, que poderia comprometer a construção do modelo preditivo. Foram geradas 10 amostras sintéticas adicionais, todas utilizadas na construção do modelo, sendo adicionadas às amostras originais do conjunto de treino, totalizando 23 amostras por classe. O conjunto de teste foi composto exclusivamente por amostras reais, utilizadas para predição como um conjunto externo para avaliar a qualidade do modelo.

Em seguida, os algoritmos de aprendizado de máquina LR, SVM e LDA foram treinados com 46 amostras (23 de cada classe) e validados com 14 amostras (7 de cada classe) no conjunto de teste.

Figura 15 - Fluxograma do processamento realizado no conjunto de dados - LRA.



2.4.5. Identificação dos metabólitos

Definido o melhor modelo, por meio das figuras de mérito, para prosseguir com as análises das variáveis mais importantes na classificação foram investigadas. O processo de atribuição dos possíveis metabólitos foi realizado utilizando os bancos de dados eletrônicos como HMDB (do inglês, *Human Metabolome Database*) (WISHART *et al*., 2022), BMRB (do inglês, *Biological Magnetic Resonance Bank*) (ULRICH *et al*., 2008) e artigos publicados. A busca pelas possíveis rotas metabólicas afetadas foi realizada utilizando o *Pathways Analysis* da plataforma *Metaboanalyst* 6.0.

2.5. Resultados e Discussão - LRA

Como descrito por Mercier e colaboradores (2017), as características demográficas dos pacientes não apresentaram diferenças estatisticamente significativas (Tabela 3), mais informações podem ser encontradas no artigo.

Tabela 3 - Características dos indivíduos entre controles e casos da LRA.

	Sem LRA (média ± DP)	*LRA (média ± DP)*	*p-valor*
N	20	20	-
Peso ao nascer (g)	834,0 ± 291,9	815,8 ± 288,2	0,84
Idade Gestacional (semanas)	26,3 ± 2,3	26,1 ± 2,3	0,84
Sexo (F/M)	11,0/9,0	13,0/7,0	0,52

Fonte: MERCIER et al., 2017

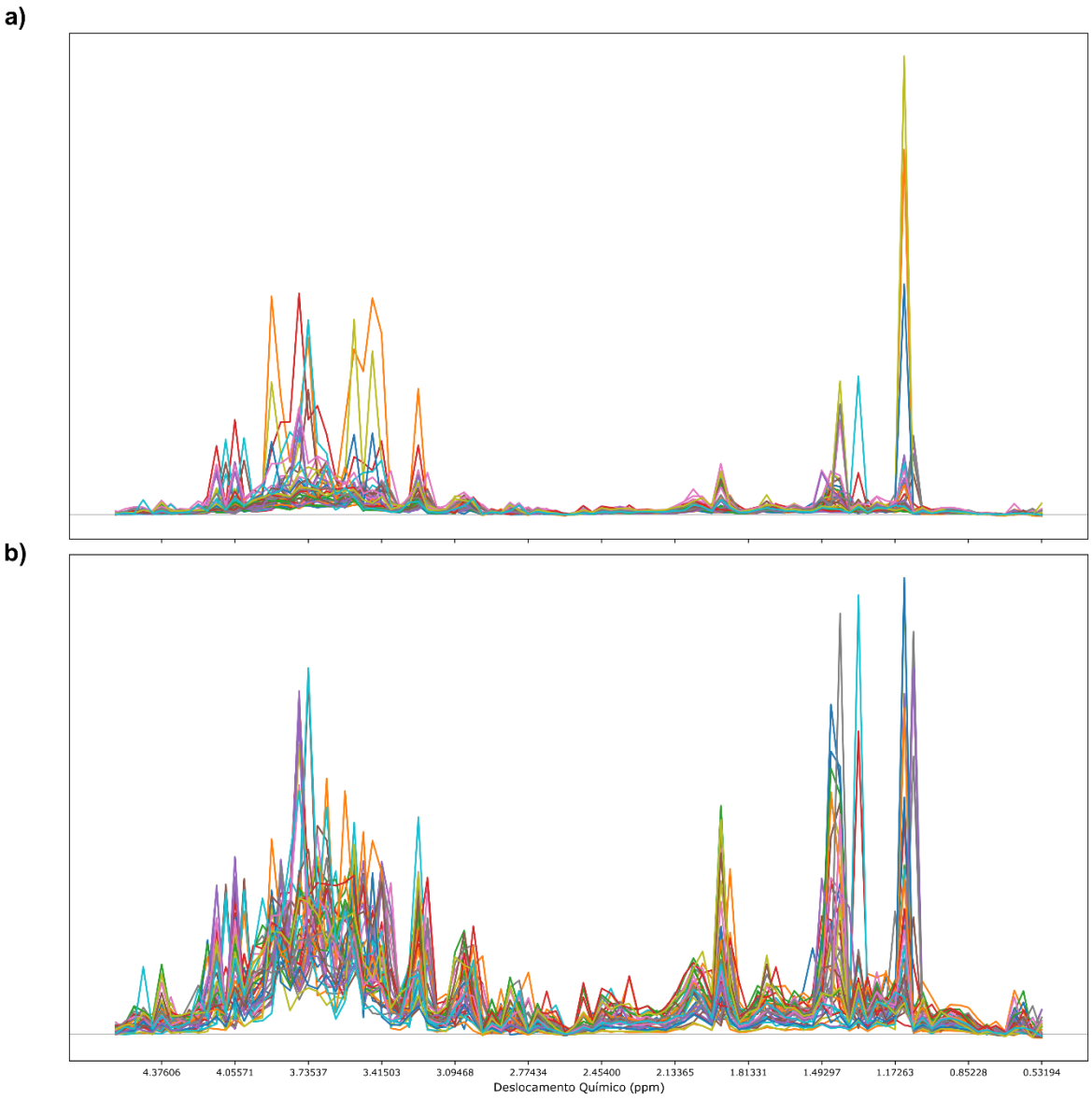
Portanto, a análise dos dados foi focada na discriminação entre bebês prematuros com e sem o diagnóstico de LRA, desconsiderando outros fatores de intervenção. Os algoritmos de aprendizado de máquina tiveram seus parâmetros otimizados (Apêndice A), foram treinados e tiveram seus desempenhos avaliados.

Toda a análise de dados foi realizada com as variáveis na faixa de 0,50 – 4,60 ppm e apesar de existir sinais além dessa faixa de deslocamento químico, a variação do pH na amostra de urina deslocou alguns sinais nessa região, que não foi possível ajustar com o sinal de referência. Além disso, trabalhos na literatura como o de Gronwald e colaboradores (2011), mostram que a maior parte dos metabólitos responsáveis por discriminar pacientes com doença renais em amostra de urina, estão nessa faixa de deslocamento químico.

2.5.1. Visualização dos dados

Inicialmente, a matriz de dados pós processamento dos espectros foi normalizada e, em seguida, auto-escalada. A Figura 16 apresenta os espectros antes e após a normalização, sendo possível notar que alguns sinais estavam muito mais intensos em algumas poucas amostras do que nas demais. Vale ressaltar que as amostras passaram por um processo de preparo que as tornam propensas a erros sistemáticos. Após a normalização, essas variações, possivelmente de origem experimental ou características do próprio material biológico, são minimizadas, preservando a informação qualitativa e quantitativa que distingue uma amostra da outra.

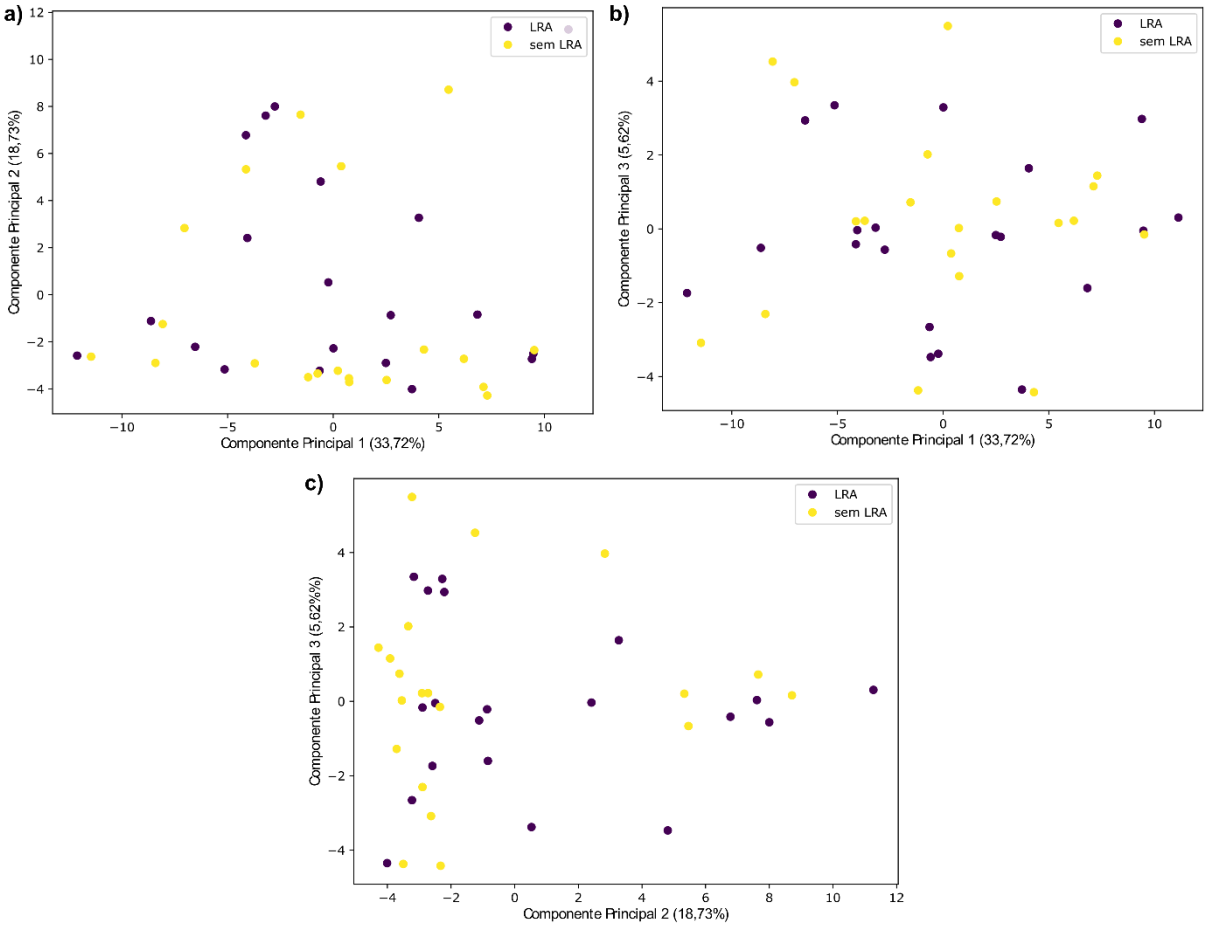
Figura 16 – Espectros de RMN de ^{1}H – análise de amostras de urina. a) originais; b) normalizados (norma 2 - euclidiana).



Fonte: A autora (2025)

Em seguida, os dados foram auto-escalados e a análise exploratória foi realizada com a PCA. As três primeiras componentes principais (PC1, PC2 e PC3) explicaram juntas 58,07% da variância dos dados e não foram observadas tendências de separação entre as classes de interesse. Esse comportamento corrobora com o que foi observado por Marcier e colaboradores (2017) em seu artigo, apesar dos autores não terem especificado se algum método de escalamento foi aplicado, mencionando apenas a normalização. A Figura 17 apresenta os gráficos de escores para as combinações entre PC1, PC2 e PC3.

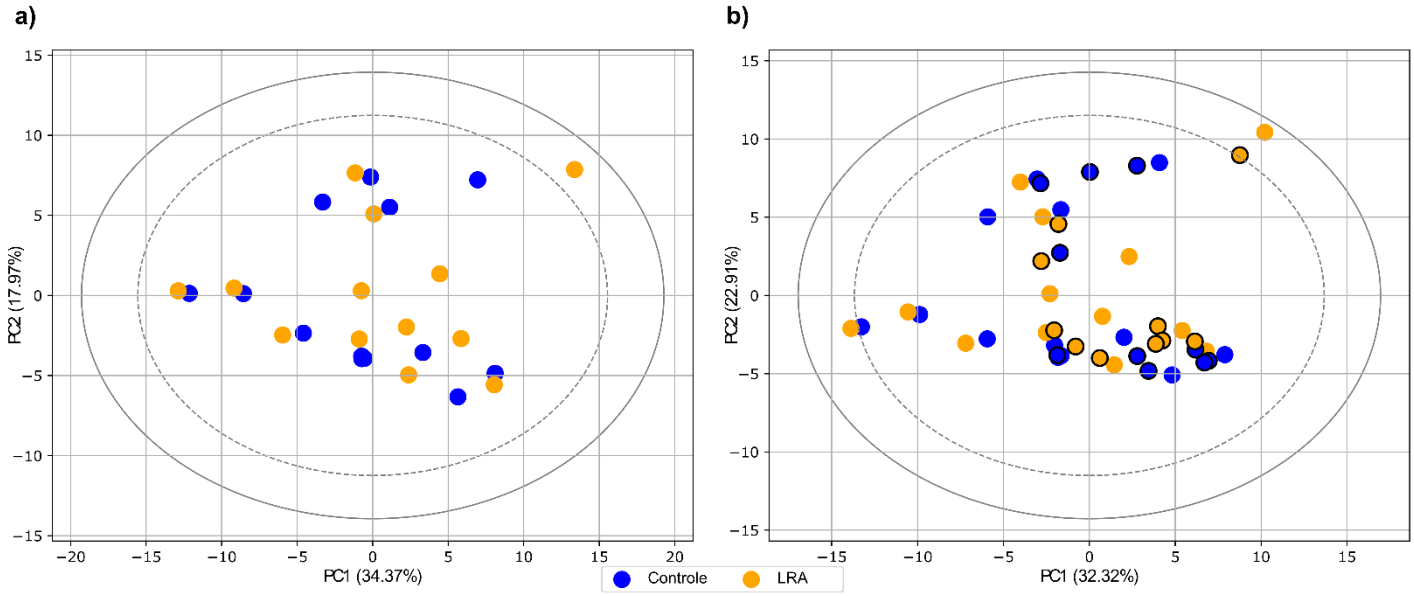
Figura 17 - Gráficos de escores. a) PC1 vs PC2; b) PC1 vs PC3; c) PC2 vs PC3.



Fonte: A autora (2025)

Os gráficos de pesos, que indicam a influência das variáveis na distribuição das amostras em relação a cada PC, estão disponíveis no Apêndice B (Figura B1). Para a construção dos modelos de classificação o conjunto de dados foi dividido em treino e teste como descrito na metodologia. Antes da modelagem, para lidar com o baixo número de amostras, ambas as classes do conjunto de treinamento foram sobreamostradas utilizando o algoritmo SMOTE, número de vizinhos igual a 5. A PCA foi então gerada um conjunto de treinamento antes e depois do SMOTE, conforme presente na Figura 18.

Figura 18 - Gráfico de escores da PCA. a) antes e b) depois do SMOTE.



O SMOTE cria amostras sintéticas posicionando-as no eixo formado entre pares de amostras reais da classe minoritária. Observa-se inicialmente o aumento do número de amostras entre as elipses de 95% e 99% confiança do T^2 de Hotelling. Como nenhuma delas ficou fora da elipse não houve exclusão no conjunto de dados. Além disso, notou-se que as amostras sintéticas apresentam um padrão mais contraído em relação à distribuição verdadeira dos dados. Esse comportamento é resultado do processo de geração do SMOTE, que utiliza interpolação linear, fazendo com que as novas amostras tendam a ser posicionadas mais próximas do centro da nuvem de dados originais (ELREEDY; ATIYA, 2019).

Apesar da distribuição observada na PCA, é importante verificar se os padrões gerados pelo SMOTE permanecem consistentes com a distribuição original das amostras. Para investigar se as distribuições das populações antes e depois do SMOTE diferem significativamente, foi aplicado o teste de Kolmogorov–Smirnov (KS) (KOLMOGOROV, 1933) às 102 variáveis. A diferença máxima entre as duas distribuições foi quantificada pela estatística D. Os resultados indicaram que as variáveis não apresentaram mudanças estatisticamente significativas, não havendo evidências para rejeitar a hipótese nula de igualdade das distribuições ($p > 0,05$). Os valores individuais do teste podem ser consultados no Apêndice C (Tabela C1). Em seguida, os modelos de classificação foram treinados com as amostras.

2.5.2. Modelos de Classificação

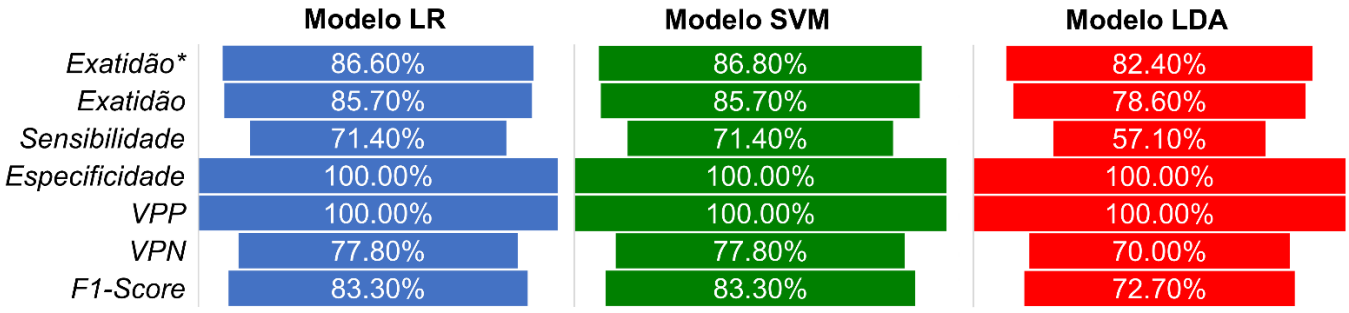
A escolha dos algoritmos de aprendizado de máquina para os modelos de classificação foi realizada com o objetivo de testar classificadores de fácil interpretação, visto que, no contexto clínico, entender como os metabólitos estão associados à doença é uma necessidade.

A matriz de dados após as etapas de pré-processamento ficou com dimensão de 60 x 102. A divisão em conjunto de treinamento e teste resultou em 46 amostras para o treinamento do modelo e 14 amostras de teste para obter as figuras de mérito. A Tabela 4 apresenta a matriz de contingência para cada um dos modelos e a Figura 19 e a Tabela 5 trazem as figuras de mérito, que são responsáveis por informar a capacidade de classificar e discriminar as classes investigadas.

Tabela 4 – Matriz de contingência dos modelos LR, SVM e LDA.

		Diagnóstico Padrão	
		LRA	**Controle**
LR	**LRA**	5	0
	Controle	2	7
SVM	**LRA**	5	0
	Controle	2	7
LDA	**LRA**	4	0
	Controle	3	7

Figura 19 - Resultados da validação dos modelos – LR, SVM e LDA.



*Validação Cruzada

Fonte: A autora (2025)

Tabela 5 – Resultados da validação dos modelos – LR, SVM e LDA.

Figuras de Mérito	***LR***	***SVM***	***LDA***
Kappa	0,714	0,714	0,571
AUROC	0,857	0,857	0,786
Valor de p	0,02	0,02	0,066

Os modelos treinados neste capítulo pertencem a dois grupos de algoritmos de aprendizado de máquina: paramétricos (LR, LDA e SVM linear) e não paramétricos (SVM kernel). Modelos paramétricos, como a Regressão Logística (LR) e a Análise Discriminante Linear (LDA), requerem a estimação de parâmetros diretamente a partir dos dados, tornando-se suscetíveis a problemas de multicolinearidade e à alta dimensionalidade (HAYATI et al., 2024). No caso do presente estudo, em que o conjunto de treinamento continha apenas 46 amostras e 102 variáveis, essas limitações se tornam mais evidentes. De fato, o LDA não apresentou desempenho no nível dos demais, possivelmente devido à instabilidade da matriz de covariância estimada em um cenário onde o número de variáveis excede o número de amostras (NAM; KIM; CHUNG, 2020).

Por outro lado, tanto o SVM linear quanto a Regressão Logística alcançaram bons resultados nas validações cruzadas e externa. Embora a LR também seja paramétrica, sua formulação regularizada permite lidar melhor com conjuntos de dados de alta dimensionalidade. Já o SVM, por ser um modelo não paramétrico, apresenta maior robustez frente à multicolinearidade, pois sua estratégia de encontrar o hiperplano de separação ótimo não depende da inversão de matrizes de covariância (SHARMA et al., 2024). Assim, as características intrínsecas de cada algoritmo impactaram diretamente a performance observada, favorecendo os classificadores lineares SVM e LR no contexto avaliado.

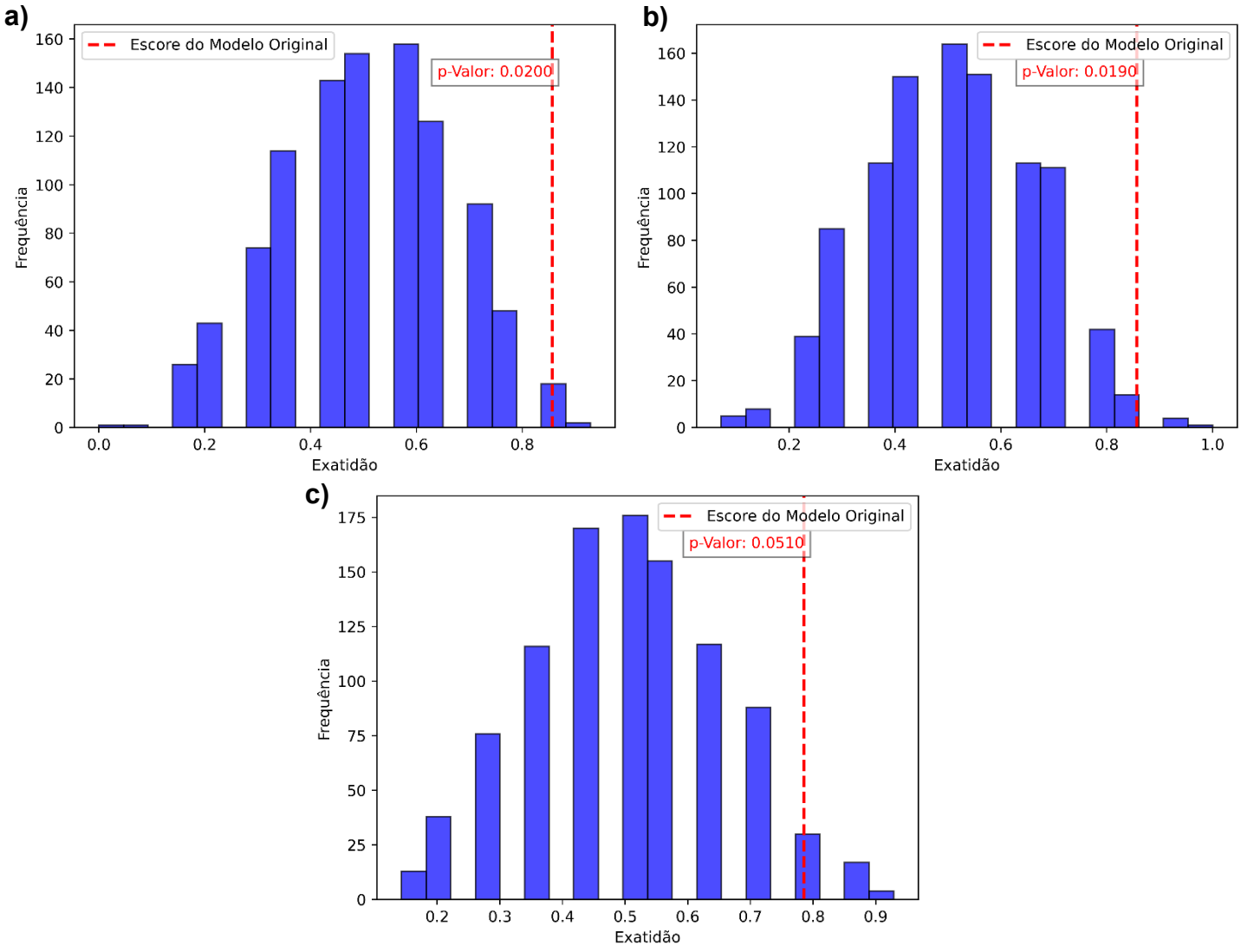
Trabalhos recentes, como o de Huang e colaboradores (2025), investigaram quatro algoritmos de aprendizado de máquina em pacientes com sepse que apresentavam ou não LRA. Utilizando amostras de soro, cromatografia líquida e três metabólitos séricos identificados, alcançaram valores de AUC de 0,90 e 0,83 para LR e SVM, respectivamente. Já no trabalho de Lee e colaboradores (2022), com doença renal crônica pediátrica e utilizando a cromatografia como técnica de análise, o SVM apresentou os melhores resultados de AUC entre os algoritmos empregados; entretanto, o F1-score, média harmônica entre precisão e sensibilidade, não ultrapassou 0,51. Esses resultados evidenciam o uso cada vez mais frequente de aprendizado de máquina em estudos envolvendo doenças renais e destacam os resultados do presente trabalho como promissores.

Embora as figuras de mérito tenham mostrado robustez dos modelos no presente trabalho, é importante considerar que eles podem apresentar deficiências na capacidade de generalização devido a *overfitting* (superajuste) (DE ANDRADE et al.,

2020). Mesmo que os modelos se ajustem bem aos dados de treinamento e prevejam com precisão os dados de teste, o conjunto de amostras do teste é pequeno e pode ter sido classificado por uma modelagem com um ajuste aos dados aleatórios, em vez de uma verdadeira capacidade de generalização.

O teste de permutação é uma técnica estatística usada para avaliar a significância do desempenho de um modelo de aprendizado de máquina. Neste estudo, foi aplicado um teste de permutação de exatidão. Os modelos foram treinados e re-treinados com amostras cujas etiquetas foram embaralhadas aleatoriamente. Para cada permutação, a exatidão do modelo no conjunto de teste foi calculada. Esse processo foi repetido 1000 vezes para cada um dos três modelos. A Figura 20 traz o histograma e o p-valor para cada teste.

Figura 20 – Histogramas referente aos testes de permutação: a) LR, b) SVM e c) LDA.



Fonte: A autora (2025)

O histograma acima mostra a distribuição das exatidões obtidas a partir de 1000 permutações aleatórias das etiquetas de treinamento. Cada barra no histograma representa a frequência das acurácias obtidas em cada intervalo específico de acurácia. A linha vertical vermelha indica a acurácia do modelo original obtida no conjunto de teste. Observa-se que a maioria das acurácias permutadas estão concentradas em torno de valores mais baixos com poucas permutações resultando em acurácias tão altas quanto a do modelo original, isso para os modelos SVM e LR. Isso sugere que o desempenho do modelo original é superior ao que seria esperado ao acaso. Para o modelo LDA, a acurácia do modelo original encontra-se em valor mais baixo, região que há uma frequência maior de acurácias permutadas.

O valor de p calculado, que é a proporção de permutações que resultaram em uma acurácia maior ou igual à do modelo original, é baixo ($p < 0,05$) para os modelos SVM e LR. Isso indica que é improvável que a alta acurácia do modelo original seja devida ao acaso para esses modelos, sugerindo que eles possuem um desempenho significativo. Entretanto, o modelo LDA apresentou $p > 0,05$, logo, não possui um desempenho estatisticamente significativo.

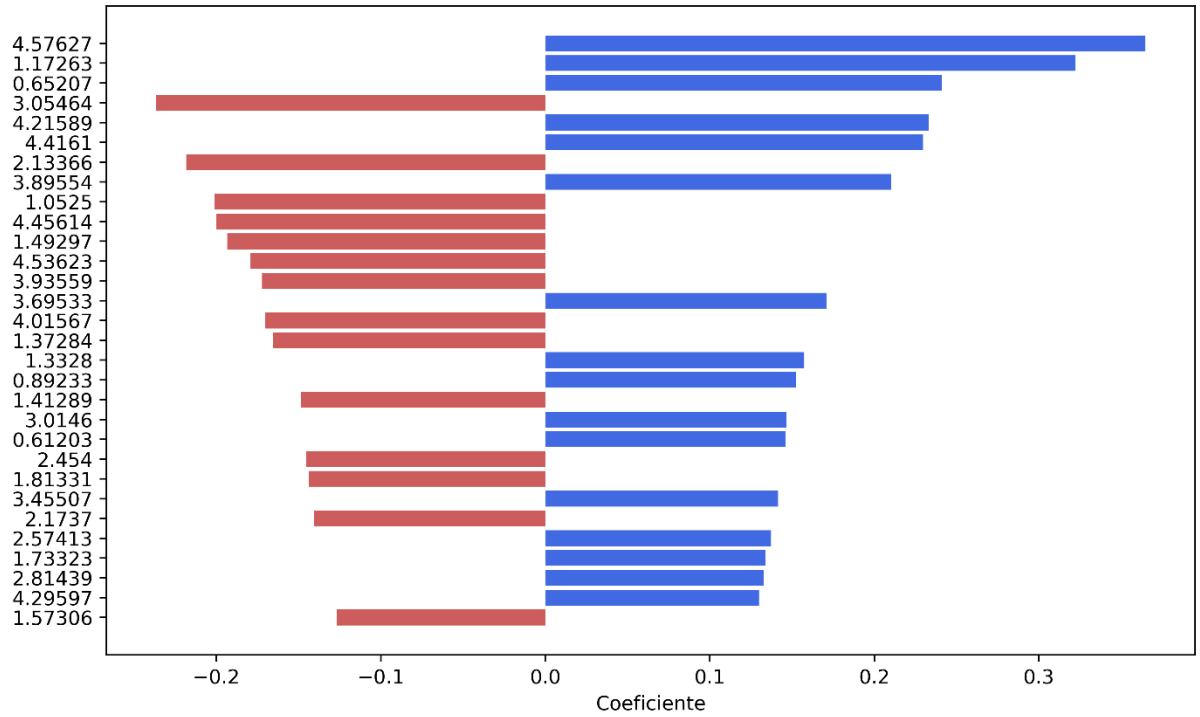
Apesar do alto desempenho do modelo SVM, mesmo considerando a matriz sem nenhuma redução da dimensionalidade, esses resultados foram obtidos com um conjunto de dados pequeno. Deve-se ser cauteloso, pois a performance observada pode não se confirmar em um conjunto de dados maior e mais realista, sendo necessário investigações futuras com maiores conjuntos de dados. Ainda assim, mesmo de forma preliminar, foram obtidos resultados promissores utilizando os algoritmos de aprendizado de máquina investigados neste estudo, associados a dados obtidos por espectroscopia de RMN de ^1H, para discriminar indivíduos com lesão renal aguda de indivíduos saudáveis.

2.5.3. Identificação dos metabólitos

A etapa de identificação dos metabólitos é de grande importância nos estudos metabonômicos, a partir dela é possível identificar as rotas metabólicas que são perturbadas como resultado da homeostase. Para encontrar as variáveis mais influentes nas previsões, foram usados os valores dos coeficientes absolutos do modelo, referentes às variáveis que têm maior impacto na decisão do SVM. Ao observar a Figura 21, podemos separar as variáveis em dois grupos: coeficientes positivos e negativos. Os coeficientes positivos sugerem que a variável está

fortemente associada à classe positiva (1 - paciente com LRA), enquanto os coeficientes negativos indicam que está associada à classe negativa (0 - controle).

Figura 21 – Importância das variáveis no modelo SVM.



Fonte: A autora (2025)

Com uma análise inicial, foi possível identificar seis metabólitos diferenciais utilizando o banco de dados HMDB (do inglês, Human Metabolome Database) e artigos da literatura. As estruturas químicas (Figura 22) e os deslocamentos químicos δ1,05, δ1,33, δ3,01, δ3,05, δ3,45, δ3,93, foram atribuídos à valina, lactato, lisina, creatinina, taurina, creatina, respectivamente (Figura 23).

Figura 22 - Estrutura dos seis metabólitos identificados. Posição dos Hidrogênios ligados a carbonos primários e secundários referente aos deslocamentos químicos em destaque.

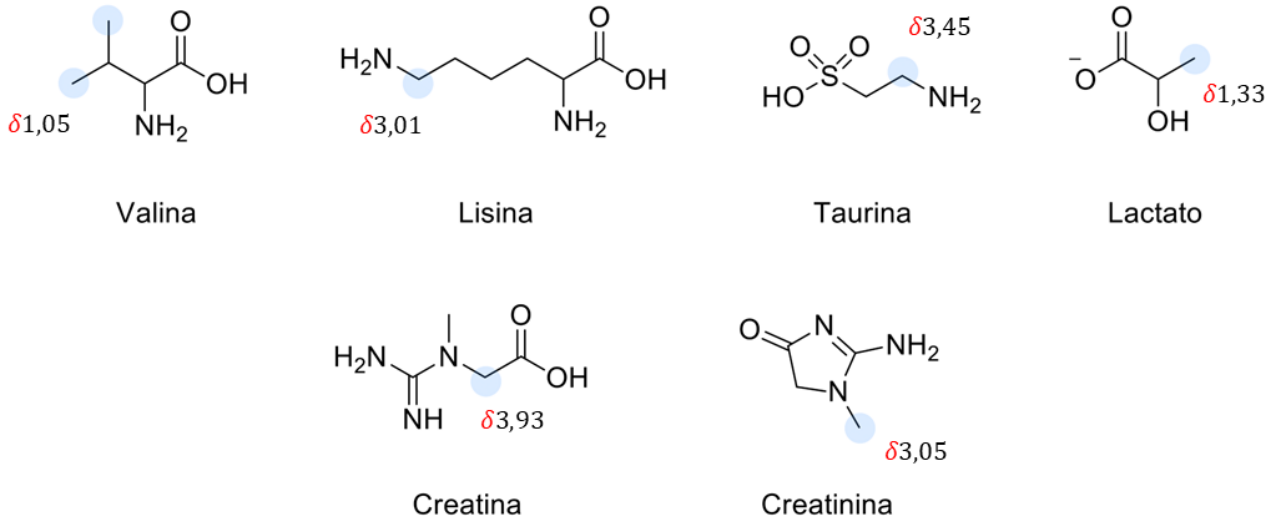
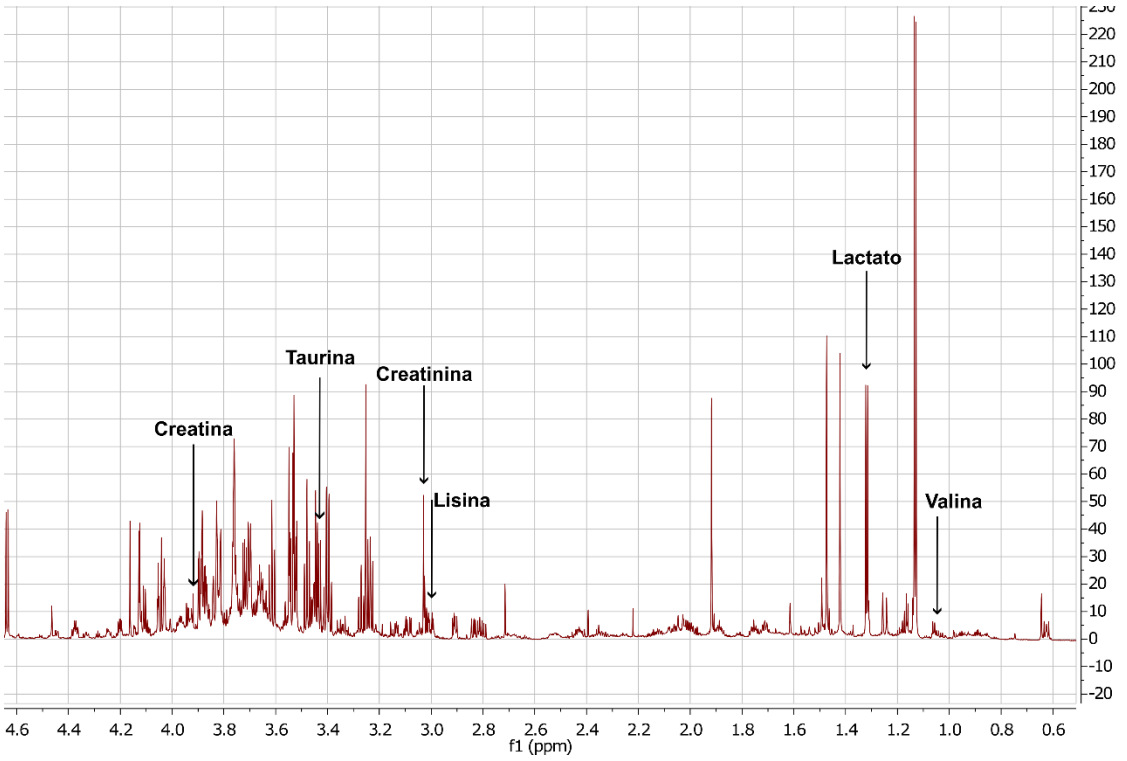


Figura 23 – Corte de um espectro de urina de pacientes com LRA na região dos metabólitos identificados.



Fonte: A autora (2025)

O coeficiente positivo para o sinal do lactato (Figura 21) indica que altos valores desse metabólito estão associados à LRA. Esses metabólitos também foram responsáveis por discriminar casos de controle no estudo de Fonseca e colaboradores (2023), que investigaram a DRC empregando a estratégia metabonômica. Kim e

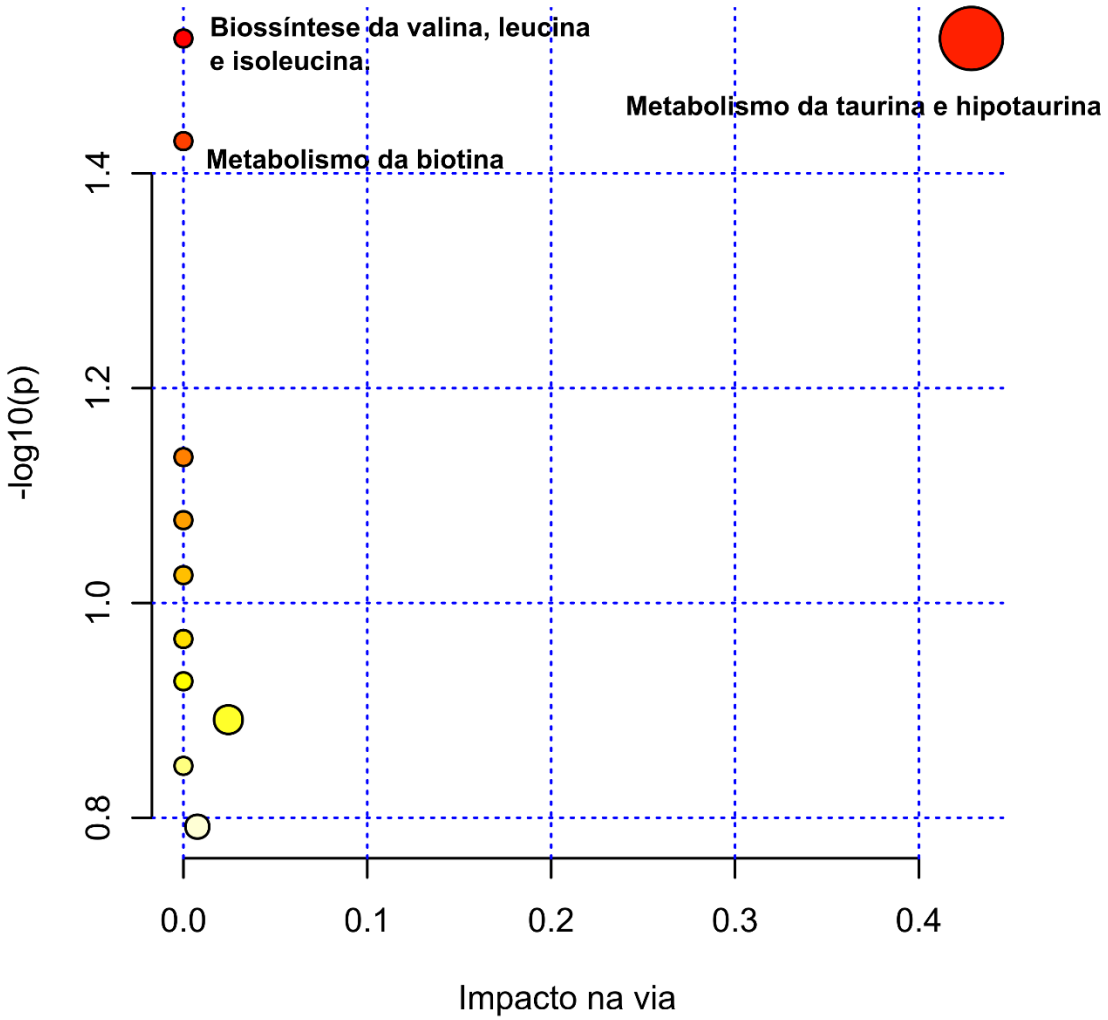
colaboradores (2014), que trabalharam com ratos com DRC, encontraram o sinal do lactato substancialmente maior nos ratos com DRC.

O coeficiente negativo para o sinal da creatinina indica que altos valores desse metabólito estão associados ao grupo de controle, mas a creatinina sérica já é utilizada como biomarcador no diagnóstico de LRA e está associada ao seu aumento em pacientes que têm a doença. Entretanto, há baixa sensibilidade e especificidade atreladas à creatinina para avaliar e determinar a LRA precoce, pois condições clínicas podem comprometer a determinação de suas concentrações (FONSECA et al., 2023). Ainda, no artigo dos autores responsáveis pelo banco de dados (MERCIER et al., 2017), tanto a creatinina quanto a creatina foram metabólitos correlacionados à idade gestacional, exclusivos de neonatos sem diagnóstico de LRA, o que corrobora os resultados da análise do presente trabalho, visto que esses dois metabólitos estão associados ao grupo de controle.

No que diz respeito à valina e à lisina, há relatos na literatura que apresentaram ambos os metabólitos como variáveis importantes na construção de modelos PLS envolvendo doenças renais (CHASAPI et al., 2021; FRANIEK et al., 2022). Franiek e colaboradores (2022), empregaram metabolômica urinária para desenvolver preditores de lesão renal aguda pediátrica utilizando cromatografia gasosa. A taurina se destacou como o segundo metabólito mais importante para o modelo de classificação de LRA *vs.* Controle.

A análise de vias metabólicas foi realizada no *MetaboAnalyst 6.0*, sendo representada por um gráfico de bolhas (Figura 24). As três vias destacadas na Figura 24 apresentaram significância estatística (p-valor < 0,05), mas apenas o metabolismo da taurina apresentou-se como uma via que sofreu impacto. As demais estão posicionadas no zero do eixo de impacto, logo, o impacto dessas vias é praticamente nulo e não devem ser consideradas relevantes para interpretações.

Figura 24 - Gráfico de bolhas da análise das vias metabólicas na LRA.



Fonte: A autora (2025)

Segundo Franiek e colaboradores (2022), diversos estudos associam a taurina à proteção renal, destacando seu papel na regulação do fluxo sanguíneo renal e da função endotelial vascular. Em lesões renais induzidas por isquemia/reperfusão (I/R), a taurina atua atenuando os danos iniciais por meio da eliminação de espécies reativas de oxigênio durante a inflamação glomerular, além de ser excretada na urina após a lesão, o que justifica sua alteração observada em casos de disfunção renal.

O SVM se mostra mais uma vez como um algoritmo de aprendizado de máquina que pode ser explorado em estudos metabonômicos.

2.6. Conclusão do Estudo 1 – Lesão Renal Aguda

O presente capítulo teve como objetivo inicial investigar e avaliar o emprego de algoritmos de aprendizado de máquina em um conjunto de dados da literatura para discriminar bebês com diagnóstico de LRA de bebês prematuros sem o diagnóstico. Utilizando a estratégia metabonômica baseada em espectros de RMN de ^{1}H, três algoritmos foram investigados: Regressão Logística (LR), Máquina de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA), empregando SMOTE e sem etapas prévias de redução da dimensionalidade da matriz, seus desempenhos foram avaliados por meio das figuras de mérito. A validação dos modelos foi realizada com um conjunto de dados extraídos da matriz original.

Entre os três modelos avaliados, o SVM e a LR apresentaram os melhores desempenhos, ambos errando a classificação de duas amostras, resultando em exatidão de 86% e VPP de 100%. Esses resultados sugerem que a informação está contida no conjunto de dados, sendo acessível pela estratégia metabonômica. A LDA obteve desempenho inferior aos demais, além disso, com base nos testes de permutação, o modelo não se mostrou estatisticamente significativo, estando sobreajustado.

Considerando que foram usados algoritmos de classificação fáceis de interpretar, foram selecionadas as 30 variáveis mais importantes para o desempenho do SVM linear na discriminação das classes. Seis deslocamentos químicos foram atribuídos a metabólitos: valina, lactato, lisina, creatinina, taurina e creatina, os quais corroboram resultados da literatura e cuja variação está associada a doenças renais, com destaque para a taurina, cujo metabolismo foi a principal rota alterada na análise das vias metabólicas.

Estudo 2 – Estadiamento da Nefrite Lúpica Proliferativa com ou sem lesão membranosa.

2.7. Objetivos específicos

- Obter espectros de amostras de soro de pacientes com Nefrite Lúpica por Ressonância Magnética Nuclear de ^{1}H.
- Investigar e otimizar algoritmos de aprendizado de máquina supervisionado, combinados com técnicas de seleção de variáveis, para o conjunto de dados de Nefrite Lúpica.
- Construir modelos metabonômicos para discriminar pacientes com Nefrite Lúpica proliferativa com ou sem lesão membranosa.
- Avaliar o desempenho dos modelos e suas combinações com os métodos de seleção de variáveis com base nas figuras de mérito.
- Identificar os metabólitos referentes aos sinais de maior importância para construção do melhor modelo.

2.8. Materiais e Métodos

2.8.1. Amostragem – Nefrite Lúpica (NL)

A seleção dos pacientes incluídos neste estudo e as coletas de sangue foram realizadas no ambulatório do setor de Nefrologia do Hospital das Clínicas da Universidade Federal de Pernambuco (HC–UFPE), em pacientes acompanhados por pesquisadores parceiros e pela equipe do ambulatório de Glomerulonefrite do próprio hospital. As amostras de sangue foram coletadas e armazenadas no Laboratório de Imunopatologia Keizo Asami (LIKA/UFPE).

Foram incluídos 36 pacientes, 25 com NL classificada como III/IV e 11 com NL classificada como III/IV+V, com mais de 18 anos de idade, com diagnóstico de lúpus e nefrite lúpica confirmada por biópsia renal (classes III a V). Como critérios de exclusão, foram considerados: pacientes que não preenchiam os critérios para LES ou biópsia de nefrite lúpica; aqueles sem laudo histopatológico confirmatório de nefrite lúpica; e pacientes com sorologias positivas para hepatite B, hepatite C, HIV ou sífilis.

Todos os pacientes estavam sob terapia de indução com metilprednisolona intravenosa, seguida de prednisona oral e seis doses únicas e rápidas de ciclofosfamida intravenosa (500 mg a 1 g) ou micofenolato de mofetila (2 a 3 g/dia).

Após a fase de indução, os pacientes receberam terapia de manutenção com azatioprina ou micofenolato, conforme protocolos previamente estabelecidos.

2.8.2. Considerações Éticas – NL

Este estudo foi aprovado pelo comitê de ética da Universidade Federal de Pernambuco (CAAE 11401219.2.0000.8807) e está de acordo com a Declaração de Helsinque. Todos os pacientes recrutados assinaram um termo de consentimento livre e esclarecido.

2.8.3. Espectroscopia de RMN de ^1H - NL

Para a realização da análise, as amostras foram descongeladas e uma alíquota de 400 µL da amostra de soro foi transferida para um tubo de RMN, com 5 mm de diâmetro interno. Em seguida, foram adicionados 200 µL de água deuterada (D_2O) e o conteúdo foi homogeneizado manualmente. Os espectros de RMN de ^1H foram obtidos usando a sequência de pulsos CPMG (Carr-Purcell-Meiboom-Gill), com pré-saturação (presat) do sinal da água em 4,70 ppm, no espectrômetro modelo Ascend (Bruker®) operando a 400 MHz. Foram utilizados os seguintes parâmetros experimentais: i) tempo de pré-saturação igual a 3 s, ii) janela espectral de 5,88 kHz, iii) número de ciclos de 126, iv) tau igual a 0,0003 s, v) bigtau igual a 0,076 s e vi) 128 transientes. Os espectros foram processados usando line broadening de 0,3 Hz.

2.8.4. Processamento dos dados - NL

O processamento dos espectros foi conduzido utilizando o software MestReNova 12.0.0, no qual a linha de base foi corrigida automaticamente e a fase foi ajustada no modo manual. Em seguida, o sinal correspondente ao grupo metil do lactato foi utilizado como referência e o espectro alinhado ao seu sinal em δ 1,33 ppm. Todos os espectros foram sobrepostos graficamente para verificação da qualidade dos ajustes realizados. O espectro foi cortado para remover regiões que não continham os compostos de interesse, ficando definida a faixa entre δ 0,5 e 4,50 para continuar com as análises. A região então definida, foi dividida em intervalos (bins) de 0,04 ppm para a construção da matriz de dados. Por fim, a matriz de dados foi exportada para o excel constituída por 36 amostras e 100 variáveis, a qual foi em seguida submetida às análises quimiométricas.

2.8.5. Análise Quimiométrica

A análise quimiométrica do conjunto de dados, incluindo as etapas de pré-processamento, visualização, análise exploratória e modelos de classificação, foram desenvolvidos em linguagem *Python 3*, utilizando o ambiente interativo de programação *Google Colaboratory (Colab),* uma plataforma gratuita oferecida pelo Google para execução em *Jupyter Notebook.* Para a execução dos algoritmos de aprendizados de máquina e todas as etapas necessária, uma série de bibliotecas foi utilizada:

Scikit-learn (PEDREGOSA et al., 2011), *numpy* (HARRIS *et al.*, 2020), p*andas* (MCKINNEY, 2010), *seaborn* (WASKOM, 2021), *scipy* (VIRTANEN *et al.*, 2020), *matplotlib* (HUNTER, 2007), *statsmodels* (SEABOLD; PERKTOLD, 2010), *imbalanced-learn* (LEMAITRE; NOGUEIRA; ARIDAS, 2017) e *tqdm* (MATIYASEVICH, 2015).

Inicialmente, os dados foram normalizados utilizando a normalização pelo comprimento do vetor (norma 2) e autoescalados para a PCA e métodos de classificação. Durante a PCA, os outliers identificados foram removidos. LR, SVM e LDA foram os algoritmos de classificação escolhidos e tiveram seus parâmetros otimizados utilizando o *GridSearchCV.* O desempenho dos modelos de classificação foi avaliado utilizando 8 figuras de mérito calculadas a partir da matriz de contingência obtida da validação do conjunto de teste. As figuras de mérito foram: exatidão (Eq. 24), VPP (Eq. 22), VPN (Eq. 23), sensibilidade (Eq. 20), especificidade (Eq. 21), *F1-Score* (Eq. 25), Kappa (Eq. 26) e AUROC. O teste de permutação foi empregado em todos os modelos de classificação para avaliar a significância estatística da exatidão dos modelos.

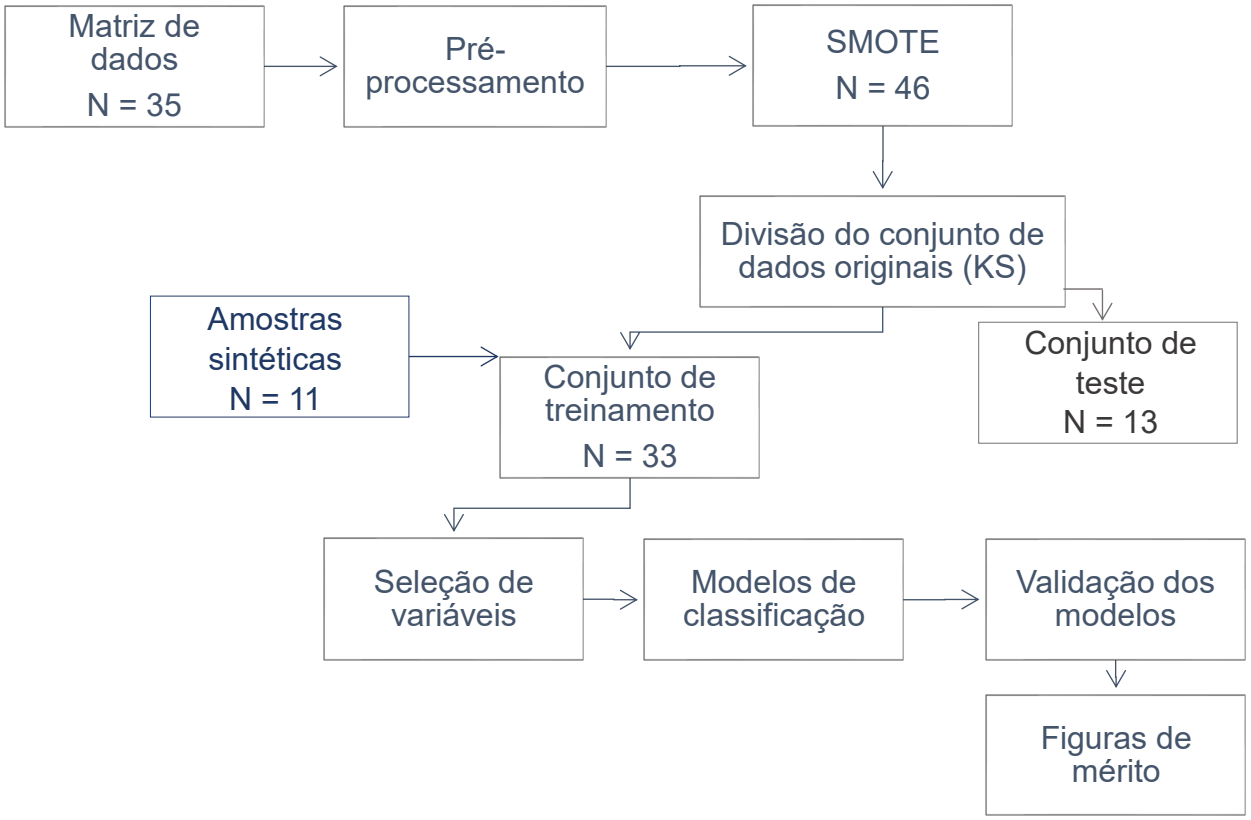
Inicialmente, foi empregada a técnica de SMOTE com o objetivo de sintetizar e aumentar o número de amostras da classe minoritária, uma vez que o desbalanceamento compromete a construção de modelos preditivos robustos. A classe minoritária (III/IV+V), que anteriormente possuía 11 amostras, teve sua quantidade dobrada. Ao final, a matriz de dados foi composta por 24 amostras da Classe III/IV e 22 da Classe III/IV+V (11 verdadeiras e 11 sintéticas). Antes da divisão em grupos de treino e teste, as amostras sintéticas foram removidas do conjunto de dados, e a divisão, utilizando o método de Kennard-Stone, foi realizada apenas nas amostras originais. A distribuição das amostras encontra-se apresentada na Tabela 6.

Tabela 6 - Divisão dos conjuntos de treino e teste para cada classe da NL.

	Classe III/IV	**Classe III/IV+V**
Treino	17	16
Teste	7	6

O conjunto de teste foi composto exclusivamente por amostras reais, utilizadas para predição como um conjunto externo para avaliar a qualidade do modelo. Durante a etapa de treinamento dos modelos, foram testadas combinações entre métodos de seleção de variáveis e de classificação, com objetivo de reduzir a dimensionalidade. No total, foram 18 combinações compostas por três algoritmos de classificação (SVM, LDA e LR) sem seleção de variáveis e combinados com três métodos de seleção. Os métodos de seleção utilizados foram o SFM (*Select From Model*), SFS (*Sequential Forward Selection*), usando o classificador LR (*Logistic Regression*) e RF (*Random Forest*), e o GA (*Genetic Algorithm*).

Figura 25 - Fluxograma do processamento realizado no conjunto de dados - NL.



2.8.6. Identificação dos metabólitos

Definido o melhor modelo, por meio das figuras de mérito, para prosseguir com as análises as variáveis mais importantes na classificação foram investigadas. O processo de atribuição dos possíveis metabólitos foi realizado utilizando os bancos de dados eletrônicos como HMDB (do inglês, *Human Metabolome Database*) (WISHART *et al*., 2022), BMRB (do inglês, *Biological Magnetic Resonance Bank*) (ULRICH *et al*., 2008) e artigos publicados. A busca pelas possíveis rotas metabólicas afetadas foi realizada utilizando o *Pathways Analysis* da plataforma *Metaboanalyst* 6.0.

2.9. Resultados e Discussão - NL

2.9.1. Dados clínicos

Foram incluídas 36 amostras, classificadas em dois grupos segundo os resultados da biópsia renal: classe III/IV e classe III/IV+V. Para avaliar diferenças significativas entre as classes, foi utilizado o teste de Mann–Whitney para as variáveis contínuas e o teste qui-quadrado para as variáveis categóricas. A Tabela 7 apresenta as características demográficas e parâmetros clínicos, divididos conforme a distribuição das classes.

Tabela 7 - Características demográficas e parâmetros clínicos de acordo com as classes da NL.

	Classe III/IV (média ± DP)	*Classe III/IV+V (média ± DP)*	*p-valor*
N	25	11	-
Idade (anos ± DP)	33,2 ± 10,2	28,8 ± 5,2	0,409*
Sexo (F/M)	21,0/4,0	10,0/1,0	0,977**
IMC (média ± DP)	25,5 ± 5,8	23,3 ± 5,2	0,410*
Tempo de diagnóstico do LES (meses)	61,7 ± 67,1	107,5 ± 69,7	0,022*
Tempo de diagnóstico da NL (meses)	42,9 ± 46,01	69,9 ± 63,2	0,056*
Proteinúria	3,0 ± 1,4	6,6 ± 5,2	0,020*
Nº Glomérulos	17,4 ± 10,6	20,7 ± 8,6	0,310*

*Teste de Mann-Whitney/**Teste qui-quadrado

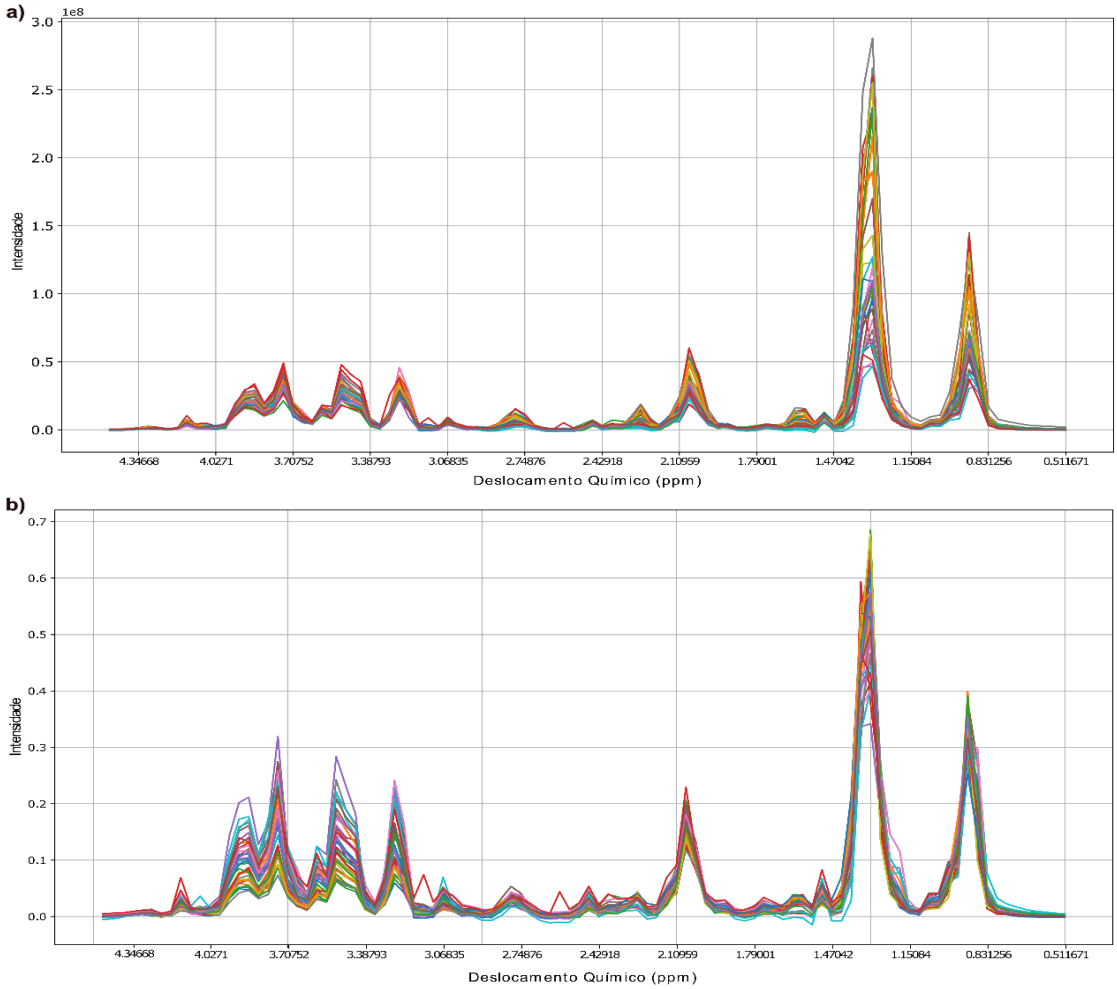
O tempo de diagnóstico do LES e a proteinúria foram as únicas variáveis que apresentaram diferenças estatisticamente significativas. Pacientes da classe III/IV+V apresentaram tempo médio de diagnóstico do LES significativamente maior (107,5 ± 69,7 meses) comparado a III/IV (61,7 ± 67,1 meses; p = 0,022). A proteinúria também foi significativamente maior na classe III/IV+V (6,6 ± 5,2 g/24h) em comparação à classe III/IV (3,0 ± 1,4 g/24h; p = 0,020), indicando maior comprometimento renal. Os resultados sugerem que pacientes com a lesão membranosa (classe V) associada à lesão glomerular podem apresentar um quadro clínico mais prolongado e maior gravidade renal, o que pode influenciar no tratamento e prognóstico (WANG et al., 2025).

2.9.2. Visualização dos dados

Inicialmente, a matriz de dados (36x100) composta por amostras de pacientes com nefrite lúpica (NL) proliferativa e nefrite lúpica mista (proliferativa e membranosa)

passou por etapas de pré-processamento, que incluem a normalização nas amostras e autoescalamento nas variáveis. A Figura 26 traz o espectro antes e depois da normalização.

Figura 26 - Espectros amostras de soro de pacientes com Nefrite Lúpica: a) sem normalização; b) normalizados.

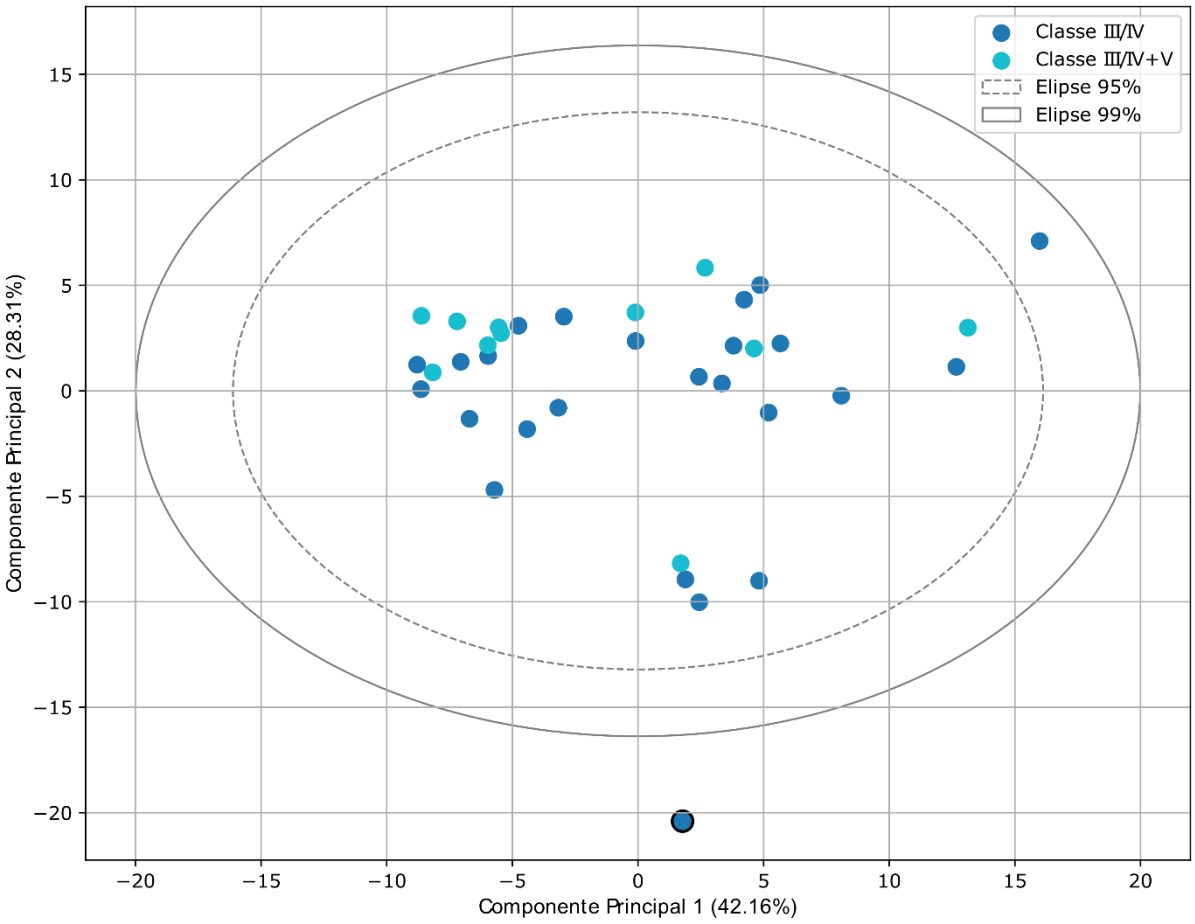


Fonte: A autora (2025)

Como já descrito anteriormente, a etapa de pré-tratamento na matriz de dados visa minimizar possíveis variações oriundas de erros durante a obtenção e preparo da amostra, ou até mesmo na aquisição dos espectros. Em seguida, foi realizada a PCA, com a finalidade de observar tendências de separação e identificar possíveis amostras anômalas. As três primeiras componentes principais (PCs, do inglês *Principal Components*) explicam cerca de 80% da variância dos dados, mas não foi possível observar uma tendência de separação entre as classes. O gráfico de escores (Figura 27), formado pelas duas primeiras PCs, mostra o grupo da NL de classe mista

(III/IV+V) menos disperso do que o grupo da NL proliferativa (III/IV); o mesmo padrão foi observado nas demais combinações de PCs (Apêndice D).

Figura 27 - Gráfico de escores da PCA formado por PC1 e PC2 das amostras de pacientes com NL.

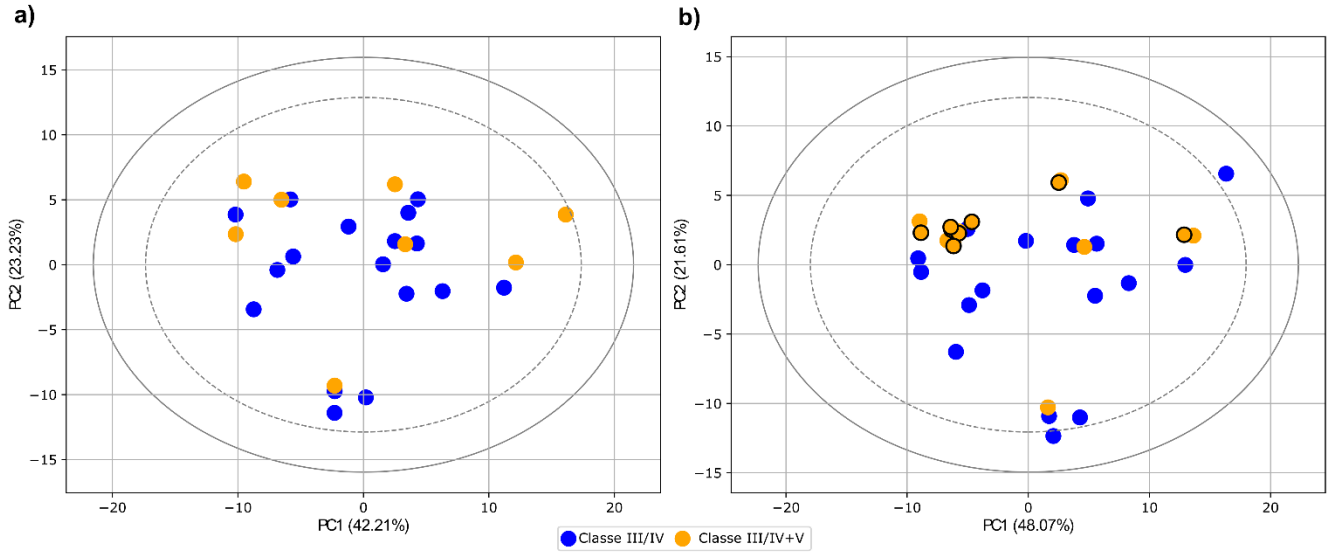


Fonte: A autora (2025)

No que diz respeito à presença de amostras anômalas, apenas uma, pertencente à classe III/IV+V, ficou fora das elipses de confiança do T^2 de Hotelling e foi removida da matriz de dados para dar continuidade às análises seguintes.

Considerando o desbalanceamento entre as classes investigadas, a etapa de sobreamostragem foi empregada utilizando o SMOTE, dobrando o número de amostras da classe minoritária, com o objetivo de alcançar figuras de mérito robustas e previsões eficientes durante o processo de construção dos modelos (RODRIGUES; LUNA; PINTO, 2023). Na Figura 28 estão os gráficos de escores da PCA antes e depois do SMOTE.

Figura 28 - Gráfico de escores NL: a) antes do SMOTE; b) depois do SMOTE.



Fonte: A autora (2025)

A distribuição das amostras sintéticas segue o padrão das demais, entretanto, é necessário investigar se, estatisticamente, isso de fato ocorre. Para isso, aplicou-se o teste de Kolmogorov-Smirnov (KS) (KOLMOGOROV, 1933) às 100 variáveis, comparando as distribuições originais com as obtidas após o SMOTE. A diferença entre elas foi quantificada pela estatística D, e os resultados indicaram que não há diferença estatisticamente significativa, com p-valor > 0,05 para todas as variáveis, não havendo evidências para rejeitar a hipótese nula de igualdade das distribuições. Os valores de D para cada variável, bem como os gráficos das 10 variáveis com maiores valores de D, estão apresentados no Apêndice E. Na próxima seção, os resultados dos modelos de classificação treinados com as amostras originais e sintéticas.

2.9.3. Modelos de classificação

Inicialmente, todas as amostras foram utilizadas para criar as amostras sintéticas com o SMOTE. Em seguida, sete amostras originais da classe III/IV e seis da III/IV+V, foram cuidadosamente selecionadas para o conjunto de teste usando o Kennard-Stone. Desse modo, apenas as amostras originais foram incluídas, garantindo que o diagnóstico do modelo não seja tendencioso. As demais amostras originais e todas as amostras sintéticas foram utilizadas como conjunto de treinamento, resultando em 17 amostras da classe III/IV e 16 amostras da classe III/IV+V. Seguindo com a escolha

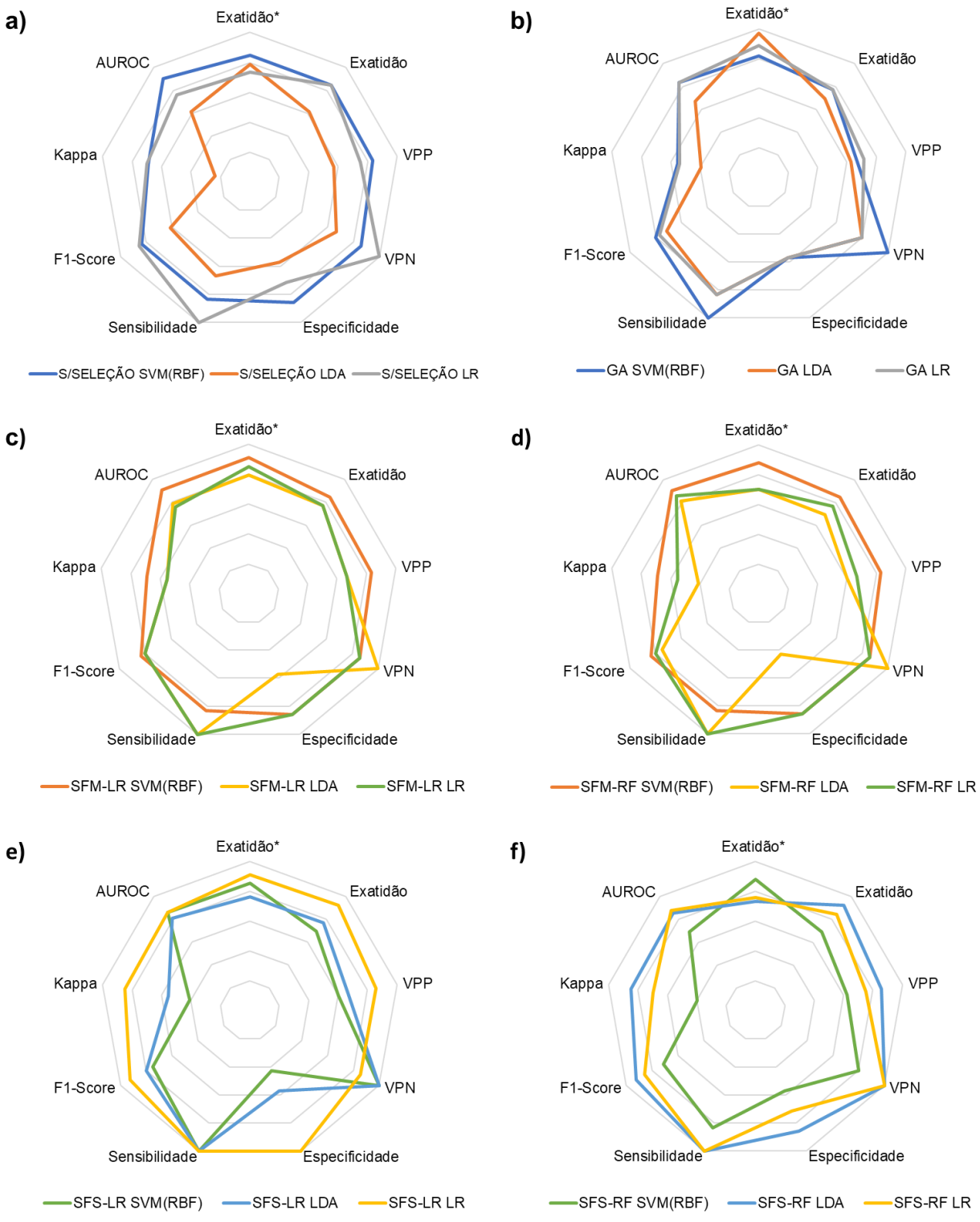
dos algoritmos de aprendizado de máquina com facilidade de interpretação, foram empregados o SVM, a LDA e o LR.

O conjunto de treinamento foi utilizado para construir os modelos de classificação com cada um dos métodos de seleção de variáveis que foram descritos na metodologia (SFM, SFS e GA) e tiveram seus parâmetros otimizados com o *GridSearchCV* (Apêndice A).

Com o intuito de facilitar a visualização e comparação do desempenho dos modelos em múltiplas figuras de mérito, utilizou-se um gráfico radar, que permite representar simultaneamente as diferentes métricas de forma intuitiva. No entanto, para maior rigor científico, os valores exatos das figuras de mérito estão apresentados na Tabela 8.

O gráfico radar (Figura 29) apresenta simultaneamente as figuras de mérito dos modelos avaliados (exatidão, VPP, VPN, especificidade, sensibilidade, f1-score, kappa e AUROC). Cada eixo do gráfico representa uma métrica, e o polígono formado pelos valores indica o desempenho geral do modelo. Quanto maior a área preenchida e mais próximo do limite máximo em todos os eixos, melhor o desempenho global. Essa representação facilita a comparação visual entre modelos, permitindo identificar rapidamente quais apresentam comportamento equilibrado em todas as métricas ou desempenho destacado em métricas específicas.

Figura 29 – Gráficos de radar construídos a partir das figuras de mérito dos modelos SVM, LDA e LR, considerando os métodos de seleção de variáveis**: a) Sem seleção; b) GA; c) SFM-LR; d) SFM-RF; e) SFS-LR; f) SFS-RF.



Fonte: A autora (2025)

*Validação cruzada (etapa de treinamento).
**Linhas de referência de 0 (menor raio) a 100% (maior raio), em intervalos de 20%.

Tabela 8 - Resultados da validação dos modelos.

	Sem seleção de variáveis			GA			SFM-LR			SFM-RF		
	SVM(RBF)	**LDA**	**LR**	**SVM(RBF)**	**LDA**	**LR**	**SVM(RBF)**	**LDA**	**LR**	**SVM(RBF)**	**LDA**	**LR**
*Exatidão**	84,8%	78,6%	73,3%	81,9%	97,1%	88.6%	91,4%	79,5%	85,2%	88,1%	70,0%	70,0%
Exatidão	84,6%	61,5%	84,6%	76,9%	69,2%	76.9%	84,6%	76,9%	76,9%	84,6%	69,2%	76,9%
VPP	83,3%	57,1%	75,0%	66,7%	62,5%	71.4%	83,3%	66,7%	66,7%	83,3%	60,0%	66,7%
VPN	85,7%	66,7%	100,0%	100,0%	80,0%	80.0%	85,7%	100,0%	85,7%	85,7%	100,0%	85,7%
Especificidade	85,7%	57,1%	71,4%	57,1%	57,1%	57.1%	85,7%	57,1%	85,7%	85,7%	42,9%	85,7%
Sensibilidade	83,3%	66,7%	100,0%	100,0%	83,3%	83.3%	83,3%	100,0%	100,0%	83,3%	100,0%	100,0%
F1-Score	83,3%	61,5%	85,7%	80,0%	71,4%	76.9%	83,3%	80,0%	80,0%	83,3%	75,0%	80,0%
Kappa	0,69	0,235	0,698	0,552	0,395	0.541	0,69	0,552	0,552	0,690	0,409	0,552
AUROC	0,905	0,615	0,762	0,833	0,667	0.833	0,905	0,786	0,762	0,905	0,81	0,857
*p-Valor***	0,029	0,198	0,005	0,069	0,118	0.04	0,012	0,029	0,015	0,018	0,151	0,081

	SFS-LR			***SFS-RF***		
	SVM(RBF)	**LDA**	**LR**	**SVM(RBF)**	**LDA**	**LR**
*Exatidão**	85,2%	76,2%	91,0%	88,1%	73,3%	76,2%
Exatidão	69,2%	76,9%	92,3%	69,2%	92,3%	84,6%
VPP	60,0%	66,7%	85,7%	62,5%	85,7%	75,0%
VPN	100,0%	100,0%	100,0%	80,0%	100,0%	100,0%
Especificidade	42,9%	57,1%	85,7%	57,1%	85,7%	71,4%
Sensibilidade	100,0%	100,0%	100,0%	83,3%	100,0%	100,0%
F1-Score	75,0%	80,0%	92,3%	71,4%	92,3%	85,7%
Kappa	0,409	0,552	0,847	0,395	0,847	0,698
AUROC	0,857	0,81	0,857	0,692	0,857	0,881
*p-Valor***	0,142	0,037	0,000	0,158	0,001	0,004

*Validação cruzada (etapa de treinamento).

**Teste de permutação

Ao observar os gráficos da Figura 29 e os resultados da Tabela 8, nota-se quais modelos de classificação apresentam melhor desempenho para cada método de seleção de variáveis. Entre todas as combinações avaliadas, o modelo LR aliado ao método SFS com classificador LR (SFS-LR) destacou-se, apresentando o melhor desempenho geral.

Oliveira e colaboradores (2024), que investigaram diferentes formalismos quimiométricos para o diagnóstico de câncer de próstata, concluíram que as combinações GA-LDA e GA-LR obtiveram os melhores resultados em relação às figuras de mérito. No presente estudo, a combinação SFS-LR com classificador LR não apenas apresentou desempenho superior às demais combinações testadas, como também alcançou valores de figuras de mérito (Tabela 9) compatíveis ou superiores aos relatados na literatura, indicando robustez do método empregado (ANEKTHANAKUL et al., 2021). O teste de permutação resultou em p-valor < 0,01, demonstrando a significância estatística do modelo.

Tabela 9 - Figuras de mérito do modelo LR combinado com o seletor SFS-LR.

Figuras de Mérito	***SFS-LR-LR***
Exatidão*	91,0%
Exatidão	92,3%
VPP	85,7%
VPN	100,0%
Especificidade	85,7%
Sensibilidade	100,0%
F1-Score	92,3%
Kappa	0,847
AUROC	0,857
Valor de p	0,002

*Validação cruzada (etapa de treinamento).

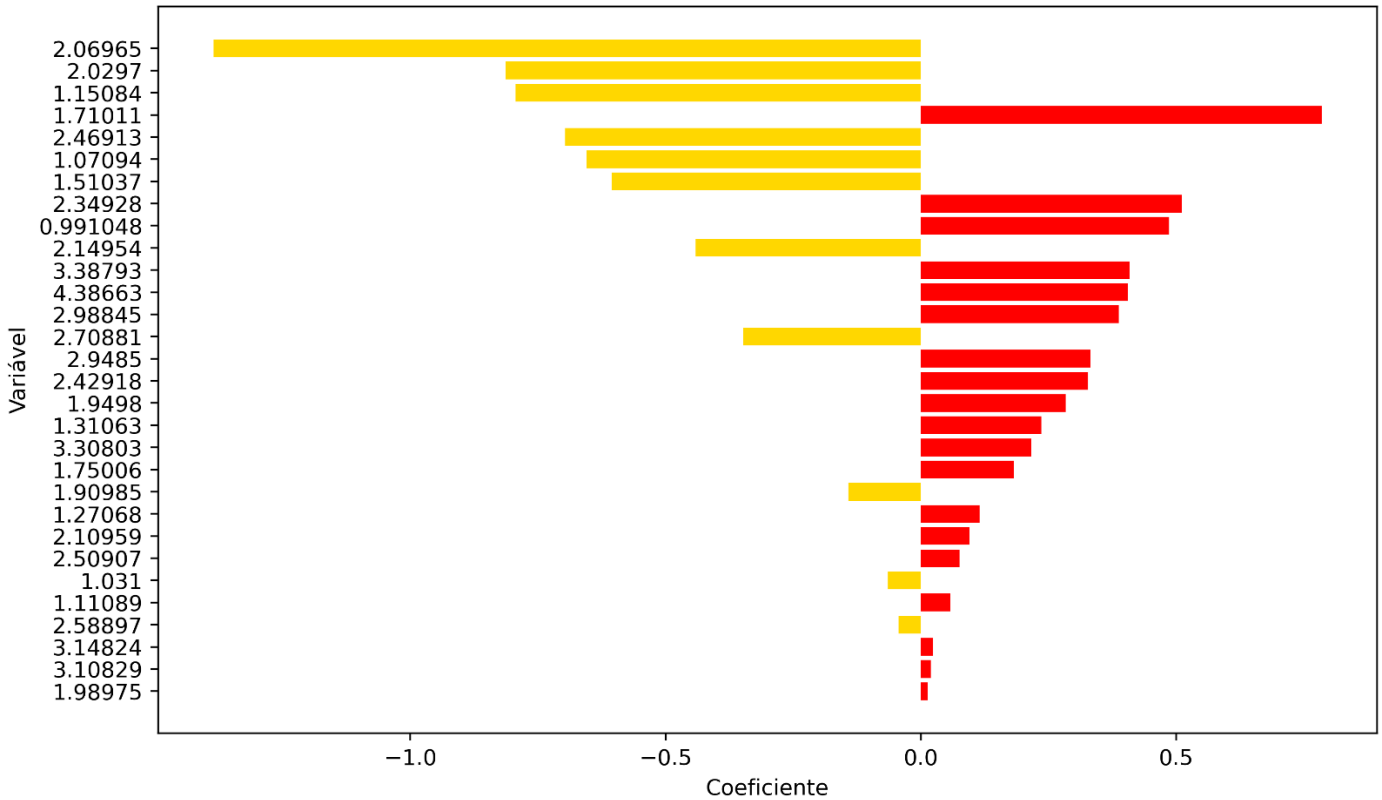
Anekthanakul e colaboradores (2021), que teve como objetivo a predição da nefrite lúpica membranosa utilizando biomarcadores urinários, testaram combinações de ácido picolínico (Pic), triptofano (Trp), TFGe e UPCR (Proporção de Proteína e Creatinina na Urina) para propor sete modelos de regressão logística, cujo desempenho foi avaliado por meio da AUROC. O modelo com melhor desempenho combinou a razão [Pic/Trp], TFGe e UPCR, alcançando uma AUROC de 0,91 na classificação de NL da classe III/IV em relação à classe V. Embora o estudo tenha fornecido informações valiosas e, assim como o presente trabalho, tenha mostrado

bom potencial na identificação de biomarcadores, ambos foram conduzidos em populações relativamente pequenas. Portanto, a utilização clínica desses metabólitos como biomarcadores depende de validações adicionais em conjuntos maiores.

2.9.4. Identificação dos metabólitos

A combinação SFS-LR selecionou 30 variáveis para os modelos de classificação. Em seguida, a importância de cada variável no modelo LR foi extraída e apresentada na Figura 30. Os coeficientes responsáveis pela construção do modelo podem ser divididos em positivos e negativos: os positivos sugerem que a variável está fortemente associada à classe positiva (1 – pacientes com NL classe III/IV+V), enquanto os negativos indicam associação à classe negativa (0 – pacientes com NL classe III/IV).

Figura 30 - Importância das variáveis na combinação SFS-LR com classificador LR.

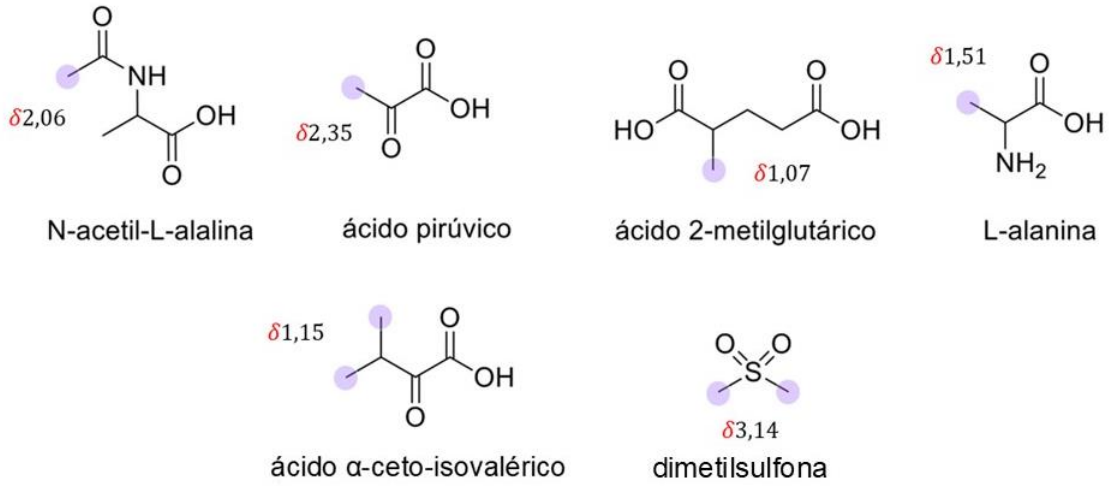


Fonte: A autora (2025)

A partir de buscas nos bancos de dados BRMB, HMDB e em artigos da literatura, oito compostos foram atribuídos aos deslocamentos químicos das variáveis importantes para o modelo. Os compostos (Figura 31) são: N-acetil-L-alanina (δ 2,06),

ácido pirúvico (δ 2,35), ácido 2-metilglutárico (δ 1,07), L-alanina (δ 1,51), ácido α-ceto-isovalérico (δ 1,15), lactato (δ 1,31), valina (δ 1,03) e dimetilsulfona (δ 3,14).

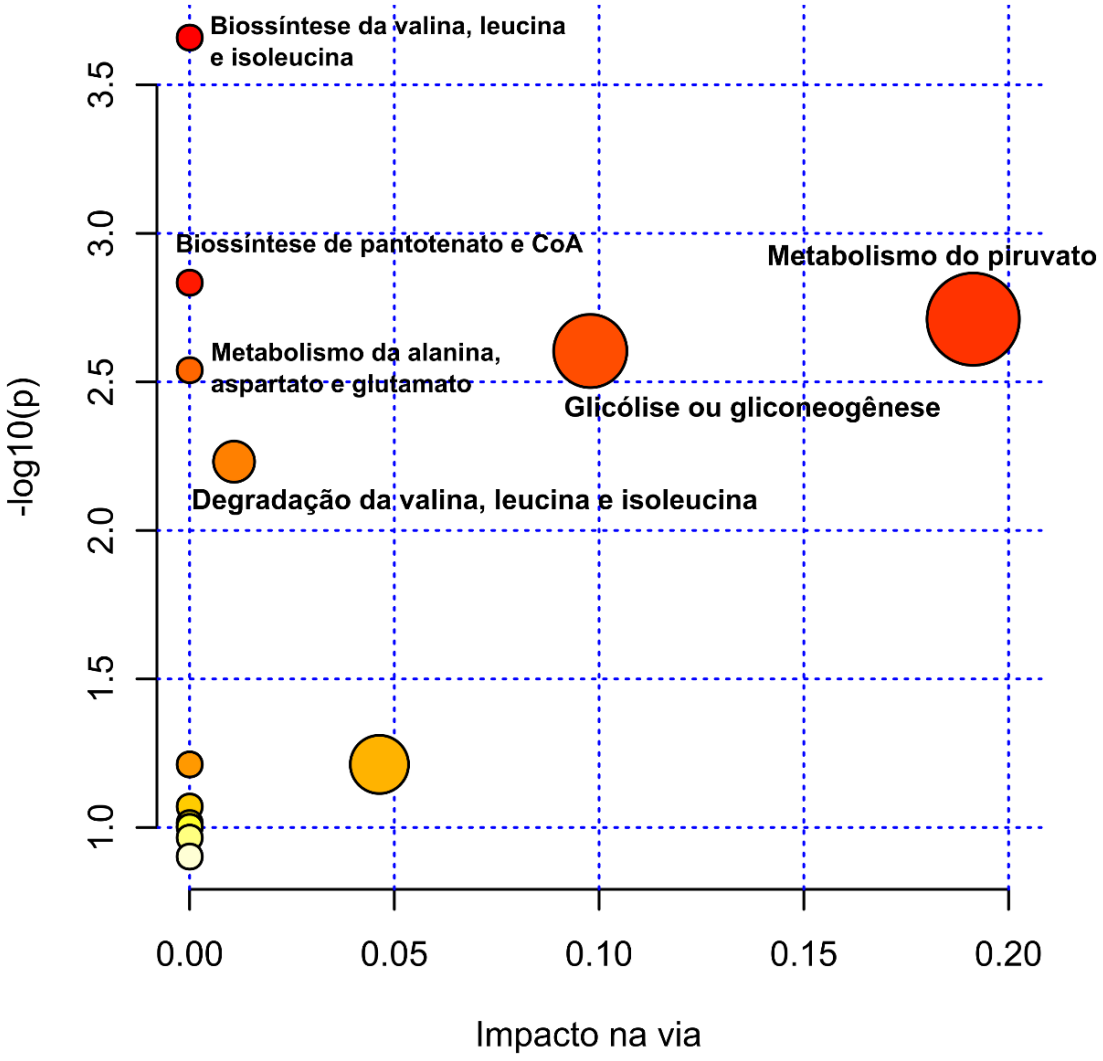
Figura 31 - Estrutura química dos metabólitos identificados. Posição dos Hidrogênios ligados aos carbonos primários referente aos deslocamentos químicos em destaque.



No que diz respeito à análise da variação de metabólitos entre as classes de NL, existem poucos trabalhos publicados. No entanto, a valina e a dimetilsulfona já foram empregadas na literatura para avaliação precisa da insuficiência renal, sendo a valina considerada um indicador do metabolismo ácido-base e a dimetilsulfona um marcador de estresse oxidativo. Como resultado, a degradação da valina associada à acidose metabólica reduz seus níveis na DRC, enquanto a dimetilsulfona tende a apresentar valores elevados (EHRICH et al., 2021). No contexto da NL, o presente estudo mostra que o coeficiente negativo da valina reflete seu leve aumento nos pacientes com classe III/IV, enquanto a dimetilsulfona aparece levemente associada à classe III/IV+V.

Para auxiliar a interpretação em relação a influência dos metabólitos na NL, a análise de vias metabólicas foi realizada no *MetaboAnalyst 6.0*. A análise indicou a glicólise e o metabolismo do piruvato como rotas significativamente enriquecidas (Figura 32). Em pacientes com lúpus eritematoso, as células T apresentam anormalidades metabólicas, incluindo aumento da glicólise e do estresse oxidativo, processos que promovem a geração de células inflamatórias (SHARABI; TSOKOS, 2020). Esses resultados estão de acordo com os níveis séricos observados de ácido pirúvico e lactato. O ácido pirúvico foi encontrado em maior concentração nos pacientes com NL classe III/IV+V, enquanto o lactato esteve levemente mais associado aos casos de NL classe III/IV.

Figura 32 - Gráfico de bolhas da análise das vias metabólicas na NL.



Fonte: A autora (2025)

Ainda, a biossíntese e degradação da valina, leucina e isoleucina, a biossíntese de pantotenato e CoA e o metabolismo da alanina, aspartato e glutamato também se apresentaram como rotas metabólicas significativamente enriquecidas, entretanto, o impacto nessas vias foi praticamente nulo. A N-acetil-L-alanina foi o metabólito que apresentou maior importância na construção do modelo, seu coeficiente negativo está associado aos pacientes com classe III/IV. Apesar de alguns aminoácidos como a alanina, glutamina e leucina serem descritas como essenciais no metabolismo das células T, não há menção específica à N-acetil-L-alanina (GUO et al., 2024). A presença desse metabólito na nossa análise pode representar a descoberta de uma derivação metabólica pouco explorada, o que justifica investigações futuras para entender melhor sua relação com a NL.

2.10. Conclusão do Estudo 2 – Nefrite Lúpica

Como continuação do capítulo na investigação de doenças renais, os algoritmos LR, LDA e SVM foram testados para discriminar pacientes com NL de classe III/IV dos pacientes de classe III/IV+V. Além do emprego do SMOTE para equilibrar as classes, três métodos de seleção de variáveis foram avaliados: SFS, SFM e GA. Por meio da análise das figuras de mérito, a maioria das combinações de modelos e seleção de variáveis, e até mesmo o modelo sem seleção, apresentaram bons desempenhos. Entre eles, o modelo LR após o SFS-LR apresentou os melhores resultados de classificação, com exatidão de 92,3% e sensibilidade de 100%, resultado de apenas uma amostra classificada de maneira incorreta.

Como resultado do modelo LR para NL, das 30 variáveis mais importantes, oito deslocamentos químicos foram atribuídos aos metabólitos: N-acetil-L-alanina, ácido α-ceto-isovalérico, ácido 2-metilglutárico, L-alanina, ácido pirúvico, lactato, valina e dimetilsulfona. Com destaque para o ácido pirúvico e lactato que podem ser associados com o aumento da glicólise e processos que promovem a geração de células inflamatórias. Com base na análise das vias metabólicas e de trabalhos da literatura, os metabólitos encontrados podem estar associados a variações no processo inflamatório da NL grave, podendo ser alvos de investigações futuras.

CAPÍTULO 3

3. Estudo 3. Aprendizado de Máquina empregado na Avaliação da Fibrose Periportal em Pacientes com Esquistossomose mansoni

A esquistossomose, considerada um grave problema de saúde pública, está entre as doenças tropicais negligenciadas que afetam, principalmente, países subdesenvolvidos da África e da América Latina (SANTOS et al., 2022). No Brasil, segundo o Ministério da Saúde (2024), a esquistossomose está presente intensamente numa faixa de terras ao longe de quase toda costa litorânea da Região Nordeste, do Rio Grande do Norte em direção ao sul, incluídas as zonas quentes e úmidas dos estados da Paraíba, de Pernambuco, Alagoas, Sergipe, do Maranhão e da Bahia, onde se interioriza alcançando Minas Gerais, no Sudeste, seguindo o trajeto de importantes bacias hidrográficas. A transmissão ocorre por meio do contato com águas contaminadas, que afetam especialmente populações vulneráveis ou que entram em contato com águas por conta da pesca e da agricultura. A falta de saneamento básico, fatores demográficos, socioeconômicos, ambientais, presença do caramujo hospedeiro e o contato com corpos d'água são os principais riscos para a infecção por *Schistosoma mansoni* no Brasil (BEZERRA et al., 2021).

Causada pela infecção por vermes do gênero *Schistosoma*, a esquistossomose é uma doença parasitária crônica que pode se manifestar de duas formas, intestinal e urogenital, e depende da espécie com qual o ser humano foi infectado. A *S. mansoni*, *S. japonicum*, *S. mekongi*, *S. guineensis* e *S. intercalatum* são responsáveis por causar a esquistossomose intestinal e a *S. haematobium* por causar a urogenital (WHO, 2022). A evolução da doença da fase aguda para a crônica é caracterizada pela resposta imunológica à deposição cumulativa de ovos. Os ovos liberados pelas fêmeas nas veias mesentéricas migram para o fígado e ficam retidos no espaço porta. O granuloma se forma ao redor dos ovos, no espaço porta, levando à formação da fibrose periportal (FPP) ou fibrose de Symers. A FPP é característica da doença e pode levar à hipertensão portal e suas manifestações e complicações, como esplenomegalia, hiperesplenismo, formação de circulação colateral e sangramento digestivo. A FPP é considerada uma complicação grave e comum da fase crônica por *S. mansoni* e da hipertensão portal, estima-se que o tempo entre o início da infecção e a FPP avançada é de 5 a 15 anos (GUNDA et al., 2020; EWUZIE et al., 2025).

A FPP em estágio avançado pode levar à morbidade e mortalidade devido ao acúmulo de tecido fibrótico, que reduz a elasticidade das veias e contribui para a

obstrução do fluxo sanguíneo portal, resultando em hipertensão portal, formação de varizes esofágicas, sangramento gastrointestinal e, por fim, morte prematura (BARRETO et al., 2017; EWUZIE et a., 2025). Sendo assim, avaliar a presença e o grau da FPP é fundamental na prevenção de complicações da doença, permitindo desenvolver estratégias de tratamento, bem como monitorar a resposta à terapia (Barreto et al., 2022).

Entre as técnicas mais empregadas para o diagnóstico da FPP, está a ultrassonografia (US). Apesar da utilização de outras técnicas, como a tomografia computadorizada e a imagem por ressonância magnética, a US destaca-se devido ao seu baixo custo e sensibilidade comparável a da biópsia hepática. Entretanto, há limitações, principalmente no que diz respeito à disponibilidade de aparelhos em unidades de atenção primária em áreas endêmicas e à necessidade de um examinador especialista com experiência no diagnóstico e classificação ultrassonográfica da FPP, pelo protocolo de Niamey, preconizado pela Organização Mundial de Saude (HASHIM; BERZIGOTTI, 2021). Alternativas minimamente invasivas e de fácil acesso são uma necessidade, principalmente, em áreas endêmicas de difícil acesso. Sendo assim, este capítulo da tese será voltado à utilização de biomarcadores séricos obtidos através de exames de rotina na construção de modelos quimiométricos visando a discriminação entre FPP leve e avançada.

No que diz respeito a literatura sobre o uso de biomarcadores séricos na investigação minimamente invasiva da FPP, no trabalho de Barreto e colaboradores (2022), a contagem de plaquetas e o índice Coutinho foram apresentados como testes promissores para avaliar a FPP em áreas endêmicas para esquistossomose. Ambos os testes utilizam biomarcadores séricos e, além de serem simples e baratos, permitem uma interpretação objetiva dos resultados.

O índice Coutinho (IC) foi desenvolvido por Barreto e colaboradores (2017), que realizaram uma análise multivariada por regressão logística (LR) com sete biomarcadores. No modelo final, apenas a fosfatase alcalina (FAL) e a contagem de plaquetas (PLT) foram significativas na predição da FPP. O índice consiste em uma equação simples (27): a razão entre os valores de FAL e a contagem de PLT.

$$IC = \frac{(FAL/LSN)}{PLT} x100 \quad (27)$$

Apesar dos avanços no uso de biomarcadores séricos e da evolução observada nos estudos disponíveis na literatura, ainda há limitações para o diagnóstico preciso da FPP em diferentes estágios. No que diz respeito ao diagnóstico por US, variações entre equipamentos e diferenças na interpretação entre analistas continuam sendo problemas persistentes, que podem impactar diretamente a predição.

O aprendizado de máquina surge como um método capaz de processar grandes quantidades de dados complexos. Liu et al. (2024) apresentaram um dos poucos estudos na literatura que utilizaram algoritmos de aprendizado de máquina em conjunto com indicadores séricos de rotina para desenvolver um modelo preditivo de FPP, especificamente em pacientes infectados com *Schistosoma japonicum*. Os autores relataram um bom desempenho preditivo dos modelos testados. A proposta central foi explorar as vantagens dos métodos de aprendizado de máquina na análise de dados complexos, permitindo um diagnóstico mais preciso da fibrose hepática na esquistossomose.

Considerando as questões levantadas, a necessidade de alternativas para a investigação da FPP e a utilização de algoritmos de aprendizado de máquina e sua capacidade em lidar com dados complexos, o presente capítulo tem como objetivo avaliar o emprego de algoritmos de aprendizado de máquina para predizer FPP leve e avançada em pacientes infectados por *Schistosoma mansoni.* Ainda, para fins de complementação, a próxima seção será dedicada a uma análise na literatura dos estudos metabonômicos/metabolômicos sobre esquistossomose.

3.1. Ensaios Metabolômicos e Metabonômicos em estudos sobre esquistossomose

A busca por alternativas minimamente invasivas, além da US, que possam ser utilizadas na avaliação da FPP relacionada ao esquistossomo vem sendo relatada na literatura (BARRETO et al., 2017; RODRIGUES et al., 2022; LIU et al., 2024). Entre elas, estão modelos metabonômicos utilizando espectros de RMN de ^{1}H associados a aprendizado de máquina (RODRIGUES et al., 2022) e o uso de biomarcadores séricos, seja para o desenvolvimento de um índice (BARRETO et al., 2017) ou empregados em algoritmos de aprendizado de máquina (LIU et al., 2024).

No cenário mundial, em uma busca na literatura nos últimos 10 anos, a maior parte dos estudos metabonômicos/metabolômicos envolvendo esquistossomose foi conduzida no continente asiático (ZHU et al., 2017; HU et al., 2017; RONG et al., 2019;

HU et al., 2020; HUANG et al., 2020; CHIENWICHAI et al., 2022; ZHOU et al., 2023; Li et al.,2024; CHIENWICHAI et al.,2024), seguido pelo Brasil (GOUVEIA et al., 2017; LOYO et al., 2021; RODRIGUES et al., 2022), pela África (TAWANA-NDOLO et al., 2023) e pelos Estados Unidos (CORTES-SELVA et al., 2021). Uma característica recorrente nesses trabalhos é o número reduzido de amostras analisadas. Exceção a esse padrão é o estudo de Tawana-Ndolo e colaboradores (2023), que investigou a infecção por *Schistosoma haematobium* em 527 crianças. Nos demais trabalhos, mesmo aqueles realizados em camundongos, o número de participantes não ultrapassou 70.

Considerando a diversidade de amostras que podem ser utilizadas em estudos metabonômicos, como soro, urina, fezes, tecidos e até medula óssea, a maioria das investigações envolvendo esquistossomose concentrou-se na análise de soro e urina. Para a caracterização desses biofluidos, diferentes técnicas analíticas podem ser empregadas, destacando-se a cromatografia líquida como a mais utilizada nos trabalhos envolvendo a esquistossomose (HU et al., 2017; RONG et al., 2019; HU et al., 2020; HUANG et al., 2020; CHIENWICHAI et al., 2022; ZHOU et al., 2023; Li et al.,2024; CHIENWICHAI et al.,2024).

No que diz respeito às análises estatísticas multivariadas aplicadas aos 14 estudos encontrados envolvendo metabolômica/metabonômica e esquistossomose, a maioria empregou os métodos PLS-DA ou OPLS-DA, em conjunto com a PCA, enquanto apenas um deles utilizou apenas a PCA (LOYO et al., 2021). As investigações resultaram em modelos com capacidade preditiva (Q^2) variando de 0,5 até 0,991. Este último valor foi relatado por Rong e colaboradores (2019), cujo modelo também apresentou bom ajuste segundo os valores de R^2, reforçando o potencial da metabonômica como alternativa promissora para o diagnóstico precoce e a identificação de biomarcadores associados à esquistossomose. Entre os trabalhos que avaliaram seus modelos por meio da curva ROC, Tawana-Ndolo e colaboradores (2023) obtiveram uma AUROC de 0,875, resultado considerado promissor pelo que vem sendo reportado na literatura.

A possibilidade de identificação de biomarcadores com potencial clínico em estudos metabonômicos é uma grande ferramenta na investigação de alterações significativas nas vias metabólicas durante a infecção por *Schistosoma*. Cerca de 30 metabólitos, entre os trabalhos dos últimos 10 anos na literatura, com *S. mansoni* e *S. japonicum* foram identificados como alterados no soro ou urina de humanos e ratos

infectados. A análise de aminoácidos como resultados da infecção por ambas as espécies revela efeitos significativos no metabolismo de proteínas e aminoácidos de cadeia ramificada. Outro metabólito relatado foi o gliceraldeído, envolvido no metabolismo de aldeídos, reflete alterações bioquímicas associadas à fibrose hepática, particularmente em infecções causadas por *S. mansoni* (RODRIGUES et al., 2022).

A análise das alterações metabólicas, incluindo as do ciclo do TCA, metabolismo de aminoácidos e lipídios, é essencial para compreender a relação entre alterações bioquímicas e manifestações clínicas da esquistossomose, como fibrose hepática e disfunções imunológicas. A complexidade das interações hospedeiro-parasita é evidenciada pela compreensão dos mecanismos patogênicos envolvidos e destacam a necessidade de pesquisas futuras que integrem dados metabolômicos, genômicos e imunológicos. Avanços metodológicos e conceituais nesse âmbito têm o potencial de melhorar o manejo de doenças induzidas por *Schistosoma* e ampliar a compreensão das perturbações metabólicas associadas a outras infecções parasitárias, oferecendo novas oportunidades para intervenções clínicas (RODRIGUES et al., 2025).

Em suma, no que diz respeito a metabonômica, uma parte dos artigos encontrados na literatura se concentraram na identificação das potenciais vias metabólicas associadas à infecção, considerando o local da doença, enquanto outros buscaram classificar e monitorar a progressão da doença. A presente seção foi escrita com o objetivo de trazer uma visão geral sobre ensaios metabolômicos e metabonômicos na literatura envolvendo esquistossomose. Na sequência, a continuidade do capítulo será voltada para o uso de biomarcadores séricos e algoritmos de aprendizado de máquina para o estadiamento de FPP. Para informações mais detalhadas sobre a abordagem metabonômica na esquistossomose, recomenda-se a leitura integral do artigo de revisão *“Metabolomics assays applied to schistosomiasis studies: a scoping review”* (RODRIGUES et al., 2025), produzido no contexto deste trabalho.

3.2. Objetivos Específicos

- Investigar o uso de biomarcadores séricos (AST, ALT, FAL, GGT e PLT) na avaliação da fibrose periportal na esquistossomose mansoni.
- Desenvolver modelos quimiométricos baseados em algoritmos de aprendizado de máquina para classificação das formas leves e avançadas da fibrose periportal, utilizando a ultrassonografia como método de referência.
- Avaliar o desempenho dos modelos com base nas figuras de mérito e a importância de cada biomarcador sérico na discriminação das classes.

3.3. Materiais e Métodos

3.3.1. Amostragem

O presente estudo avaliou dados obtidos de exames laboratoriais do banco de dados do Ambulatório de Esquistossomose do HC/UFPE (HC) e da cidade de Jaboatão dos Guararapes (JG) que fazem parte da rotina de solicitações médicas para os pacientes atendidos. Os biomarcadores séricos analisados foram o aspartato aminotransferase (AST), alanina aminotransferase (ALT), fosfatase alcalina (FAL) e gama-glutamil transferase (GGT), além da contagem de plaquetas (PLT). Os pacientes são homens e mulheres residentes do estado de Pernambuco com idade entre 18 e 80 anos infectados pelo Schistosoma Mansoni com FPP evidente, recrutados nos períodos de setembro de 2015 a agosto de 2016 e março de 2019 a 2022.

O diagnóstico foi baseado na história clínica de contato com fontes de água em áreas endêmicas, relatos de tratamento com praziquantel, juntamente com os achados no US de FPP. Como critério de exclusão, estão pacientes esplenectomizados, doença hepática de outra etiologia, etilistas, pacientes com marcadores dos vírus HIV ou das hepatites B ou C, ou outras doenças hepáticas.

No que diz respeito ao tamanho da amostra, inicialmente, o banco de dados, composto por informações adquiridas no município de Jaboatão dos Guararapes e no HC-UFPE, continha 288 pacientes, sendo 172 mulheres e 116 homens. A distribuição dos padrões de FPP foi a seguinte: grupo AB = 18, grupo C = 93, grupo D = 81 e grupo E/F = 96. Para o presente trabalho, 184 pacientes foram incluídos, sendo 108 mulheres e 76 homens. A FPP foi classificada de acordo com o protocolo de Niamey

em padrão C (periférico, leve) e padrões E ou F (central e periférico, avançada ou muito avançada).

3.3.2. Análise Quimiométrica

A análise quimiométrica do conjunto de dados, incluindo as etapas de pré-processamento, visualização, análise exploratória e modelos de classificação, foram desenvolvidos em linguagem *Python 3*, utilizando o ambiente interativo de programação *Google Colaboratory (Colab),* uma plataforma gratuita oferecida pelo Google para execução em *Jupyter Notebook.* Para a execução dos algoritmos de aprendizados de máquina e todas as etapas necessária, uma série de bibliotecas foi utilizada:

Scikit-learn (PEDREGOSA et al., 2011), *Numpy* (HARRIS *et al.*, 2020), *Pandas* (MCKINNEY, 2010), *seaborn* (WASKOM, 2021), *scipy* (VIRTANEN *et al.*, 2020), *matplotlib* (HUNTER, 2007), *statsmodels* (SEABOLD; PERKTOLD, 2010) e *tqdm* (MATIYASEVICH, 2015).

Inicialmente, foram detectados dois valores ausentes que foram tratados com a imputação da mediana de cada classe, em seguida, os dados foram autoescalados para a PCA e futuras etapas de classificação.

Para os modelos de classificação, as amostras foram divididas em conjunto de treinamento e teste, 70% e 30%, respectivamente, utilizando o algoritmo Kennard-Stone, que seleciona as amostras com base nas distâncias entre elas. A reprodutibilidade dos modelos foi garantida com o parâmetro r*andom_state*. Em seguida, os algoritmos de aprendizado de máquina, LDA, SVM, LR e Árvore de Decisão, tiveram seus parâmetros otimizados com o *gridsearchCV*, treinados com 128 amostras e validados com 56 amostras de teste. O desempenho dos modelos de classificação foi avaliado utilizando seis figuras de mérito calculadas a partir da matriz de contingência obtida da validação do conjunto de teste. As figuras de mérito foram: exatidão (Eq. 21), VPP (Eq. 19), VPN (Eq. 20), sensibilidade (Eq. 17), especificidade (Eq. 18), *F1-Score* (Eq. 22), Kappa (Eq. 23) e AUROC. O teste de permutação foi empregado em todos os modelos de classificação para avaliar a significância estatística da exatidão dos modelos.

A escolha dos algoritmos de aprendizado de máquina foi realizada com o objetivo de testar tanto classificadores lineares quanto não lineares. Além disso, foram

consultados trabalhos da literatura e do próprio grupo de pesquisa em que este estudo está sendo desenvolvido.

3.3.3. Considerações Éticas

O projeto foi aprovado pelo Comitê de Ética em Pesquisa (CEP) do Hospital das Clínicas da UFPE/ EBSERH, sob o parecer 4.465.533. A pesquisa envolvendo os dados dos pacientes de Jaboatão dos Guararapes-PE também passou pelo comitê de ética e pesquisa sob o parecer 5330740 e foram cedidos após a publicação dos dados (OZAKI et al., 2024).

3.4. Resultados e Discussão

3.4.1. Dados Clínicos

Foram incluídas 184 amostras, classificadas em dois grupos conforme os critérios ultrassonográficos do Padrão de Niamey: padrão C, correspondente à forma leve de fibrose periportal, e padrões E/F, representando os casos de forma avançada. A Tabela 10 apresenta as características demográficas e os parâmetros clínico-laboratoriais dos pacientes incluídos, estratificados de acordo com a gravidade da fibrose.

Tabela 10 - Características demográficas e parâmetros clínico-laboratoriais de acordo com a gravidade da esquistossomose mansoni.

	Grupo I (Padrão C)	Grupo II (Padrão EF)	p-valor
n	88	96	-
Idade (anos ± DP)	(41,4 ± 15,0)	(51,2 ± 14,5)	0,00001*
Sexo (M/F)	(35,2%/64,8%)	(46,9%/53,1%)	0,146**
AST (média ± DP)	(13,8 ± 10,1)	(28,3±20,3)	< 0,001*
ALT (média ± DP)	(23,0 ± 20,5)	(27,3 ± 21,5)	0,16*
FAL (média ± DP)	(105,0 ± 171,9)	(178,1 ± 142,6)	< 0,0001*
GGT (média ± DP)	(34,7 ± 46,1)	(80,0 ± 78,9)	< 0,0001*
PLT (média ± DP)	(249 ± 100)	(108,8±80,6)	< 0,0001*

*Teste de Mann-Whitney/**Teste qui-quadrado

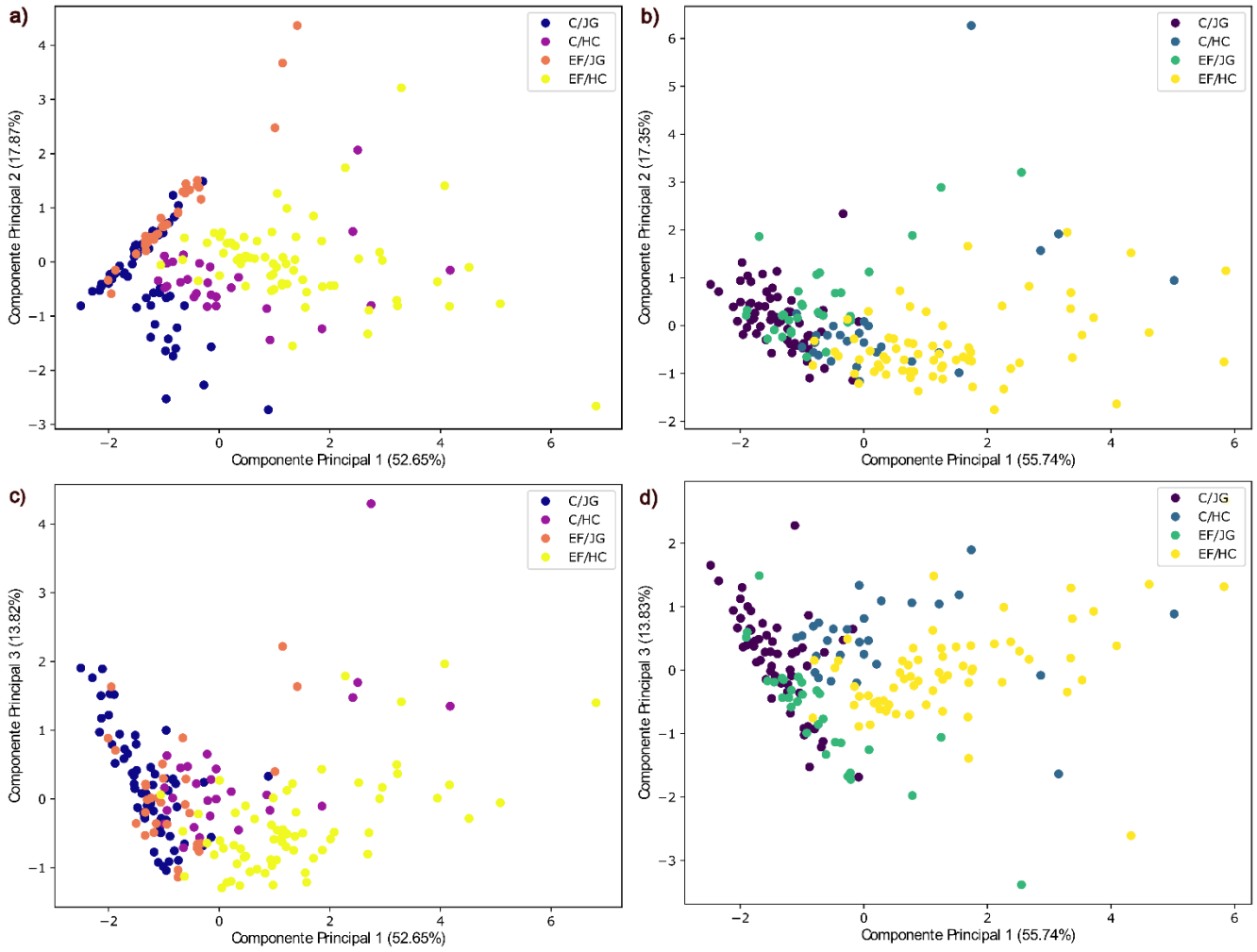
A análise estatística apresentada na Tabela 10 evidenciou diferenças significativas entre os grupos com padrão leve (C) e avançado (EF) da FPP em pacientes com esquistossomose mansoni. Utilizando-se o teste de Mann-Whitney para as variáveis contínuas e o teste qui-quadrado para a variável categórica, observou-se associação entre as formas avançadas da doença e maior idade, elevação das enzimas AST, FAL e GGT, além de uma redução significativa na contagem plaquetária. Por outro lado, embora a ALT tenha apresentado valores discretamente mais elevados no grupo com fibrose avançada, essa diferença não atingiu significância estatística.

3.4.2. Visualização dos dados

O presente estudo clínico visa avaliar o emprego de modelos preditivos para discriminar FPP, leve e avançada/muito avançada, baseados em dados clínicos de pacientes infectados pelo *Schistosoma mansoni*. Com o intuito de obter um método de triagem clínica e fornecer informações de como as vias metabólicas dos biomarcadores séricos estão associadas ao desenvolvimento dessa condição, a matriz de dados composta por 184 amostras e cinco variáveis (AST, ALT, FAL, GGT e PLT) passou por etapas de pré-processamento antes do processo de modelagem. Inicialmente, foi realizada a busca por valores ausentes, comuns em matriz de dados de análises clínicas, na qual, dois valores estavam ausentes e foram tratados com a substituição pela mediana, uma vez que é um método robusto à presença de outliers e distribuições assimétricas (ALAM et al., 2023). Em seguida, as variáveis foram padronizadas por autoescalamento antes de qualquer modelo empregado, supervisionado e não supervisionado.

É comum na rotina clínica que esses valores de AST, ALT, FAL, GGT e PLT sejam normalizados pelo limite superior da normalidade (LSN), devido a diferenças entre centros e kit análises. Sendo assim, a PCA foi empregada com os valores com e sem a normalização pelo LSN com o objetivo de investigar a distribuição das amostras considerando ambos os tratamentos. A Figura 33 apresenta os gráficos de escores como resultado da PCA.

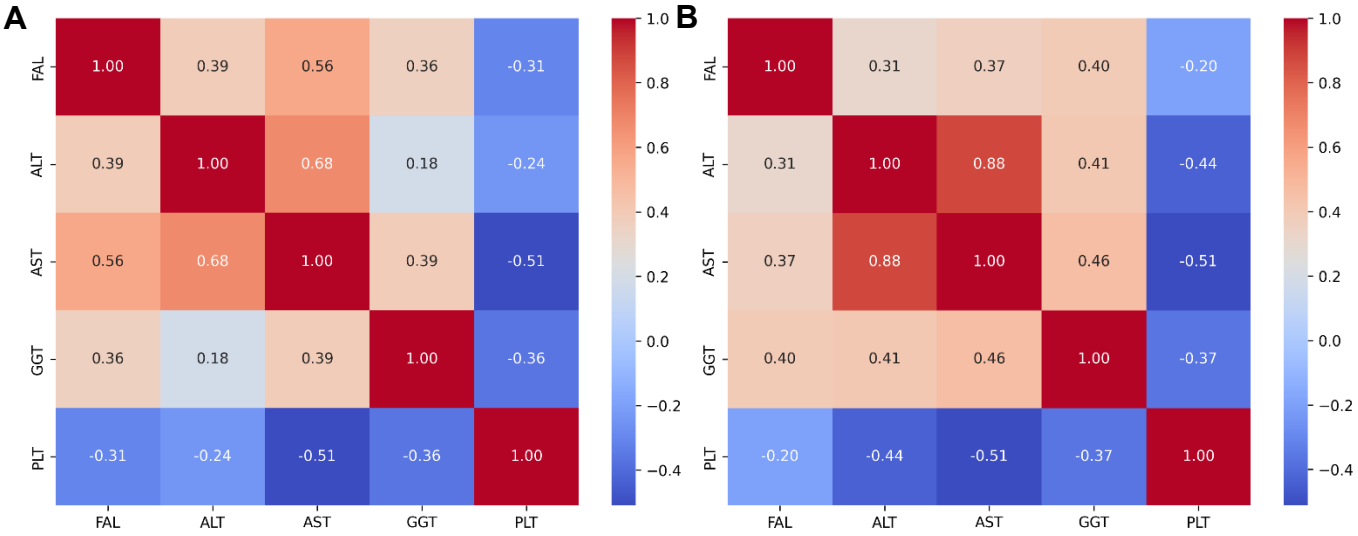
Figura 33 - Gráfico de escores comparando a distribuição das amostras de diferentes centros: Hospital das Clínicas (C/HC e EF/HC) e Jaboatão dos Guararapes (C/JG e EF/JG). (a) PC1 vs PC2 – sem normalização; b) PC1 vs PC2 – com normalização LSN; c) PC1 vs PC3 – sem normalização; d) PC1 vs PC3 – com normalização LSN.



Fonte: A autora (2025)

Foi possível observar que, tanto no conjunto de amostras normalizadas quanto no não normalizado, os padrões de agrupamento das amostras do mesmo grupo se mantêm, mesmo quando provenientes de centros diferentes. Pode-se chamar atenção para a Figura 33. a), onde é possível observar uma melhor tendência de separação entre os grupos C e EF do que nas demais. Ainda, com a normalização pelo LSN, a matriz de dados apresentou duas variáveis altamente correlacionadas, ou seja, que carregam informações semelhantes. Nas matrizes de correlação da Figura 34, é possível observar que nos dados não normalizados essa correlação entre ALT e AST diminuiu.

Figura 34 - Matriz de correlação. A) Sem normalização; B) Com normalização – LSN.



Fonte: A autora (2025)

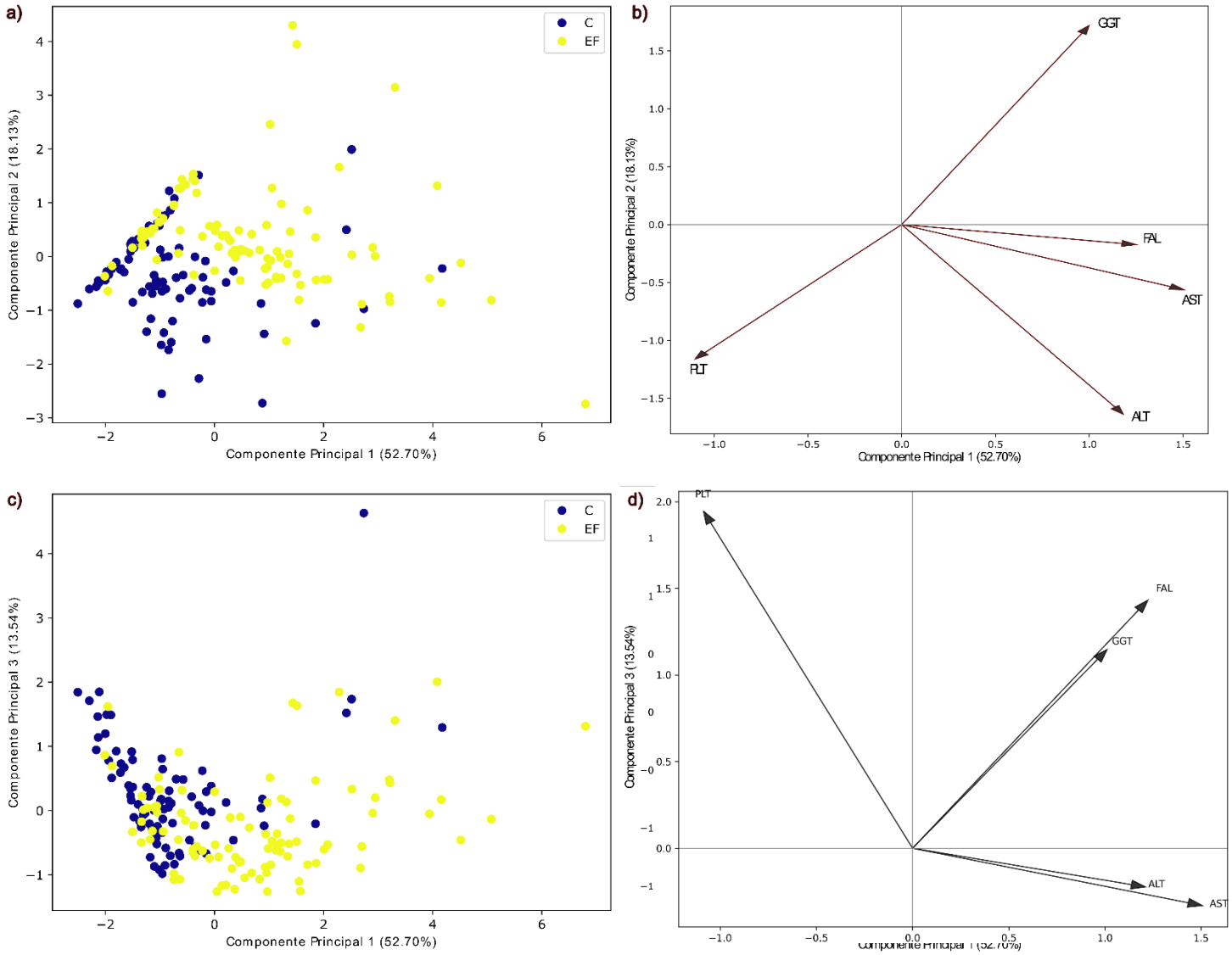
A normalização reduziu a variação observada para cada parâmetro. Ao observar os valores médio de ALT (0,60±0,56) e AST (0,62±0,51), esse comportamento fica mais claro. Sendo essas duas variáveis altamente correlacionadas, uma delas precisaria ser removida para evitar o risco de multicolinearidade nos modelos de classificação. Entretanto, considerando que o presente estudo visa investigar a influência dos cinco biomarcadores séricos e como os resultados da PCA não mostraram diferença entre os tratamentos, as análises seguiram com a matriz de dados sem a normalização pelo LSN.

Seguindo com análise da PCA, visando identificar a ocorrência de amostras anômalas e se há agrupamento natural nas classes de interesse, é possível observar uma dispersão das amostras principalmente em relação ao grupo EF. As três primeiras componentes principais explicam cerca de 84% da variação dos dados. A Figura 35 apresenta os gráficos de escore e de pesos da PCA para as três primeiras componentes, indicando uma tendência de agrupamento das amostras classificadas como grupo C com escores negativos nas três componentes principais.

O gráfico de pesos da PC1 e PC2 (Figura 35. b) indica que as amostras do grupo C apresentam níveis séricos de PLT mais elevados, já as amostras do grupo EF apresentam níveis séricos de GGT, FAL, AST e ALT mais elevados e estão associados à sua maior dispersão no sentido positivo da PC1, com exceção do GGT e ALT que também estão a maior dispersão nos sentidos positivo e negativo, respectivamente,

de PC2. No que diz respeito ao gráfico de pesos da PC1 e PC3 (Figura 35. d), FAL, GGT e PLT estão fortemente associados a dispersão das amostras no sentido da PC3.

Figura 35 - PCA. a) Gráfico de escores – PC1 vs PC2; b) Gráfico de pesos – PC1 vs PC2; c) Gráfico de escores – PC1 vs PC3; d) Gráfico de pesos – PC1 vs PC3.



Fonte: A autora (2025)

Considerando os resultados observados por meio da PCA, a etapa seguinte foi construir os modelos de classificação.

3.4.3. Modelos de Classificação

Foram aplicados quatro formalismos quimiométricos para modelar o conjunto de dados: a Regressão Logística (LR, do inglês, *Logistic Regression*), a Análise Discriminante Linear (LDA, do inglês, *Linear Discriminant Analysis*), o Máquina de

Vetores de Suporte (SVM, do inglês, *Support Vector Machine*) e Árvore de Decisão (DT, do inglês, *Decision Tree*). Para isso, o conjunto de 184 amostras foi divido em um grupo de treinamento, contendo 128 amostras, e grupo de teste, contendo 56 amostras. Os modelos foram treinados e validados com o grupo de teste, a exatidão também foi avaliada por validação cruzada (*kfold* = 10) no conjunto de treino. A significância do desempenho dos modelos foi avaliada com o teste de permutação, com 1000 permutações cada teste e resultado expresso em p-valor. A matriz de contingência para cada um dos modelos e os resultados da validação estão apresentados nas Tabelas 11 e 12, respectivamente.

Tabela 11 - Matriz de contingência dos modelos LR, SVM, LDA e DT.

		Diagnóstico Padrão	
		EF	**C**
LR	**EF**	18	4
	C	10	24
SVM	**EF**	18	4
	C	10	24
LDA	**EF**	19	5
	C	9	23
DT	**EF**	21	1
	C	7	27

Tabela 12 - Figuras de mérito calculadas para cada modelo.

Figuras de Mérito	***LR***	***SVM***	***LDA***	***DT***
Exatidão*	79%	78%	81%	81%
Exatidão	75%	75%	75%	86%
Sensibilidade	64%	64%	68%	75%
Especificidade	86%	86%	82%	96%
VPP	82%	82%	79%	96%
VPN	71%	71%	72%	79%
F1-score	72%	72%	73%	84%
AUROC	0,879	0,875	0,884	0,875
Kappa	0,500	0,500	0,500	0,714
Valor de p*	0,052	0,002	0,045	0,000

*Validação Cruzada
**Teste de permutação

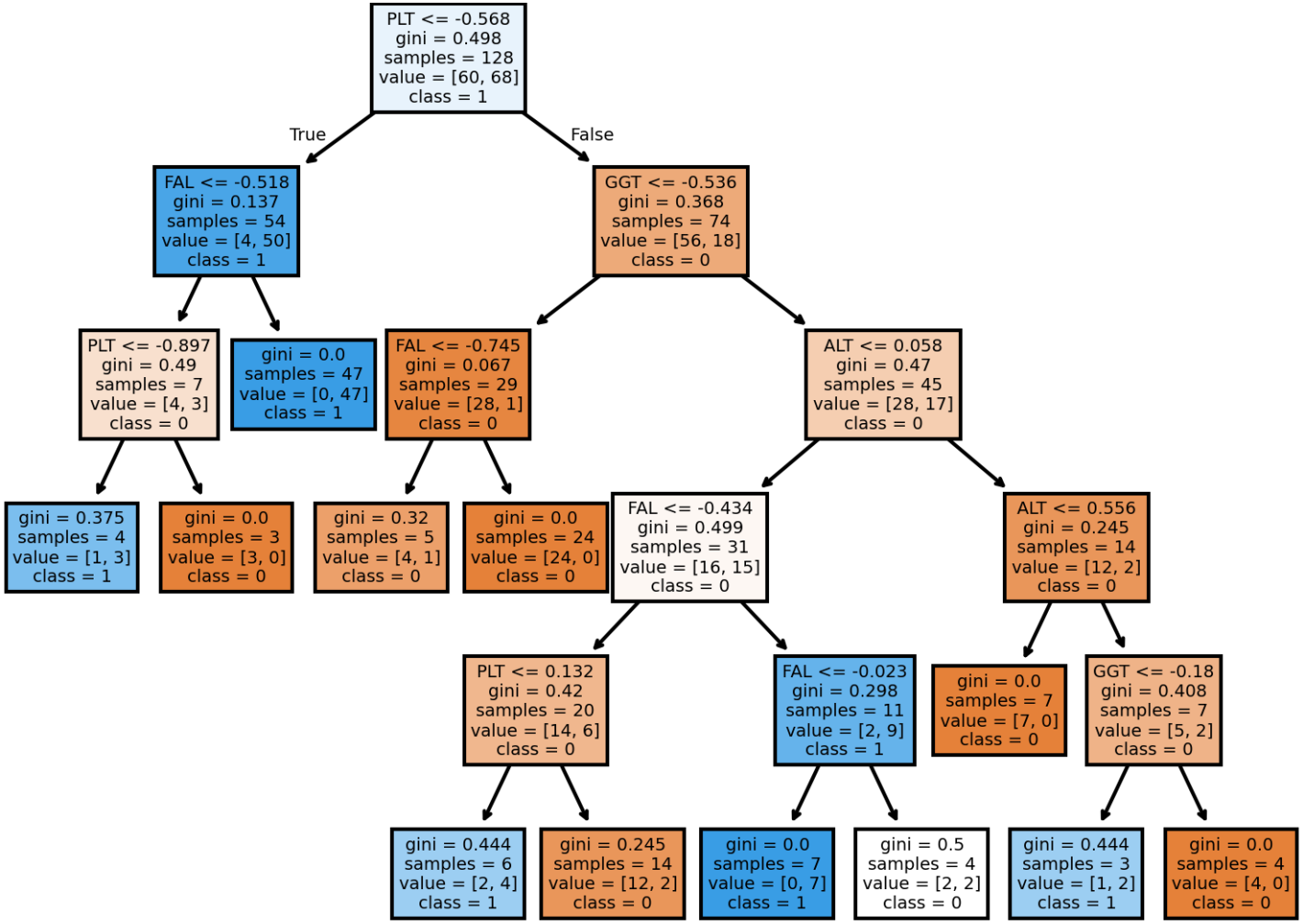
Inicialmente, três modelos paramétricos lineares de aprendizado de máquina foram empregados: LR, LDA e SVM (com kernel linear). Os dois primeiros modelos apresentaram o mesmo desempenho na classificação das amostras do conjunto de teste, entretanto, a regressão logística (LR) não passou no teste de permutação (p-

valor > 0,05). O modelo LDA classificou corretamente a mesma quantidade de amostras dos modelos anteriores, o que resultou em uma exatidão de 75% nos três. No entanto, a LDA acertou uma amostra a mais no grupo com FPP avançada.

É importante destacar que os três modelos apresentaram maior taxa de erro na classificação das amostras do grupo com FPP avançada, resultando em valores de sensibilidade na faixa de 60%.

Como alternativa não paramétrica, o algoritmo de árvore de decisão (DT) foi avaliado, apresentando o melhor desempenho de classificação entre os modelos testados e, consequentemente, os melhores resultados nas figuras de mérito, especialmente nos valores de especificidade, valor preditivo positivo (VPP) e índice Kappa. Este último indicou uma concordância substancial entre a predição do modelo e a classificação real, considerando também a possibilidade de acerto ao acaso. As áreas sob as curvas ROC apresentaram valores muito próximos entre os quatro modelos em torno de 0,88, indicando uma boa capacidade dos modelos em distinguir amostras positivas de negativas. A Figura 36 traz a DT para a classificação das amostras de FPP, construída com profundidade máxima igual a 5 e número mínimo de amostras por folha igual a 3.

Figura 36 - Árvore de decisão para classificação da FPP. Em cada nó encontra-se a impureza de Gini, o número de amostras, a distribuição por classes: [0 (C), 1 (EF)], e a classe com a maioria das amostras.



Fonte: A autora (2025).

A profundidade da árvore (max_depth = 4) e o valor mínimo de 3 amostras por folha favoreceram a geração de um modelo robusto, evitando segmentações que poderiam se ajustar demais ao ruído dos dados de treino e impedindo a formação de decisões com base em pequenos grupos. Ainda, foi possível identificar os erros ao observar os nós terminais, onde 9 amostras, de um total de 128, foram classificadas de forma incorreta.

No que diz respeito à literatura, são limitados os trabalhos que empregam biomarcadores séricos a modelos de aprendizado máquina para prever FPP. Entre eles, o estudo de Liu e colaboradores (2024) usaram exames de rotina sanguínea e informações básicas de pacientes com *Schistosoma japonicum* para estabelecer um modelo de aprendizado de máquina capaz de prever precocemente a FPP. Seis algoritmos diferentes de aprendizado de máquina foram avaliados para estabelecer os modelos de predição, onde, a LR, SVM e KNN apresentaram os piores desempenhos, com a AUC menor que 0,75. No caso dos autores, o LightGBM foi o algoritmo que melhor classificou as amostras, com a AUC de 0,84. Ao comparar com o melhor modelo do presente trabalho, pode-se afirmar resultados estão de acordo com o que vem sendo encontrado na literatura (Tabela 13)

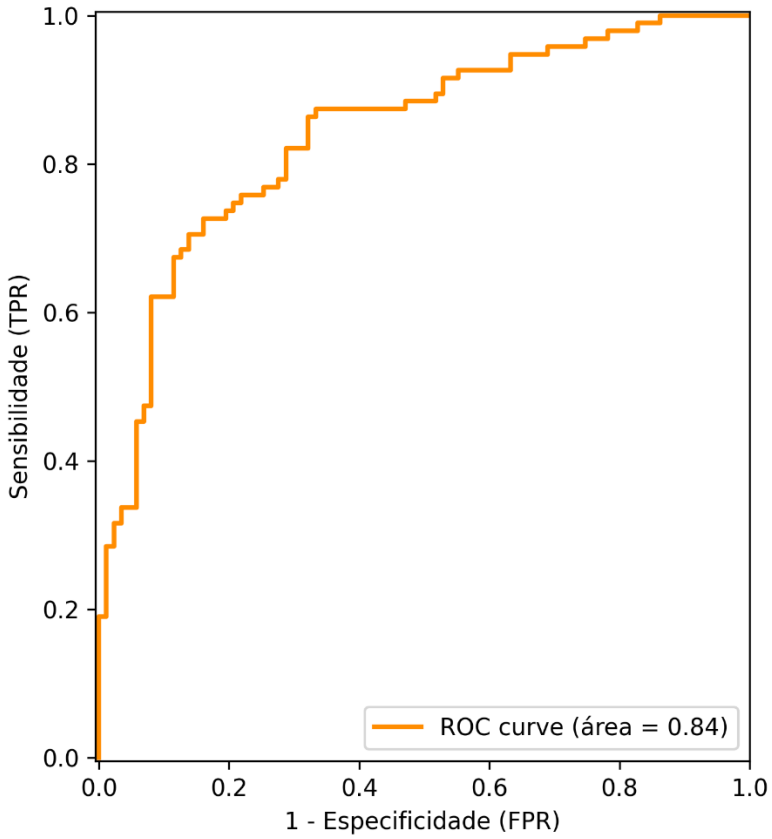
Tabela 13 - Figuras de mérito do melhor modelo de classificação do presente trabalho e do estudo de Liu e colaboradores (2024).

Figuras de Mérito	***DT***	***LightGBM (Liu et al., 2024)***
Exatidão	86%	81%
Sensibilidade	75%	71%
Especificidade	96%	84%
VPP	96%	84%
VPN	79%	80%
F1-score	84%	77%
Kappa	0,71	0,394

Além dos modelos de aprendizado de máquina, o índice de Coutinho foi calculado para as amostras e, em seguida, uma curva ROC foi construída de C *vs.* EF, utilizando o índice para discriminar as duas classes e avaliando, por meio da área sob a curva, quão bem as amostras são discriminadas em diferentes limiares. A Figura 37 apresenta o gráfico da curva ROC, bem como o valor da área sob a curva, que foi de 0,84, valor próximo ao dos modelos apresentado neste presente trabalho. O limiar que apresentou melhor relação entre sensibilidade e especificidade foi em 0,351. Com esse valor estabelecido, foi possível gerar uma matriz de contingência, na qual os

valores menores que 0,351 foram classificados como C e valores maiores que 0,351 foram classificados como EF (Tabela 14).

Figura 37 - Curva ROC: índice de Coutinho



Fonte: A autora (2025)

Tabela 14 - Matriz de contingência e figuras de mérito: Índice de Coutinho.

	Diagnóstico padrão		Cutoff (Coutinho-index) = 0,351 AUC = 0,84 Exatidão 78% Sensibilidade 72% Especificidade 84% VPP 83% VPN 74% F1-Score 77%
	EF	C	
EF	68	14	
C	26	73	

Os resultados referentes ao desempenho do índice de Coutinho estão entre os achados encontrados no presente trabalho utilizando aprendizado de máquina. No trabalho de Barreto e colaboradores (2022), o índice de Coutinho foi capaz de prever a FPP avançada na maioria dos indivíduos avaliados. Com ponto de corte ≥ 0,316, o estudo revelou uma curva AUROC de 0,70, sensibilidade de 67,4% e especificidade de 68,3%, indicando um desempenho inferior. Os autores obtiveram esses resultados

comparando pacientes sem FPP (A+B) com pacientes de FPP avançada (EF), apesar da maior diferença entre os pacientes, o desempenho do índice, neste caso, pode ser atribuído ao menor tamanho da amostra (n=87) utilizada.

É importante destacar que tanto os modelos de classificação quanto o índice Coutinho, fazem parte da busca por alternativas pouco invasivas capazes de diagnosticar e/ou estadiar a presença de FPP, principalmente em suas formas graves. Os modelos de classificação apresentados, com destaque para LDA e DT, obtiveram bons desempenho e demonstraram robustez ao classificar as amostras com FPP leve e avançada/muito avançada. Eles podem, portanto, ser vistos como uma alternativa para a discriminação de pacientes nessas condições, principalmente, por utilizarem biomarcadores séricos que estão disponíveis por meio de exames laboratoriais de rotina, facilitando o acesso de pacientes de áreas endêmicas e rurais.

3.4.4. Importância das Variáveis

O índice de Coutinho foi desenvolvido por Barreto e colaboradores (2017) por meio de um modelo de regressão logística (LR) utilizando 7 biomarcadores séricos na construção do modelo, entre eles, estavam os 5 avaliados no presente trabalho (AST, ALT, FAL, GGT e PLT). Os autores chegaram em um modelo final onde apenas a FAL e o número de PLT foram variáveis significativas para a FPP.

Com o intuito de investigar como essas variáveis se comportaram no presente trabalho, esta seção será dedicada a apresentar e avaliar suas importâncias para a construção dos modelos LDA e DT.

Iniciando pela LDA, para entender como cada variável influenciou na construção do modelo, foram usados os coeficientes absolutos (Tabela 15).

Tabela 15 - Importância das variáveis no modelo LDA.

Variável	**Coeficiente - LDA**
FAL	-0,051357
ALT	-0,552133
AST	0,782260
GGT	0,289840
PLT	-1,753637
Intercepto	0,0436

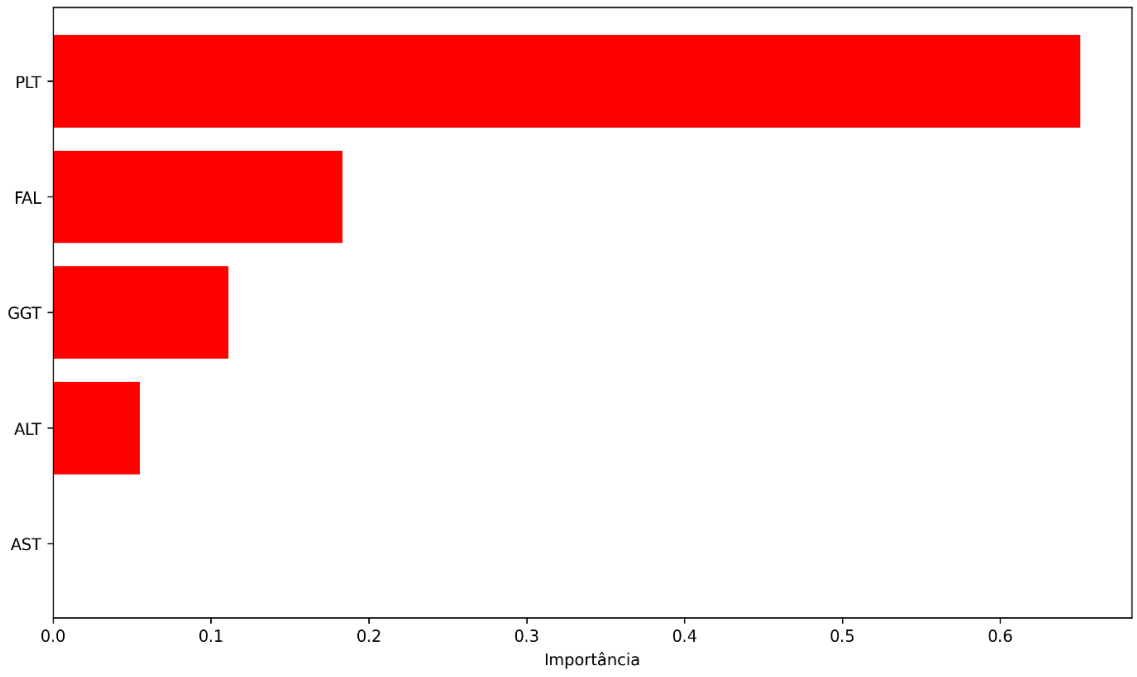
O indicador sérico que apresenta maior influência na construção do modelo é o número de PLT, apresentando o maior coeficiente entre os demais. Como o modelo

LDA foi construído de tal forma que os voluntários com padrão C tenham escores menores que aqueles com padrão EF, os níveis séricos mais elevados de plaquetas (PLT), alanina aminotransferase (ALT) e fosfatase alcalina (FAL) tendem a diminuir o escore LDA para a amostras, classificando-a no grupo C. A contagem de PLT tem sido usada como indicador da hipertensão portal, varizes esofágicas e sangramento do trato digestivo superior, sintomas da FPP avançada, que resultam na redução do número de PLT (BARRETO et al., 2021; LEITE et al., 2013).

Por outro lado, níveis séricos elevados de aspartato aminotransferase (AST) e gama glutamil transferase (GGT) aumentam o escore LDA da amostra, classificando-a no grupo EF. O segundo indicador sérico mais importante foi a AST, uma enzima encontrada no fígado, mas também presente em outros tecidos do corpo humano, por isso, seus níveis anormais devem ser interpretados com cautela. Em casos de lesão hepatocelular, ocorre liberação de aminotransferases pelos hepatócitos, resultando em níveis séricos elevados (KALAS et al., 2021). Apesar de se esperar que na FPP a função hepatocelular esteja preservada, alterações nas enzimas hepáticas podem refletir danos funcionais a nível celular (LEITE et al., 2013; SILVA et al., 2018).

Seguindo com as investigações, o modelo gerado com o algoritmo DT não gera coeficientes assim como o LDA, mas fornece as variáveis mais importantes na sua construção. A Figura 38 apresenta as variáveis mais importantes na discriminação dos pacientes de grau leve dos avançados.

Figura 38 - Importância das variáveis no modelo DT.



Fonte: A autora (2025)

Nesse caso, o número de PLT, assim como na LDA, apresenta a maior influência na discriminação dos grupos e, em segundo lugar, a FAL assume a posição. O número de PLT se destaca como um parâmetro essencial no processo de classificação. É válido notar que a importância das variáveis FAL e AST estão o inverso de como elas se comportaram na LDA. A enzima FAL é um biomarcador sérico importante para avaliar anormalidades colestáticas, sendo útil no diagnóstico de doenças hepáticas crônicas, seus níveis séricos são utilizados em investigações de trabalhos na literatura para prever fibrose (hepática e periportal) significativa e apresentam resultados promissores (LEITE et al., 2013; HU et al., 2019; ZENG et al., 2024). Esse resultado tem relação direta com o índice de Coutinho no qual FAL e PLT são utilizados e contribuem de forma semelhante para a classificação.

3.5. Conclusão do Estudo 3 – Estadiamento de Fibrose Periportal

Este trabalho teve como objetivo investigar o uso de biomarcadores séricos na construção de algoritmos de aprendizado de máquina aplicados à progressão da FPP para formas mais graves, em pacientes com *Schistosoma mansoni*. Quatro algoritmos foram avaliados: Regressão Logística (LR), Máquina de Vetores de Suporte (SVM), Análise Discriminante Linear (LDA) e Árvore de Decisão (DT). Entre eles, apenas o LR não apresentou desempenho significativo no teste de permutação.

Ao avaliar os três modelos restantes, o algoritmo de Árvore de Decisão apresentou os melhores resultados com exatidão de 86%, valor predito positivo (VPP) de 96% e coeficiente Kappa de 0,710, se destaca entre os demais modelos, evidenciando sua robustez na classificação. Ao avaliar as variáveis importantes para os modelos LDA e DT, foi possível ver como os biomarcadores séricos influenciaram na construção dos modelos, com atenção para PLT e FAL que apresentaram maior importância para construção do modelo DT, e estão relacionados diretamente no cálculo do índice de Coutinho, o que reforça sua relevância clínica.

Por fim, pode-se concluir que os modelos apresentados no presente trabalho apresentaram resultados promissores na discriminação de pacientes com formas leve e avançada de FPP. Tais achados sugerem que o uso de algoritmos de aprendizado de máquina configura-se como mais uma ferramenta a ser explorada e investigada para o monitoramento da FPP, especialmente em áreas endêmicas e com difícil acesso aos métodos padrão.

4. Conclusão

O presente estudo teve como objetivo desenvolver, a partir de ensaios metabonômicos e quimiométricos, modelos para o diagnóstico de Lesão Renal Aguda em recém-nascidos, para a discriminação de Nefrite Lúpica proliferativa associada ou não à membranosa e para a discriminação entre Fibrose Periportal leve e avançada, em pacientes com esquistossomose mansoni. O emprego dos algoritmos de aprendizado de máquina seguiu por dois caminhos, um deles na abordagem metabonômica em pacientes com Lesão Renal Aguda e Nefrite Lúpica, enquanto o outro se baseou em parâmetros bioquímicos séricos de pacientes com Fibrose Periportal.

Para todas as doenças investigadas, os modelos de classificação empregados demonstraram alta acurácia, com destaque para os modelos SVM e LR. Além disso, foi possível identificar possíveis metabólitos associados às condições investigadas, os quais revelam informações sobre vias metabólicas perturbadas e oferecem novas perspectivas sobre a doença.

Os resultados indicam que a hipótese de que o perfil de metabólitos endógenos presentes em biofluidos de uma pessoa se altera em função de seu status clínico (doente/saudável) mostrou-se correta. Essa perturbação, acessada por meio de ferramentas quimiométricas aplicadas a dados espectrais obtidos desses biofluidos, permitiu investigar sistemas complexos de forma minimamente invasiva, oferecendo ferramentas promissoras para o diagnóstico e estadiamento de doenças renais e hepáticas, mesmo diante da limitação no número de amostras.

Perspectivas

Os resultados do presente estudo demonstraram-se promissores, o emprego de modelos metabonômicos e quimiométricos na investigação de doenças renais e hepáticas, apresentaram bons desempenhos. Entretanto, assim como observado na literatura, o presente estudo encontrou limitações relacionadas ao número de amostras e, mesmo com a aplicação da sobreamostragem, ainda seria importante trabalhar com conjuntos maiores para avaliar melhor a robustez dos modelos.

Aumentar o número de amostras segue como uma das principais perspectivas futuras, não apenas para continuidade deste estudo, mas também para trabalhos futuros. Além disso, três artigos, um em etapa de submissão e dois em produção, são esperados como resultado desta tese, somando-se ao artigo de revisão já publicado.

Como projeto futuro, pretende-se analisar amostras de pacientes com doença renal terminal submetidos a transplante por meio de RMN de ^{1}H, integrando abordagem metabonômica, parâmetros clínicos e aprendizado de máquina. O objetivo é investigar os perfis metabólicos de pacientes que necessitam de diálise pós-transplante em comparação àqueles que não apresentam essa necessidade, utilizando os resultados e métodos desta tese como direcionamento.

O crescente desenvolvimento de estudos metabonômicos e a aplicação de aprendizado de máquina em dados clínicos podem contribuir para um futuro baseado em técnicas não invasivas como forma de diagnóstico, trazendo não apenas maior conforto aos pacientes, mas também a possibilidade de identificar doenças de forma precoce e em larga escala.

REFERÊNCIAS

ABBISS, H.; MAKER, G. L.; TRENGOVE, R. D. Metabolomics approaches for the diagnosis and understanding of kidney diseases. **Metabolites**, v. 9, n. 2, 34, 2019. Disponível em: https://doi.org/10.3390/metabo9020034

AGUIAR, L. K.; LADEIRA, R. M.; MACHADO, Í. E.; IVATA BERNAL, R. T.; MOURA, L.; MALTA, D. C. Fatores associados à doença renal crônica segundo critérios laboratoriais da Pesquisa Nacional de Saúde. **Revista Brasileira de Epidemiologia**, v. 23, 2020. Disponível em: https://doi.org/10.1590/1980-549720200101

AHMAD, G.N.; SHAFIULLAH; FATIMA, H.; ABBAS, M.; RAHMAN, O.; Imdadullah; Alqahtani, M.S. Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. **Applied Sciences**, v. 12, 7449, 2022. Disponível em: https://doi.org/10.3390/app12157449

ALAM, S.; AYUB, M. S.; ARORA, S.; KHAN, M. A. Uma investigação das técnicas de imputação de valores ausentes em dados ordinais, aprimorando a validade da análise de agrupamento e classificação. **Decision Analytics Journal**, v. 9, p. 100341, 2023. Disponível em: https://doi.org/10.1016/j.dajour.2023.100341

ALMAGHTHAWI, Y.; AHMAD, I.; ALSAADI, F.E. Performance Analysis of Feature Subset Selection Techniques for Intrusion Detection. **Mathematics**, v. *10*, 4745, 2022. https://doi.org/10.3390/math10244745

ANDERS, H. J.; SAXENA, R.; ZHAO, M. H.; PARODIS, I.; SALMON, J. E.; MOHAN, C. Nefrite lúpica. **Nature Reviews Disease Primers**, v. 6, p. 7, 2020. Disponível em: https://doi.org/10.1038/s41572-019-0141-9

ANEKTHANAKUL, K.; MANOCHEEWA, S.; CHIENWICHAI, K.; POUNGSOMBAT, P.; LIMJIASAHAPONG, S.; WANICHTHANARAK, K.; JARIYASOPIT, N.; MATHEMA, V. B.; KUHAKARN, C.; REUTRAKUL, V.; PHETCHARABURANIN, J.; PANYA, A.; PHONSATTA, N.; VISESSANGUAN, W.; POMYEN, Y.; SIRIVATANAUKSORN, Y.; WORAWICHAWONG, S.; SATHIRAPONGSASUTI, N.; KITIYAKARA, C.; KHOOMRUNG, S. Predicting lupus membranous nephritis using reduced picolinic acid to tryptophan ratio as a urinary biomarker. **iScience**, v. 24, 103355, 2021. Disponível em: 10.1016/j.isci.2021.

ANOWAR, F.; SADAOUI, S.; SELIM, B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). **Computer Science Review**, v. 40, 100378, 2021. Disponível em: https://doi.org/10.1016/j.cosrev.2021.100378

ARAÚJO JÚNIOR, A. S.; SATO, E. I.; SILVA DE SOUZA, A. W.; JENNINGS, F.; KIRSZTAJN, G. M.; SESSO, R.; DOS REIS-NETO, E. T. Development of an instrument to predict proliferative histological class in lupus nephritis based on clinical and laboratory data. **Lupus**, v. 32, n. 2, p. 216–224, 2023. Disponível em: https://doi.org/10.1177/09612033221143933

BAJEMA, I. M.; WILHELMUS, S.; ALPERS, C. E.; BRUIJN, J. A.; COLVIN, R. B.; COOK, H. T.; D'AGATI, V. D.; FERRARIO, F.; HAAS, M.; JENNETTE, J. C.; JOH, K.; NAST, C. C.; NOËL, L.; RIJNINK, E. C.; ROBERTS, I. S. D.; SESHAN, S. V.; SETHI, S.; FOGO, A. B. Revision of the International Society of Nephrology/Renal Pathology Society classification for lupus nephritis: clarification of definitions, and modified National Institutes of Health activity and chronicity indices. **Kidney International**, v. 93, p. 789-796, Issue 4, 2018. Disponível em: https://doi.org/10.1016/j.kint.2017.11.023.

BANSAL, M.; GOYAL, A.; CHOUDHARY, A. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. **Decision Analytics Journa**l, v. 3, 100071, 2022. Disponível em: https://doi.org/10.1016/j.dajour.2022.100071.

BARRETO, A. V. M. S.; ALECRIM, V. M.; MEDEIROS, T. B.; DOMINGUES, A. L. C.; LOPES, E. P.; MARTINS, J. R. M.; NADER, H. B.; DINIZ, G. T. N.; MONTENEGRO, S. M. L.; MORAIS, C. N. L. New index for the diagnosis of liver fibrosis in Schistosomiasis mansoni. **Arquivos de Gastroenterologia**, v. 54, n. 1, p. 51–56, 2017. Disponível em: https://doi.org/10.1590/S0004-2803.2017v54n1-10

BARRETO, A. V. M. S.; DOMINGUES, A. L. C.; DINIZ, G. T. N.; CAVALCANTI, A. M. S.; LOPES, E. P.; MONTENEGRO, S. M. L.; MORAIS, C. N. L. The Coutinho index as a simple tool for screening patients with advanced forms of Schistosomiasis mansoni: a validation study. **Transactions of The Royal Society of Tropical Medicine and Hygiene**, v. 116, n. 1, p. 19–25, 2022. Disponível em: https://doi.org/10.1093/trstmh/trab040 [acesso em: 12 ago. 2025].

BEZERRA, D. V. F.; QUEIROZ, J. W.; CÂMARA, V. A. V.; MACIEL, B. L. L.; NASCIMENTO, E. L. T.; JERÔNIMO, S. M. B. Factors Associated with Schistosoma mansoni Infestation in Northeast Brazil: A Need to Revisit Individual and Community Risk Factors. **Am J Trop Med Hyg**, v. 104, p. 1404-1411, 2021. Disponível em: 10.4269/ajtmh.19-0513.

BJERRUM, J. T.; WANG, Y. L.; SEIDELIN, J. B.; NIELSEN, O. H. IBD metabonomics predicts phenotype, disease course, and treatment response. **eBioMedicine**, v. 71, 103551, 2021. Disponível em: https://doi.org/10.1016/j.ebiom.2021.103551

BOATENG, E. Y.; ABAYE, D. A. A Review of the Logistic Regression Model with Emphasis on Medical Research. **Journal of Data Analysis and Information Processing**, v. 07, n. 04, p. 190–207, 2019. Disponível em: https://doi.org/10.4236/jdaip.2019.74012

BOEDEKER, P.; KEARNS, N. T. Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer. **Advances in Methods and Practices in Psychological Science**, v. 2, p. 250-263, 2019. Disponível em:10.1177/2515245919849378

BRAGG, F.; TRICHIA, E.; AGUILAR-RAMIREZ, D.; BEŠEVIĆ, J.; LEWINGTON, S.; EMBERSON, J. Predictive value of circulating NMR metabolic biomarkers for type 2 diabetes risk in the UK Biobank study. **BMC Medicine**, v. 20, n. 1, 2022. Disponível em: https://doi.org/10.1186/s12916-022-02354-9

BRASIL. Ministério da Saúde. Brasil é o terceiro maior transplantador de rim do mundo. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/2022/marco/brasil-e-o-terceiro-maior-transplantador-de-rim-do-mundo>. Acesso em: 21 abr. 2024.

BRASIL. Ministério da Saúde. Vigilância da esquistossomose mansoni. 1. ed. eletrônica. Brasília: Ministério da Saúde, 2024. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/vigilancia_esquistossomose_mansoni_1ed.pdf. Acesso em: 16 jul. 2025.

CAMPOS, M. P.; REIS, M. S. Data preprocessing for multiblock modelling – A systematization with new methods. **Chemometrics and Intelligent Laboratory Systems**, v. 199, 2020. Disponível em: https://doi.org/10.1016/j.chemolab.2020.103959

CANUTO, G. A. B.; DA COSTA, J. L.; DA CRUZ, P. L. R.; DE SOUZA, A. R. L.; FACCIO, A. T.; KLASSEN, A.; RODRIGUES, K. T.; TAVARES, M. F. M. Metabolômica: definições, estado-da-arte e aplicações representativas. **Química Nova**, v. 41, n. 1, p. 75-91, 2018. Disponível em: https://doi.org/10.21577/0100-4042.20170134

CARDOSO, M. R.; SILVA, A. A. R.; TALARICO, M. C. R.; SANCHES, P. H. G.; SFORÇA, M. L.; ROCCO, S. A.; REZENDE, L. M.; QUINTERO, M.; COSTA, T. B. B. C.; VIANA, L. R.; CANEVAROLO, R. R.; FERRACINI, A. C.; RAMALHO, S.; GUTIERREZ, J. M.; GUIMARÃES, F.; TASIC, L.; TATA, A.; SARIAN, L. O.; CHENG, L. L.; PORCARI, A. M.; DERCHAIN, S. F. M. Metabolomics by NMR Combined with Machine Learning to Predict Neoadjuvant Chemotherapy Response for Breast Cancer. **Cancers**, v. 14, n. 20, 2022. Disponível em: https://doi.org/10.3390/cancers14205055

CARR, H. Y.; PURCELL, E. M. Effects of diffusion on free precession in nuclear magnetic resonance experiments. **Physical Review**, New York, v. 94, n. 3, p. 630-638, 1954.

CHAI, X.; LIU, C.; FAN, X.; HUANG, T.; ZHANG, X.; JIANG, B.; LIU, M. Combination of peak-picking and binning for NMR-based untargeted metabonomics study. **Journal of Magnetic Resonance**, v. 351, 2023. Disponível em: https://doi.org/10.1016/j.jmr.2023.107429

CHASAPI, S. A.; KARAGKOUNI, E.; KALAVRIZIOTI, D.; VAMVAKAS, S.; ZOMPRA, A.; TAKIS, P. G.; GOUMENOS, D. S.; SPYROULIAS, G. A. NMR-based metabolomics in differential diagnosis of chronic kidney disease (CKD) subtypes. **Metabolites**, v. 12, n. 6, p. 490, 2022. Disponível em: https://doi.org/10.3390/metabo12060490

CHEN, F.; DAI, X.; ZHOU, C. C.; LI, K.-X.; ZHANG, Y.-J.; LOU, X. Y.; ZHU, Y. M.; SUN, Y.-L.; PENG, B. X.; CUI, W. Integrated analysis of the faecal metagenome and serum metabolome reveals the role of gut microbiome-associated metabolites in the detection of colorectal cancer and adenoma. **Gut**, v. 71, n. 7, p. 1315–1325, 2022. https://doi.org/10.1136/gutjnl-2020-323476

CHEN, R.; WANG, H.; SONG, L.; HOU, J.; PENG, J.; DAI, H.; PENG, L. Predictors and one-year outcomes of patients with delayed graft function after deceased donor kidney transplantation. **BMC Nephrology**, v. 21, n. 1, 2020. Disponível em: https://doi.org/10.1186/s12882-020-02181-1

CHEN, T. K.; KNICELY, D. H.; GRAMS, M. E. Chronic Kidney Disease Diagnosis and Management: A Review. **American Medical Association**, v. 322, p. 1294-1304, 2019. Disponível em: https://doi.org/10.1001/jama.2019.14745

CHI, J.; SHU, J.; LI, M.; MUDAPPATHI, R.; JIN, Y.; LEWIS, F.; BOON, A.; QIN, X.; LIU, L.; GU, H. Artificial intelligence in metabolomics: a current review. **Trends in analytical chemistry: TRAC**, v. 178, 117852, 2024. Disponível em: https://doi.org/10.1016/j.trac.2024.117852

CHI, N-F.; CHANG, T-H.; LEE, C-Y.; WU, Y-W.; SHEN, T-A.; CHAN, L.; CHEN, Y-R.; CHIOU, H-Y.; HSU, C. Y.; HU, C-J. Untargeted metabolomics predicts the functional outcome of ischemic stroke. **Journal of the Formosan Medical Association**, v. 120, p. 234-241, 2021. Disponível em: 10.1016/j.jfma.2020.04.026

CHIENWICHAI, P.; NOGRADO, K.; TIPTHARA, P.; TARNING, J.; LIMPANONT, Y.; CHUSONGSANG, P. et al. Untargeted serum metabolomic profiling for early detection of Schistosoma mekongi infection in mouse model. **Frontiers in Cellular and Infection Microbiology**, v. 12, 910177, 2022. Disponível em: https://doi.org/10.3389/fcimb.2022.910177

CHIENWICHAI, P.; TIPTHARA, P.; TARNING, J.; LIMPANONT, Y.; CHUSONGSANG, P.; CHUSONGSANG, Y. et al. Identification of trans-genus biomarkers for early diagnosis of intestinal schistosomiasis and progression of gut pathology in a mouse model using metabolomics. **PLoS Neglected Tropical Diseases**, v. 18, n. 2, e0011966, 2024. Disponível em: https://doi.org/10.1371/journal.pntd.0011966

CHIESA, M.; MAIOLI, G.; COLOMBO, G.I.; PIACENTINI, L. GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. **BMC Bioinformatics,** v. 21, 54, 2020. Disponível em: https://doi.org/10.1186/s12859-020-3400-6

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1960. Disponivel em: https://doi.org/10.1177/001316446002000104

COLNAGO, L. A.; ANDRADE, F. D. RMN no domínio do tempo: fundamentos e aplicações offline e inline", p. 439 -470. In: **Biotecnologia Aplicada à Agro&Indústria**

- Vol. 4. São Paulo: Blucher, 2017.
ISBN: 9788521211150. Disponível em: 10.5151/9788521211150-12

CORTES-SELVA, D.; GIBBS, L.; MASCHEK, J. A.; NASCIMENTO, M.; VAN RY, T.; COX, J. E. et al. Metabolic reprogramming of the myeloid lineage by Schistosoma mansoni infection persists independently of antigen exposure. **PLoS Pathogens**, v. 17, n. 1, e1009198, 2021. Disponível em: https://doi.org/10.1371/journal.ppat.1009198

DAI, L. shang; TIAN, H. fei; HANG, Y.; WEN, C. wei; HUANG, Y. hao; WANG, B. feng; HU, J. wei; XU, J. ping; DENG, M. jie. 1H NMR-based metabonomic evaluation of the pesticides camptothecin and matrine against larvae of Spodoptera litura. **Pest Management Science**, v. 77, n. 1, p. 208–216, 2021. Disponível em: https://doi.org/10.1002/ps.6009

DE ANDRADE, B. M.; DE GOIS, J. S.; XAVIER, V. L.; LUNA, A. S. Comparison of the performance of multiclass classifiers in chemical data: Addressing the problem of overfitting with the permutation test. **Chemometrics and Intelligent Laboratory Systems**, v. 201, 2020. Disponível em: https://doi.org/10.1016/j.chemolab.2020.104013

DE SAN-MARTIN, B. S.; FERREIRA, V. G.; BITENCOURT, M. R.; PEREIRA, P. C. G.; CARRILHO, E.; DE ASSUNÇÃO, N. A.; DE CARVALHO, L. R. S. Metabolomics as a potential tool for the diagnosis of growth hormone deficiency (Ghd): A review. **Archives of Endocrinology and Metabolism**, v. 64, n. 6, p. 654–663, 2020. Disponível em: https://doi.org/10.20945/2359-3997000000300

DEBIK, J.; SANGERMANI, M.; WANG, F.; MADSSEN, T. S.; GISKEØDEGÅRD, G. F. Multivariate analysis of NMR-based metabolomic data. **NMR IN BIOMEDICINE**, v. 35, 2022. Disponível em: https://doi.org/10.1002/nbm.4638

DI GIOVANNI, N.; MEUWIS, M. A.; LOUIS, E.; FOCANT, J. F. Correlations for untargeted GC × GC-HRTOF-MS metabolomics of colorectal cancer. **Metabolomics**, v. 19, n. 10, 2023. Disponível em: https://doi.org/10.1007/s11306-023-02047-1

DIEHL, B. Chapter 1 - Principles in NMR Spectroscopy. In: HOLZGRABE, Ulrike; WAWER, Iwona; DIEHL, Bernd (Ed.). NMR **Spectroscopy in Pharmaceutical Analysis.** Elsevier, 2008. p. 1-41. ISBN 9780444531735. Disponível em: https://doi.org/10.1016/B978-0-444-53173-5.00001-9. Acesso em: [out. 2023].

DONG, C. *et al.* Gut microbiota combined with metabolites reveals unique features of acute myocardial infarction patients different from stable coronary artery disease. **Journal of Advanced Research**, v. 46, p. 101–112, 2023. Disponível em: https://doi.org/10.1016/j.jare.2022.06.008

DUMANCAS, G. G.; RAMASAHAYAM, S.; BELLO, G.; HUGHES, J.; KRAMER, R. Chemometric regression techniques as emerging, powerful tools in genetic association studies. **TrAC Trends in Analytical Chemistry**, v. 74, p. 79-88, 2015. Disponível em: https://doi.org/10.1016/j.trac.2015.05.007

DUNN, W. B.; ELLIS, D. I. Metabolomics: Current analytical platforms and methodologies. **TrAC - Trends in Analytical Chemistry**, v. 24, n. 4, p. 285–294, 2005. Disponível em: https://doi.org/10.1016/j.trac.2004.11.021

EHRICH, J.; DUBOURG, L.; HANSSON, S.; PAPE, L.; STEINLE, T.; FRUTH, J.; HÖCKNER, S.; SCHIFFER, E. Mio-inositol sérico, dimetilsulfona e valina em combinação com creatinina permitem avaliação precisa da insuficiência renal — uma prova de conceito. **Diagnostics**, v. 11, 2021. Disponível em: https://doi.org/10.3390/diagnostics11020234

ELDAHSHAN, K. A.; ALHABSHY, A. A.; MOHAMMED, L. T. Filter and Embedded Feature Selection Methods to Meet Big Data Visualization Challenges. **Computers, Materials & Continua**, v. 74, p. 817-839, 2023. Disponível em: https://doi.org/10.32604/cmc.2023.032287

ELREEDY, D.; ATIYA, A. F. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. **Information Sciences**, v. 505, p. 32-64, 2019. doi.org/10.1016/j.ins.2019.07.070.

ELREEDY, D.; ATIYA, A. F.; KAMALOV, F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. **Machine Learning**, v. 113, p. 4903–4923, 2024. Disponível em: https://doi.org/10.1007/s10994-022-06296-4

EMWAS, A. H.; ROY, R.; MCKAY, R. T.; TENORI, L.; SACCENTI, E.; NAGANA GOWDA, G. A.; RAFTERY, D.; ALAHMARI, F.; JAREMKO, L.; JAREMKO, M.; WISHART, D. S. Nmr spectroscopy for metabolomics research. **Metabolites**, v. 9, 2019. Disponível em: https://doi.org/10.3390/metabo9070123

ERNSTMEYER, K.; CHRISTMAN, E., eds. Chapter 8: Renal and Urinary System Alterations. In: Open Resources for Nursing (Open RN). **Health Alterations [internet]**. Eau Claire (WI): Chippewa Valley Technical College, 2024. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK613065/

EROGLU, E. C.; KUCUKGOZ GULEC, U.; VARDAR, M. A.; PAYDAS, S. GC-MS based metabolite fingerprinting of serous ovarian carcinoma and benign ovarian tumor. **European Journal of Mass Spectrometry**, v. 28, n. 1–2, p. 12–24, 2022. Disponível em: https://doi.org/10.1177/14690667221098520

EWUZIE, A.; WILBURN, L.; THAKRAR, D. B.; CHENG, H.; REITZUG, F.; ROBERTS, N.; MALOUF, R.; CHAMI, G. F. Association of current *Schistosoma mansoni*, *Schistosoma japonicum*, and *Schistosoma mekongi* infection status and intensity with periportal fibrosis: a systematic review and meta-analysis. **The Lancet Global Health**, v. 13, p. e69–e80, 2025. Disponível em: 10.1016/S2214-109X(24)00425-X.

FERREIRA, M. M. C. **Quimiometria: conceitos, métodos e aplicações**. Editora da Unicamp, 2015. Disponível em: https://doi.org/10.7476/9788526814714

FONSECA, R. I. D.; MENEZES, L. R. A.; SANTANA-FILHO, A. P.; SCHIEFER, E. M.; PECOITS-FILHO, R.; STINGHEN, A. E. M.; SASSAKI, G. L. Untargeted plasma 1H NMR-based metabolomic profiling in different stages of chronic kidney disease. **Journal of Pharmaceutical and Biomedical Analysis**, v. 229, 2023. Disponível em: https://doi.org/10.1016/j.jpba.2023.115339

FORESTO, R. D.; PESTANA, J. O. M.; SILVA, H. T. Brasil: The leading public kidney transplant program worldwide. *[S. l.]*: **Associacao Medica Brasileira**, 2020. Disponível em: https://doi.org/10.1590/1806-9282.66.6.708

FRANIEK, A.; SHARMA, A.; COCKOVSKI, V.; WISHART, D. S.; ZAPPITELLI, M.; BLYDT-HANSEN, T. D. Urinary metabolomics to develop predictors for pediatric acute kidney injury. **Pediatric Nephrology**, v. 37, n. 9, p. 2079–2090, 2022. Disponível em: https://doi.org/10.1007/s00467-021-05380-6

GALVAN, D.; DE AGUIAR, L. M.; BONA, E.; MARINI, F.; KILLNER, M. H. M. **Successful combination of benchtop nuclear magnetic resonance spectroscopy and chemometric tools: A review**. *[S. l.]*: Elsevier B.V., 2023. Disponível em: https://doi.org/10.1016/j.aca.2023.341495

GARCIA-ALVAREZ, D.; BREGON, A.; PULIDO, B.; ALONSO-GONZALEZ, C. J. Integrating PCA and structural model decomposition to improve fault monitoring and diagnosis with varying operation points. **Engineering Applications of Artificial Intelligence**, v. 122, 2023. Disponível em: https://doi.org/10.1016/j.engappai.2023.106145

GARZARELLI, V.; FERRARA, F.; PRIMICERI, E.; CHIRIACÒ, M. S. Biofluids manipulation methods for liquid biopsy in minimally-invasive assays. **MethodsX**, v. 9, 101759, 2022. Disponível em: https://doi.org/10.1016/j.mex.2022.101759.

GHINI, V.; MEONI, G.; VIGNOLI, A.; DI CESARE, F.; TENORI, L.; TURANO, P.; LUCHINAT, C. **Fingerprinting and profiling in metabolomics of biosamples**. *[S. l.]*: Elsevier B.V., 2023. Disponível em: https://doi.org/10.1016/j.pnmrs.2023.10.002

GOUVEIA, L. R.; SANTOS, J. C.; SILVA, R. D.; BATISTA, A. D.; DOMINGUES, A. L. C.; LOPES, E. P. D. A. et al. Diagnosis of coinfection by schistosomiasis and viral hepatitis B or C using 1H NMR-based metabonomics. **PLoS ONE**, v. 12, n. 8, e0182196, 2017. Disponível em: https://doi.org/10.1371/journal.pone.0182196

GRAF, R.; ZELDOVICH, M.; FRIEDRICH, S. Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study. **Biometrical Journal**, v. 66, n. 1, 2024. Disponível em: https://doi.org/10.1002/bimj.202200098.

GRASSO, D.; PILLOZZI, S.; TAZZA, I.; BERTELLI, M.; CAMPANACCI, D. A.; PALCHETTI, I.; BERNINI, A. An improved NMR approach for metabolomics of intact serum samples. **Analytical Biochemistry**, v. 654, 2022. Disponível em: https://doi.org/10.1016/j.ab.2022.114826

GRONWALD, W.; KLEIN, M. S.; ZELTNER, R.; SCHULZE, B.-D.; REINHOLD, S. W.; DEUTSCHMANN, M.; IMMERVOLL, A.-K.; BÖGER, C. A.; BANAS, B.; ECKARDT, K.-U.; OEFNER, P. J. Detection of autosomal dominant polycystic kidney disease by NMR spectroscopic fingerprinting of urine. **Kidney International**, v. 79, n. 11, p. 1244-1253, 2011. ISSN 0085-2538. DOI: 10.1038/ki.2011.30.

GUNDA, D. W.; KILONZO, S. B.; MANYIRI, P. M.; PECK, R. N.; MAZIGO, H. D. Morbidity and Mortality Due to Schistosoma mansoni Related Periportal Fibrosis: Could Early Diagnosis of Varices Improve the Outcome Following Available Treatment Modalities in Sub Saharan Africa? A Scoping Review. **Trop Med Infect Dis,** v. 5, 2020. Disponível em: 10.3390/tropicalmed5010020.

GUO, Z. S.; LU, M. M.; LIU, D. W.; ZHOU, C. Y.; LIU, Z. S.; ZHANG, Q. Identification of amino acids metabolomic profiling in human plasma distinguishes lupus nephritis from systemic lupus erythematosus. **Amino Acids**, v. 56, 2024. Disponível em: 10.1007/s00726-024-03418-1.

HABIB, M.; VICENTE-PALACIOS, V.; GARCÍA-SÁNCHEZ, P. Bio-inspired optimization of feature selection and SVM tuning for voice disorders detection. **Knowledge-Based Systems**, v. 310, 112950, 2025. doi.org/10.1016/j.knosys.2024.112950.

HAHN, E. L. Spin Echoes. **Physical Review**, [s.l.], v. 80, p. 580, 1950.

HALL, J. E, GUYTON, A. C. Guyton and Hall Textbook of Medical Physiology. 12. ed. [s.l.: s.n.], 2011. (**Rev. ed. of: Textbook of medical physiology**. 11. ed. c2006). ISBN 978-1-4160-4574-8.

HANG, J.; CHEN, Y.; LIU, L.; CHEN, L.; FANG, J.; WANG, F.; WANG, M. Antitumor effect and metabonomics of niclosamide micelles. **Journal of Cellular and Molecular Medicine**, v. 26, n. 18, p. 4814–4824, 2022. Disponível em: https://doi.org/10.1111/jcmm.17509

HAO, Z.; YAO, J.; ZHAO, X.; LIU, R.; CHANG, B.; SHAO, H. Preliminary observational study of metabonomics in patients with early and late-onset type 2 diabetes mellitus based on UPLC-Q-TOF/MS. **Scientific Reports**, v. 13, n. 1, 2023. Disponível em: https://doi.org/10.1038/s41598-023-41883-y

HARRIS, C. R.; MILLMAN, K. J.; VAN DER WALT, S. J.; GOMMERS, R.; PÉREZ, F.; NIMROD, N.; WALT, S. J.; WATKINS, A.; BÉRAUD, J. L.; RØNNE, J.; WEYLAND, R. L.; PLIS, S.; PÉREZ, F.; KAPPA, M. Array programming with NumPy. **Nature**, [s.l.], v. 585, p. 357-362, 2020. DOI: 10.1038/s41586-020-2649-2.

HASSELBALCH, R. B.; KRISTENSEN, J. H.; STRANDKJÆR, N.; JØRGENSEN, N.; BUNDGAARD, H.; MALMENDAL, A.; IVERSEN, K. K. Metabolomics of early myocardial ischemia. **Metabolomics**, v. 19, n. 4, 2023. Disponível em: https://doi.org/10.1007/s11306-023-01999-8

HASUBEK, A. L. *et al.* Differentiation of patients with and without prostate cancer using urine 1H NMR metabolomics. **Magnetic Resonance in Chemistry**, v. 61, n. 12, p. 740–747, 2023. Disponível em: https://doi.org/10.1002/mrc.5391

HAYATI, R.; MUNAWAR, A. A.; LUKITANINGSIH, E.; EARLIA, N.; KARMA, T.; IDROES, R. Combination of PCA with LDA and SVM classifiers: A model for determining the geographical origin of coconut in the coastal plantation, Aceh Province, Indonesia. **Case Studies in Chemical and Environmental Engineering**, v. 9, 2024. Disponível em: https://doi.org/10.1016/j.cscee.2023.100552

HU, J.; SHEN, Y.; CHEN, Y. Metabonomics Application on Screening Serum Biomarkers of Golden Hamsters with Nonalcoholic Steatohepatitis Induced by High-Fat Diet. **Combinatorial chemistry & high throughput screening,** v. 26, p. 2280-2292, 2023. Disponível em: 10.2174/1386207326666230223095745

HU, J.; ZHANG, X.; GU, J.; YANG, M.; ZHANG, X.; ZHAO, H.; LI, L. Serum alkaline phosphatase levels as a simple and useful test in screening for significant fibrosis in treatment-naive patients with hepatitis B e-antigen negative chronic hepatitis B. **European Journal of Gastroenterology & Hepatology**, v. 31, n. 7, p. 817–823, 2019. Disponível em: https://doi.org/10.1097/MEG.0000000000001336

HU, Y.; CHEN, J.; XU, Y.; ZHOU, H.; HUANG, P.; MA, Y. et al. Alterations of gut microbiome and metabolite profiling in mice infected by Schistosoma japonicum. **Frontiers in Immunology**, v. 11, 569727, 2020. Disponível em: https://doi.org/10.3389/fimmu.2020.569727

HU, Y.; SUN, L.; YUAN, Z.; XU, Y.; CAO, J. High throughput data analyses of the immune characteristics of Microtus fortis infected with Schistosoma japonicum. **Scientific Reports**, v. 7, n. 1, 11311, 2017. Disponível em: https://doi.org/10.1038/s41598-017-11644-3

HUANG, P.; LIU, Y.; LI, Y.; XIN, Y.; NAN, C.; LUO, Y.; FENG, Y.; JIN, N.; PENG, Y.; WANG, D.; ZHOU, Y.; LUAN, F.; WANG, X.; WANG, X.; LI, H.; ZHOU, Y.; ZHANG, W.; LIU, Y.; YUAN, M.; ZHANG, Y.; SONG, Y.; XIAO, Y.; SHEN, L.; YU, K.; ZHAO, M.; CHENG, L.; WANG, C. Metabolomics- and proteomics-based multi-omics integration reveals early metabolite alterations in sepsis-associated acute kidney injury. **BMC Med**, v. 23, 143, 2025. Disponível em: https://doi.org/10.1186/s12916-025-03980-9

HUANG, S.; HU, D.; YUAN, S.; HE, Y.; LI, C.; ZHU, Y.; WU, X. The serum metabolomics study of liver failure and artificial liver therapy intervention. **Medical Science Monitor**, v. 27, 2021. Disponível em: https://doi.org/10.12659/MSM.930638

HUANG, Y.; WU, Q.; ZHAO, L.; XIONG, C.; XU, Y.; DONG, X. et al. UHPLC-MS-based metabolomics analysis reveals the process of schistosomiasis in mice. **Frontiers in Microbiology**, v. 11, 1517, 2020. Disponível em: https://doi.org/10.3389/fmicb.2020.01517

HULJANAH, M.; RUSTAM, Z.; UTAMA, S.; SISWANTINING, T. Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. **IOP Conf. Ser.: Mater. Sci. Eng.** v. 546, 052031, 2019. Disponível em: 10.1088/1757-899X/546/5/052031

HUNG, C. I.; LIN, G.; CHIANG, M. H.; CHIU, C. Y. Metabolomics-based discrimination of patients with remitted depression from healthy controls using 1H-NMR spectroscopy. **Scientific Reports**, v. 11, n. 1, 2021. Disponível em: https://doi.org/10.1038/s41598-021-95221-1

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science \& Engineering**, v. 9, n. 3, p. 90–95, 2007. Disponivel em: https://doi.org/10.1109/MCSE.2007.55

HUSSAIN, A.; PAUKOVICH, N.; HENEN, M. A.; VÖGELI, B. Advances in the exact nuclear Overhauser effect 2018–2022. **Methods**, v. 206, p. 87–98, 2022. Disponível em: https://doi.org/10.1016/j.ymeth.2022.08.006

IKEUCHI, H.; IMAI, Y.; MARUYAMA, S.; SUGIYAMA, H.; SATO, H.; YOKOYAMA, H.; HIROMURA, K. Long-term prognosis of pure membranous lupus nephritis: a comparison with proliferative lupus nephritis in Japan. **Clinical and Experimental Nephrology**, 2025. Disponível em: https://doi.org/10.1007/s10157-025-02709-5

IMAM, F.; MUSILEK, P.; REFORMAT, M. Z. Parametric and nonparametric machine learning techniques for increasing power system reliability: a review. **Information**, v. 15, 2024. Disponível em: https://doi.org/10.3390/info15010037

INTERNATIONAL SOCIETY OF NEPHROLOGY. ISN–Global Kidney Health Atlas: A report by the International Society of Nephrology: An Assessment of Global Kidney Health Care Status focussing on Capacity, Availability, Accessibility, Affordability and Outcomes of Kidney Disease. 2023.

IUPAC. Gold book: the IUPAC compendium of chemical terminology. 2. ed. [s.l.]: IUPAC, 2019.

JACOB, M.; NIMER, R. M.; ALABDALJABAR, M. S.; SABI, E. M.; AL-ANSARI, M. M.; HOUSIEN, M.; SUMAILY, K. M.; DAHABIYEH, L. A.; ABDEL RAHMAN, A. M. Metabolomics Profiling of Nephrotic Syndrome towards Biomarker Discovery. **International Journal of Molecular Sciences**, v. 23, n. 20, 2022. Disponível em: https://doi.org/10.3390/ijms232012614

JIA, W.; SUN, M.; LIAN, J.; HOU, S. Feature dimensionality reduction: a review. **Complex and Intelligent Systems**, v. 8, 3, p. 2663–2693, 2022. Disponivel em: https://doi.org/10.1007/s40747-021-00637-x

JIN, W.; BI, J.; XU, S.; RAO, M.; WANG, Q.; YUAN, Y.; FAN, B. Metabolic regulation mechanism of Aconiti Radix Cocta extract in rats based on 1H-NMR metabonomics. **Chinese Herbal Medicines**, v. 14, n. 4, p. 602–611, 2022. Disponível em: https://doi.org/10.1016/j.chmed.2022.07.002

JUNG, B. A.; WEIGEL, M. Spin echo magnetic resonance imaging. **Journal of magnetic resonance imaging**, v. 37, 2013. Disponível em: https://doi.org/10.1002/jmri.24068

KALANTAR-ZADEH, K.; JAFAR, T. H.; NITSCH, D.; NEUEN, B. L.; PERKOVIC, V. **Chronic kidney disease**. Elsevier B.V., 2021. Disponível em: https://doi.org/10.1016/S0140-6736(21)00519-5

KALAS, M. A.; CHAVEZ, L.; LEON, M.; TAEWESEDT, P. T.; SURANI, S. Abnormal liver enzymes: a review for clinicians. **World Journal of Hepatology**, v. 13, n. 11, p. 1688–1698, 2021. Disponível em: https://doi.org/10.4254/wjh.v13.i11.1688

KAWAMURA, T. Interpretação de um teste sob a visão epidemiológica: eficiência de um teste. **Arq Bras Cardiol**, v. 79, p. 437–441, 2002. Disponível em: https://doi.org/10.1590/S0066-782X2002001300015

KDIGO. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. **Official Journal of the International Society of Nephrology**, [s.l.], 2012. Disponível em: https://www.publicationethics.org. Acesso em: [mar. 2024]

KDIGO. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease, 2024. Disponível em: www.kidney-international.org. Acesso em: [abr. 2024].

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics,** v. 11, n. 1, p. 137–148, 1969. Disponível em: https://doi.org/10.2307/1266770

KHALID, U.; NEWBURY, L. J.; SIMPSON, K.; JENKINS, R. H.; BOWEN, T.; BATES, L.; SHEERIN, N. S.; CHAVEZ, R.; FRASER, D. J. A urinary microRNA panel that is an early predictive biomarker of delayed graft function following kidney transplantation. **Scientific Reports**, v. 9, n. 1, 2019. Disponível em: https://doi.org/10.1038/s41598-019-38642-3

KHAROUF, F.; MEHTA, P.; LI, Q.; GLADMAN, D. D.; TOUMA, Z.; GARCIA, L. P. W. Does the time to the onset of lupus nephritis impact renal disease presentation and outcomes?. **Seminars in Arthritis and Rheumatism**, v. 73, 152724, 2025. Disponível em: https://doi.org/10.1016/j.semarthrit.2025.152724.

KHW AJA, A. KDIGO clinical practice guidelines for acute kidney injury. **Nephron Clinical Practice**, v. 120, n. 4, p. c179–c184, 2012. Disponível em: https://doi.org/10.1159/000339789

KIM, D. W.; KIM, H. J.; SEONG, E. Y.; KIM, S. S.; LEE, S.; KIM, S.; KWON, C. H.; SONG, S. H. Virtual diagnosis of diabetic nephropathy using metabolomics in place of kidney biopsy: The DIAMOND study. **Diabetes Research and Clinical Practice**, v. 205, 2023. Disponível em: https://doi.org/10.1016/j.diabres.2023.110986

KIM, J. A.; CHOI, H. J.; KWON, Y. K.; RYU, D. H.; KWON, T. H.; HWANG, G. S. 1H NMR-based metabolite profiling of plasma in a rat model of chronic kidney disease. **PLoS ONE**, v. 9, n. 1, 2014. Disponível em: https://doi.org/10.1371/journal.pone.0085445

KISELEVA, O.; KURBATOV, I.; ILGISONIS, E.; POVERENNAYA, E. Defining blood plasma and serum metabolome by GC-MS. **Metabolites**, v. 12, n. 1, 15, 2021. Disponível em: https://doi.org/10.3390/metabo12010015

KOLMOGOROV, A. Sulla determinazione empirica di una legge di distribuzione. **Giornale dell'Istituto Italiano degli Attuari,** v. 4, p. 83–91, 1933.

KOU, J.; HE, C.; CUI, L.; ZHANG, Z.; WANG, W.; TAN, L.; LIU, D.; ZHENG, W.; GU, W.; XIA, N. Discovery of Potential Biomarkers for Postmenopausal Osteoporosis Based on Untargeted GC/LC-MS. **Frontiers in Endocrinology**, v. 13, 2022. Disponível em: https://doi.org/10.3389/fendo.2022.849076

KUMAR, A.; RANI GRACE, R. C. Nuclear overhauser effect. *In*: **Encyclopedia of Spectroscopy and Spectrometry**. *[S. l.]*: Elsevier, 2016. p. 423–431. Disponível em: https://doi.org/10.1016/B978-0-12-803224-4.00236-3

LABORY, J.; NJOMGUE-FOTSO, E.; BOTTINI, S. Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. **Computational and Structural Biotechnology Journal**, v. 23, p. 1274–1287, 2024. Disponível em: https://doi.org/10.1016/j.csbj.2024.03.016

LAI, Q.; ZHOU, B.; CUI, Z.; AN, X.; ZHU, L.; CAO, Z.; LIU, S.; YU, B. Development of a metabolite-based deep learning algorithm for clinical precise diagnosis of the progression of diabetic kidney disease. **Biomedical Signal Processing and Control**, v. 83, 104625, 2023. Disponível em: https://doi.org/10.1016/j.bspc.2023.104625.

LEE, A. M.; HU, J.; XU, Y.; ABRAHAM, A. G.; XIAO, R.; CORESH, J.; REBHOLZ, C.; CHEN, J.; RHEE, E. P.; FELDMAN, H. I.; RAMACHANDRAN, V. S.; KIMMEL, P. L.; WARADY, B. A.; FURTH, S. L.; DENBURG, M. R. Using Machine Learning to Identify Metabolomic Signatures of Pediatric Chronic Kidney Disease Etiology. **Journal of the American Society of Nephrology**, v. 33, n. 2, p. 375–386, 2022. Disponível em: https://doi.org/10.1681/ASN.2021040538

LEE, A.M.; HU, J.; XU, Y.; ABRAHAM, A.G.; XIAO, R.; CORESH, J.; REBHOLZ, C.; CHEN, J.; RHEE, E.P.; FELDMAN, H.I.; RAMACHANDRAN, V. S.; KIMMEL, P.L.; WARADY, B.A.; FURTH, S.L.; DENBURG, M.R.; CKD Biomarkers Consortium. Using Machine Learning to Identify Metabolomic Signatures of Pediatric Chronic Kidney Disease Etiology. **J Am Soc Nephrol**, v. 33, p. 375-386, 2022. Disponível em: https://doi.org/10.1681/ASN.2021040538.

LEE, L. C.; JEMAIN, A. A. On overview of PCA application strategy in processing high dimensionality forensic data. **Microchemical Journal**, v. 169, 106608, 2021. Disponível em: https://doi.org/10.1016/j.microc.2021.106608

LEITE, L. A.; DOMINGUES, A. L.; LOPES, E. P.; FERREIRA, R. C.; PIMENTA, A. A. F.; DA FONSECA, C. S.; DOS SANTOS, B. S.; LIMA, V. L. Relationship between splenomegaly and hematologic findings in patients with hepatosplenic schistosomiasis. **Revista Brasileira de Hematologia e Hemoterapia**, v. 35, n. 5, p. 332–336, 2013. Disponível em: https://doi.org/10.5581/1516-8484.20130098

LEITE, L. A.; PIMENTA FILHO, A. A.; MARTINS DA FONSECA, C. S.; SANTANA DOS SANTOS, B.; FERREIRA, R. C.; MONTENEGRO, S. M.; LOPES, E. P.; DOMINGUES, A. L.; OWEN, J. S.; LIMA, V. L. Hemostatic dysfunction is increased in patients with hepatosplenic schistosomiasis mansoni and advanced periportal fibrosis. **PLoS Neglected Tropical Diseases**, v. 7, n. 7, e2314, 18 jul. 2013. Disponível em: https://doi.org/10.1371/journal.pntd.0002314

LEMAITRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, **Journal of Machine Learning Research**, v. 18, p. 1-5, 2017. Disponivel em: http://arxiv.org/abs/1609.06570

LEVITT, M. H. Spin dynamics: basics of nuclear magnetic resonance. **Concepts in Magnetic Resonance Part A**, v. 34, p. 60–61, 2008. Disponível em: 10.1002/cmr.a.20130

LI, J.; JIANG, J.; ZHU, Y.; ZHANG, Y.; ZHU, J.; MING, Y. Metabolomics analysis of patients with Schistosoma japonicum infection based on UPLC-MS method. **Parasites & Vectors**, v. 17, n. 1, 350, 2024. Disponível em: https://doi.org/10.1186/s13071-024-06106-9

LI, M.; LIU, M.; WANG, B.; SHI, L. Metabonomics Analysis of Stem Extracts from Dalbergia sissoo. **Molecules**, v. 27, n. 6, 2022. Disponível em: https://doi.org/10.3390/molecules27061982

LIN, Y., MA, C., LIU, C., WANG, Z., YANG, J., LIU, X., SHEN, Z., WU, R. NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal câncer. **Oncotarget**, v. 7, n. 20, p. 29454-29464, 2016.

LIU, A.; XUAN, W.; XIAO, Y. State-of-the-art review on applications of various machine learning models in biodiesel production. **Chemometrics and Intelligent Laboratory Systems**, v. 262, 105391, 2025. doi.org/10.1016/j.chemolab.2025.105391.

LIU, Y.; XIE, S.; ZHOU, J.; CAI, Y.; ZHANG, P.; LI, J.; MING, Y. Using blood routine indicators to establish a machine learning model for predicting liver fibrosis in patients with *Schistosoma japonicum*. **Scientific Reports**, v. 14, 11485, 2024. Disponível em: https://doi.org/10.1038/s41598-024-62521-1

LIU, Y.; XIE, S.; ZHOU, J.; CAI, Y.; ZHANG, P.; LI, J.; MING, Y. Using blood routine indicators to establish a machine learning model for predicting liver fibrosis in patients with Schistosoma japonicum**. Scientific Reports**, v. 14, n. 1, p. 11485, 2024. Disponível em: https://doi.org/10.1038/s41598-024-62521-1

LOYO, R. M.; ZARATE, E.; BARBOSA, C. S.; SIMOES-BARBOSA, A. Gas chromatography-mass spectrometry (GC/MS) reveals urine metabolites associated to light and heavy infections by Schistosoma mansoni in mice. **Parasitology International**, v. 80, 102239, 2021. Disponível em: https://doi.org/10.1016/j.parint.2020.102239

MA, J.; PATHIRANA, C.; LIU, D. Q.; MILLER, S. A. **NMR spectroscopy as a characterization tool enabling biologics formulation development**. *[S. l.]*: Elsevier B.V., 2023. Disponível em: https://doi.org/10.1016/j.jpba.2022.115110

MANNON, R. B. Delayed graft function: The AKI of kidney transplantation. **Kidney Diseases**, v. 4, n. 4, p. 211–220, 2018. Disponível em: https://doi.org/10.1159/000491558

MARINO, C.; GRIMALDI, M.; SABATINI, P.; AMATO, P.; PALLAVICINO, A.; RICCIARDELLI, C.; D'URSI, A. M. Fibromyalgia and depression in women: An 1h-nmr metabolomic study. **Metabolites**, v. 11, n. 7, 2021. Disponível em: https://doi.org/10.3390/metabo11070429

MATIYASEVICH, M. tqdm: A Fast, Extensible Progress Bar for Python and CLI, 2010. Disponível em: https://github.com/tqdm/tqdm.

MCKINNEY, W. Data Structures for Statistical Computing in Python. In: 2010, Anais, p. 56–61. Disponível em: https://doi.org/10.25080/Majora-92bf1922-00a

MEDEIROS, T. B.; DOMINGUES, A. L. C.; LUNA, C. F.; LOPES, E. P. Correlation between platelet count and both hepatic fibrosis and spleen diameter in patients with Schistosomiasis mansoni. **Arquivos de Gastroenterologia**, v. 51, n. 1, p. 34–38, 2014. Disponível em: https://doi.org/10.1590/S0004-28032014000100008

MEHTA, R. L.; KELLUM, J. A.; SHAH, S. V.; MOLITORIS, B. A.; RONCO, C.; WARNOCK, D. G.; LEVIN, A. Acute Kidney Injury Network: Report of an initiative to improve outcomes in acute kidney injury. **Critical Care**, v. 11, n. 2, 2007. Disponível em: https://doi.org/10.1186/cc5713

MEIBOOM, S.; GILL, D. Modified spin-echo method for measuring nuclear relaxation times. **Review of Scientific Instruments**, v. 29, n. 8, p. 688–691, 1958. Disponível em: https://doi.org/10.1063/1.1716296

MENDEZ, K. M.; REINKE, S. N.; BROADHURST, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. **Metabolomics**, v. 15, n. 12, 2019. Disponível em: https://doi.org/10.1007/s11306-019-1612-4

MERCIER, K.; MCRITCHIE, S.; PATHMASIRI, W.; NOVOKHATNY, A.; KORALKAR, R.; ASKENAZI, D.; BROPHY, P. D.; SUMNER, S. Preterm neonatal urinary renal developmental and acute kidney injury metabolomic profiling: an exploratory study. **Pediatric Nephrology**, v. 32, n. 1, p. 151–161, 2017. Disponível em: https://doi.org/10.1007/s00467-016-3439-9

MIELKO, K. A.; PUDEŁKO-MALIK, N.; TARCZEWSKA, A.; MŁYNARZ, P. NMR spectroscopy as a “green analytical method” in metabolomics and proteomics studies. **Sustainable Chemistry and Pharmacy**, v. 22, 2021. Disponível em: https://doi.org/10.1016/j.scp.2021.100474

MISHRA, P.; BIANCOLILLO, A.; ROGER, J. M.; MARINI, F.; RUTLEDGE, D. N. New data preprocessing trends based on ensemble of multiple preprocessing techniques. **TrAC - Trends in Analytical Chemistry**, v. 132, 116045, 2020. Disponível em: https://doi.org/10.1016/j.trac.2020.116045

MO, H.; RAFTERY, D. Improved residual water suppression: WET180. **Journal of Biomolecular NMR**, v. 41, n. 2, p. 105–111, 2008. Disponível em: https://doi.org/10.1007/s10858-008-9246-2

MOHAMED SAIED MOHAMED, A.; MARTIN, A. Acute kidney injury in critical care. **Anaesthesia & Intensive Care Medicine**, v. 25, n. 5, p. 308-315, 2024. ISSN 1472-0299. DOI: 10.1016/j.mpaic.2024.03.008.

MORATH, C.; DÖHLER, B.; KÄLBLE, F.; PEGO DA SILVA, L.; ECHTERDIEK, F.; SCHWENGER, V.; ŽIVČIĆ-ĆOSIĆ, S.; KATALINIĆ, N.; KUYPERS, D.; BENÖHR, P.; HAUBITZ, M.; ZIEMANN, M.; NITSCHKE, M.; EMMERICH, F.; PISARSKI, P.; KARAKIZLIS, H.; WEIMER, R.; RUHENSTROTH, A.; SCHERER, S.; TRAN, T. H.; MEHRABI, A.; ZEIER, M.; SÜSAL, C. Pre-transplant HLA antibodies and delayed graft function in the current era of kidney transplantation. **Frontiers in Immunology**, v. 11, p. 1886, 2020. DOI: 10.3389/fimmu.2020.01886

MORENO-BAREA, F. J.; FRANCO, L.; ELIZONDO, D.; GROOTVELD, M. Application of data augmentation techniques towards metabolomics. **Computers in Biology and Medicine**, v. 148, 2022, 105916, doi.org/10.1016/j.compbiomed.2022.105916.

MUSA, R.; ROUT, P.; QURIE, A. Lupus Nephritis. In: StatPearls [Internet]. Treasure Island (FL): **StatPearls Publishing**, 2025. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK499817/. Acesso em: [10 jul. 2025].

N. HAQUE, T. ISLAM, M. ERFAN. An exploration of machine learning approaches for early autism spectrum disorder detection, **Healthcare Analytics**, v. 7, 100379, 2025. Disponível em: https://doi.org/10.1016/j.health.2024.100379.

NAM, J. H.; KIM, D.; CHUNG, D. Sparse linear discriminant analysis using the prior-knowledge-guided block covariance matrix. **Chemometrics and Intelligent Laboratory Systems**, v. 206, 104142, 2020. Disponível em: https://doi.org/10.1016/j.chemolab.2020.104142.

NARDELLI, L.; SCALAMOGNA, A.; MESSA, P.; GALLIENI, M.; CACCIOLA, R.; TRIPODI, F.; CASTELLANO, G.; FAVI, E. Peritoneal dialysis for potential kidney transplant recipients: pride or prejudice? **Medicina** (Kaunas), v. 58, n. 2, p. 214, 2022. DOI: 10.3390/medicina58020214.

NETO, F. T. L.; MARQUES, R. A.; CAVALCANTI FILHO, A. de F.; FONTE, J. E. F. da; LIMA, S. V. C.; SILVA, R. O. Prediction of semen analysis parameter improvement after varicocoelectomy using 1 H NMR-based metabonomics assays. **Andrology**, v. 10, n. 8, p. 1581–1592, 2022. Disponível em: https://doi.org/10.1111/andr.13281

NEULING, N. R.; ALLERT, R. D.; BUCHER, D. B. Prospects of single-cell nuclear magnetic resonance spectroscopy with quantum sensors. **Current opinion in biotechnology**, v. 83, 102975, 2023. Disponível em: https://doi.org/10.1016/j.copbio.2023.102975

NICHOLSON, J. K., LINDON, J. C., Metabonomics. **Nature,** v. 455, p. 1054-1056, 2008. Disponível em: https://doi.org/10.1038/4551054a

NICHOLSON, J. K.; LINDON, J. C.; HOLMES, E. “Metabonomics”: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. **Xenobiotica**, v. 29, n. 11, p. 1181–1189, 1999. Disponível em: 10.1080/004982599238047.

OLIVEIRA, M. F.; DE ALBUQUERQUE NETO, M. C.; LEITE, T. S.; ALVES, P. A. A.; LIMA, S. V. C.; SILVA, R. O. Performance evaluate of different chemometrics formalisms used for prostate cancer diagnosis by NMR-based metabolomics. **Metabolomics**, v. 20, n. 1, 2024. Disponível em: https://doi.org/10.1007/s11306-023-02067-x

OZAKI, B. C.; THORP, R. D.; BATISTA, A. D.; MARIZ, C. D.; DINIZ, G. T.; BARRETO, A. V.; DE MORAIS, C. N.; DOMINGUES, A. C.; LOPES, E. P. A-250 Evaluation of Liver Fibrosis in Schistosomiasis Mansoni Using the ELF (Enhanced Liver Fibrosis) Score. **Clinical Chemistry**, v. 70, Oct. 2024. Disponível em: https://doi.org/10.1093/clinchem/hvae106.247

PAN, P. J.; LEE, C. H.; HSU, N. W.; SUN, T. L. Combining principal component analysis and logistic regression for multifactorial fall risk prediction among community-dwelling older adults. **Geriatric Nursing**, v. 57, p. 208–216, 2024. Disponível em: https://doi.org/10.1016/j.gerinurse.2024.04.021

PANG, Z.; LU, Y.; ZHOU, G.; HUI, F.; XU, L.; VIAU, C.; SPIGELMAN, A. F.; MACDONALD, P. E.; WISHART, D. S.; LI, S.; XIA, J. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. **Nucleic Acids Research**, v. 52, p. W398–W406, 2024. Disponível em: https://doi.org/10.1093/nar/gkae253

PARIHAR, R.; SHUKLA, R.; BAISHYA, B.; KALITA, J.; HALDAR, R.; MISRA, U. K. NMR based CSF metabolomics in tuberculous meningitis: correlation with clinical and

MRI findings. **Metabolic Brain Disease**, v. 37, n. 3, p. 773–785, 2022. Disponível em: https://doi.org/10.1007/s11011-021-00860-y

PEDREGOSA, F. *et al.* **Scikit-Learn**. 2011a. Disponivel em: https://scikitlearn.org/stable. Acesso em: 9 jul. 2024.

PICKKERS, P.; DARMON, M.; HOSTE, E.; JOANNIDIS, M.; LEGRAND, M.; OSTERMANN, M.; PROWLE, J. R.; SCHNEIDER, A.; SCHETZ, M. Acute kidney injury in the critically ill: an updated review on pathophysiology and management. **Intensive care medicine**, v. 47, p. 835-850, 2021. Disponível em: https://doi.org/10.1007/s00134-021-06454-7

POLO, T. C. F.; MIOT, H. A. Use of roc curves in clinical and experimental studies. **Jornal Vascular Brasileiro**, v. 19, e20200186, 2020. Disponível em: https://doi.org/10.1590/1677-5449.200186

PONTES, T. A.; BARBOSA, A. D.; SILVA, R. D.; MELO-JÚNIOR, S. R.; SILVA, R. O. Osteopenia-osteoporosis discrimination in postmenopausal women by 1H NMR-based metabonomics. **PLOS ONE**, v. 14, n. 5, e0217348, 2019. Disponível em: https://doi.org/10.1371/journal.pone.0217348.

QI, Y-S.; XIAO, H-Y.; XIE, P.; XIE, J. B.; GUO, M.; LI, F-F.; PIAO, X-L. Comprehensive serum metabolomics and network analysis to reveal the mechanism of gypenosides in treating lung cancer and enhancing the pharmacological effects of cisplatin. **Frontiers in pharmacology**, v. 13, 1070948, 2022. 10.3389/fphar.2022.1070948

QU, L.; PEI, Y. A Comprehensive Review on Discriminant Analysis for Addressing Challenges of Class-Level Limitations, Small Sample Size, and Robustness. **Processes**, v. 12, n. 7, p. 1382, 2024. Disponível em: https://doi.org/10.3390/pr12071382

QUAGLIA, M.; MERLOTTI, G.; GUGLIELMETTI, G.; CASTELLANO, G.; CANTALUPPI, V. Recent advances on biomarkers of early and late kidney graft dysfunction. **International Journal of Molecular Sciences**, v. 21, n. 15, 5404, 2020. Disponível em: https://doi.org/10.3390/ijms21155404

RAJA, G.; JUNG, Y.; JUNG, S. H.; KIM, T. J. ^{1}H-NMR-based metabolomics for cancer targeting and metabolic engineering – a review. **Process Biochemistry**, v. 99, p. 161–172, 2020. Disponível em: https://doi.org/10.1016/j.procbio.2020.08.023

RAZAVI, S. A.; MAHMANZAR, M.; NOBAKHT M. GH., B. F.; ZAMANI, Z.; NASIRI, S.; HEDAYATI, M. Plasma metabolites analysis of patients with papillary thyroid cancer: A preliminary untargeted 1H NMR-based metabolomics. **Journal of Pharmaceutical and Biomedical Analysis**, v. 241, 2024. Disponível em: https://doi.org/10.1016/j.jpba.2023.115946

REIS-NETO, E.; SEGURO, L.; SATO, E.; BORBA, E. F.; KLUMB, E. M.; COSTALLAT, L. T. L.; MEDEIROS, M. M. C.; BONFÁ, E.; ARAÚJO, N. C.; APPENZELLER, S.; MONTANDON, A. C. O. S.; YUKI, E. F. N.; TEIXEIRA, R. C. A.; TELLES, R. W.;

EGYPTO, D. C. S.; RIBEIRO, F. M.; GASPARIN, A. A.; JUNIOR, A. S. A.; NEIVA, C. L. S.; CALDERARO, D. C.; MONTICIELO, O. A. II Brazilian Society of Rheumatology consensus for lupus nephritis diagnosis and treatment. **Advances in Rheumatology**, v. 64, 2024. Disponível em: https://doi.org/10.1186/s42358-024-00386-8

REY-STOLLE, F.; DUDZIK, D.; GONZALEZ-RIANO, C.; FERNÁNDEZ-GARCÍA, M., ALONSO-HERRANZ, V., ROJO, D.; BARBAS, C.; GARCÍA, A. Low and high resolution gas chromatography-mass spectrometry for untargeted metabolomics: A tutorial**.** **Analytica chimica acta**, v. 1210, 339043, 2022. Disponível em: doi:10.1016/j.aca.2021.339043

RODRIGUES, A. P.; LUNA, A. S.; PINTO, L. An evaluation strategy to select and discard sampling preprocessing methods for imbalanced datasets: A focus on classification models. **Chemometrics and Intelligent Laboratory Systems**, v. 240, 104933, 2023. Disponível em: doi.org/10.1016/j.chemolab.2023.104933.

RODRIGUES, M. L.; DA LUZ, T. P. S. R.; PEREIRA, C. L. D.; BATISTA, A. D.; DOMINGUES, A. L. C.; SILVA, R. O.; LOPES, E. P. Assessment of periportal fibrosis in Schistosomiasis mansoni patients by proton nuclear magnetic resonance-based metabonomics models. **World Journal of Hepatology**, v. 14, n. 4, p. 719–728, 2022. Disponível em: https://doi.org/10.4254/wjh.v14.i4.719

RODRIGUES, M. L.; GOIS, A. R. S.; DOMINGUES, A. L. C.; et al. Metabolomics assays applied to schistosomiasis studies: a scoping review. **BMC Infectious Diseases**, v. 25, 2025. Disponível em: https://doi.org/10.1186/s12879-025-10606-1

RONG, L.; YE, F.; ZHONG, Q. P.; WANG, S. H.; CHAI, T.; DONG, H. F. et al. Comparative serum metabolomics between SCID mice and BALB/c mice with or without Schistosoma japonicum infection: clues to the abnormal growth and development of schistosome in SCID mice. **Acta Tropica**, v. 200, 105186, 2019. Disponível em: https://doi.org/10.1016/j.actatropica.2019.105186

ROY, A.; CHAKRABORTY, S. Support vector machine in structural reliability analysis: a review. **Reliability Engineering & System Safety**, v. 239, 109126, 2023.Disponível em: https://doi.org/10.1016/j.ress.2023.109126

RULE, G. S.; HITCHENS, T. K. NMR spectroscopy. In: RULE, G. S.; HITCHENS, T. K. (Ed.). **Fundamentals of Protein NMR Spectroscopy**, v. 5. Holanda: Springer, 2006. p. 1–27. Disponível em: https://doi.org/10.1007/1-4020-3500-4

RYU, J. H.; KIM, S. H.; KANG, S. H.; et al. Better health-related quality of life in kidney transplant patients compared to chronic kidney disease patients with similar renal function. **PLoS ONE**, v. 16, n. 10, e0257981, 2021. Disponível em: https://doi.org/10.1371/journal.pone.0257981

SAIGUSA, D.; MATSUKAWA, N.; HISHINUMA, E.; KOSHIBA, S. Identification of biomarkers to diagnose diseases and find adverse drug reactions by metabolomics.

Drug Metabolism and Pharmacokinetics, v. 36, p. 100392, 2021.Disponível em: https://doi.org/10.1016/j.dmpk.2020.11.008

SAMY, N.; HAMZA, S. H.; MABROUK, F. I.; ATYAALLAH, D. F.; MOHAMMED, M. M. Class V membranous nephropathy in patients with systemic lupus erythematosus: Clinical characteristics, outcome and prognosis. **The Egyptian Rheumatologist**, v. 47, p. 207-211, Issue 4, 2025. Disponível em: https://doi.org/10.1016/j.ejr.2025.07.003.

SANTOS, J.C.; PEREIRA, C. L. D.; DOMINGUES, A. L.C.; LOPES, E. P. Noninvasive diagnosis of periportal fibrosis in schistosomiasis mansoni: A comprehensive review. **World J Hepatol**, v. 14, p. 696-707, 2022. Disponível em: 10.4254/wjh.v14.i4.696.

SAVELIEV, M.; PANCHUK, V.; KIRSANOV, D. Math is greener than chemistry: assessing green chemistry impact of chemometrics. **TrAC - Trends in Analytical Chemistry**, v. 202, 117556, 2024.Disponível em: https://doi.org/10.1016/j.trac.2024.117556

SAVORANI, F.; TOMASI, G.; ENGELSEN, S. B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. **Journal of Magnetic Resonance**, v. 202, n. 2, p. 190–202, 2010. Disponível em: https://doi.org/10.1016/j.jmr.2009.11.012

SEABOLD, S.; PERKTOLD, J. Statsmodels: Econometric and Statistical Modeling with Python. In: Proceedings of the 9th Python in Science Conference (SciPy 2010). [S.l.]: [s.n.], 2010.

SHAO, S.; YANG, L.; HU, G.; LI, L.; WANG, Y.; TAO, L. Application of omics techniques in forensic entomology research. **Acta Tropica**, v. 249, 106985, 2023. Disponível em: https://doi.org/10.1016/j.actatropica.2023.106985

SHARABI, A.; TSOKOS, G. C. T cell metabolism: new insights in systemic lupus erythematosus pathogenesis and therapy. **Nature Reviews Rheumatology**, v. 16, p. 100–112, 2020. Disponível em: https://doi.org/10.1038/s41584-019-0356-x

SHARMA, A.; KUMAR, R.; KUMAR, N.; SAXENA, V. Machine learning driven portable Vis-SWNIR spectrophotometer for non-destructive classification of raw tomatoes based on lycopene content. **Vibrational Spectroscopy**, v. 130, 2024. Disponível em: https://doi.org/10.1016/j.vibspec.2023.103628

SHEN, Y.; WU, S. D.; CHEN, Y.; LI, X. Y.; ZHU, Q.; NAKAYAMA, K.; ZHANG, W. Q.; WENG, C. Z.; ZHANG, J.; WANG, H. K.; WU, J.; JIANG, W. Alterations in gut microbiome and metabolomics in chronic hepatitis B infection-associated liver disease and their impact on peripheral immune response. **Gut Microbes**, v. 15, n. 1, 2023. Disponível em: https://doi.org/10.1080/19490976.2022.2155018

SILVA, F. L.; DEL-REI, R. P.; MOTHÉ FRAGA, D. B.; LEONY, L. M.; GONZAGA CARLOS DE SOUZA, A. M.; SANTOS, F. L. N. Alterations in the lipid profiles and circulating liver enzymes in individuals infected by Schistosoma mansoni. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 51, n. 6, p. 795–801, 2018. Disponível em: https://doi.org/10.1590/0037-8682-0113-2018

SILVA, R. D. da. **Aplicacoes Metabonomicas usando Ressonancia Magnetica Nuclear de 1H: Diagnostico Nao-Invasivo de Cancer de Prostata e Urologico & Classificacao de Azeite de Oliva Extra Virgem de Producao Organica.** 2017. Tese (Doutorado em Química) - Universidade Federal de Pernambuco, Recife, 2017. Disponivel em: https://repositorio.ufpe.br/handle/123456789/24852

SILVA, R. E.; BALDIM, J. L.; CHAGAS-PAULA, D. A.; SOARES, M. G.; LAGO, J. H. G.; GONÇALVES, R. V.; NOVAES, R. D. Predictive metabolomic signatures of end-stage renal disease: A multivariate analysis of population-based data. **Biochimie**, v. 152, p. 14-30, 2018. Disponível em: https://doi.org/10.1016/j.biochi.2018.06.009.

SILVA, R. O. da. **A espectroscopia de RMN como ferramenta elucidativa: Estruturas moleculares, mecanismos de reação e metabonômica.** 2010. Tese (Doutorado em Química) – Universidade Federal de Pernambuco, Recife, 2010. Disponível em: https://repositorio.ufpe.br/handle/123456789/9789

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. K. Spectrometric identification of organic compounds. 8. ed. Hoboken, NJ: Wiley, 2019.

SIMON, Q.; GAILLARD, F.; TCHEN, J.; BACHELET, D.; SACRÉ, K.; PEOC'H, K.; JOURDE-CHICHE, N.; DAUGAS, E.; CHARLES, N. Immune characterization of lupus nephritis patients undergoing dialysis. **Journal of Translational Autoimmunity**, v. 10, 100290, 2025. Disponível em: https://doi.org/10.1016/j.jtauto.2025.100290.

SINGH, U.; ALSUHAYMI, S.; AL-NEMI, R.; EMWAS, A. H.; JAREMKO, M. Compound-Specific 1D 1H NMR Pulse Sequence Selection for Metabolomics Analyses. **ACS Omega**, v. 8, n. 26, p. 23651–23663, 2023. Disponível em: https://doi.org/10.1021/acsomega.3c01688

SOONG, R. *et al.* Water suppression 101 for benchtop NMR–An accessible guide and primer including fully interactive training videos. **Journal of Magnetic Resonance Open**, v. 19, 2024. Disponível em: https://doi.org/10.1016/j.jmro.2024.100150

SOTELO-OROZCO, J.; CHEN, S. Y.; HERTZ-PICCIOTTO, I.; SLUPSKY, C. M. A Comparison of Serum and Plasma Blood Collection Tubes for the Integration of Epidemiological and Metabolomics Data. **Frontiers in Molecular Biosciences**, v. 8, 2021. Disponível em: https://doi.org/10.3389/fmolb.2021.682134

STANIMIROVA, I.; DASZYKOWSKI, M.; HOPKE, P. K. The role of chemometrics in improving clinical data analysis and diagnostics. **TrAC Trends in Analytical Chemistry**, [s.l.], v. 173, p. 117642, 2024. ISSN 0165-9936. Disponível em: https://doi.org/10.1016/j.trac.2024.117642.

STOJANOV, D.; LAZAROVA, E.; VELJKOVA, E.; RUBARTELLI, P.; GIACOMINI, M. Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population – Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy. **Journal of King Saud University - Science**, [s.l.],

v. 35, n. 3, p. 102573, 2023. ISSN 1018-3647. Disponível em: 10.1016/j.jksus.2023.102573.

STRAUSS, C.; BOOKE, H.; FORNI, L.; ZARBOCK, A. Biomarkers of acute kidney injury: From discovery to the future of clinical practice. **Journal of Clinical Anesthesia**, v. 95, 2024. Disponível em: https://doi.org/10.1016/j.jclinane.2024.111458

SUN, B.; FANG, Y.; YANG, H.; MENG, F.; HE, C.; ZHAO, Y.; ZHAO, K.; ZHANG, H. The combination of deep learning and pseudo-MS image improves the applicability of metabolomics to congenital heart defect prenatal screening. **Talanta**, v. 275, 2024. Disponível em: https://doi.org/10.1016/j.talanta.2024.126109

SUN, J.; XIA, Y. Pretreating and normalizing metabolomics data for statistical analysis. **KeAi Communications Co**., 2024. Disponível em: https://doi.org/10.1016/j.gendis.2023.04.018

TANTISATTAMO, E.; MOLNAR, M. Z.; HO, B. T.; REDDY, U. G.; DAFOE, D. C.; ICHII, H.; FERREY, A. J.; HANNA, R. M.; KALANTAR-ZADEH, K.; AMIN, A. Approach and Management of Hypertension After Kidney Transplantation. *[S. l.]*: **Frontiers Media S.A.**, 2020. Disponível em: https://doi.org/10.3389/fmed.2020.00229

TAWANA-NDOLO, S. M.; ZACHARIAH, M.; PHALADZE, N. A.; SICHILONGO, K. F. A solid-phase microextraction gas chromatography–mass spectrometry technique for urinary metabolomics of human samples infected with schistosomiasis—case of the Okavango Delta, Botswana. **Biomedical Chromatography**, v. 37, n. 11, e5718, 2023. Disponível em: https://doi.org/10.1002/bmc.5718

TEPEL, M.; ALKAFF, F. F.; KREMER, D.; BAKKER, S. J. L.; THAUNAT, O.; NAGARAJAH, S.; SALEH, Q.; BERGER, S. P.; VAN DEN BORN, J.; KROGSTRUP, N. V.; NIELSEN, M. B.; NØRREGAARD, R.; JESPERSEN, B.; SPARDING, N.; GENOVESE, F.; KARSDAL, M. A.; RASMUSSEN, D. G. K. Pretransplant endotrophin predicts delayed graft function after kidney transplantation. **Scientific Reports**, v. 12, n. 1, 2022. Disponível em: https://doi.org/10.1038/s41598-022-07645-y

TESFAYE, W.; PARRISH, N.; SUD, K.; GRANDINETTI, A.; CASTELINO, R. Medication Adherence Among Patients With Kidney Disease: An Umbrella Review. **W.B. Saunders**, 2024. Disponível em: https://doi.org/10.1053/j.akdh.2023.08.003

THARWAT, A. Parameter investigation of support vector machine classifier with kernel functions. **Knowledge and Information Systems**, v. 61, n. 3, p. 1269–1302, 2019. Disponível em: https://doi.org/10.1007/s10115-019-01335-4

ULRICH, E. L.; AKUTSU, H.; DORELEIJERS, J. F.; HARANO, Y.; IOANNIDIS, Y. E.; LIN, J.; LIVNY, M.; MADING, S.; MAZIUK, D.; MILLER, Z.; NAKATANI, E.; SCHULTE, C. F.; TOLMIE, D. E.; WENGER, R. K.; YAO, H.; MARKLEY, J. L. BioMagResBank. **Nucleic Acids Research**, v. 36, p. D402-D408, 2008. Disponível em: https://doi.org/10.1093/nar/gkm957

VACH, W.; GERKE, O. Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties. **MethodsX**, v. 10, 102212, 2023, Disponível em: https://doi.org/10.1016/j.mex.2023.102212.

VIGNOLI, A.; TENORI, L.; MORSIANI, C.; TURANO, P.; CAPRI, M.; LUCHINAT, C. Serum or Plasma (and Which Plasma), That Is the Question. **Journal of Proteome Research**, v. 21, n. 4, p. 1061–1072, 2022. Disponível em: https://doi.org/10.1021/acs.jproteome.1c00935

WANG, A.; QIN, Y.; XING, Y.; YU, Z.; HUANG, L.; YUAN, J.; HUI, Y.; HAN, M.; XU, G.; ZHAO, J.; SUN, S. Clinical characteristics, prognosis, and predictive modeling in class IV ± V lupus nephritis. **Frontiers in Immunology**, v. 16, 1580146, 2025. Disponível em: https://doi.org/10.3389/fimmu.2025.1580146.

WANG, D.; ZHU, J.; LI, N.; LU, H.; GAO, Y.; ZHUANG, L.; CHEN, Z.; MAO, W. GC-MS-based untargeted metabolic profiling of malignant mesothelioma plasma. **PeerJ**, v. 11, 2023 a. Disponível em: https://doi.org/10.7717/PEERJ.15302

WANG, K.; THEEKE, L. A.; LIAO, C.; WANG, N.; LU, Y.; XIAO, D.; XU, C. Deep learning analysis of UPLC-MS/MS-based metabolomics data to predict Alzheimer's disease. **Journal of the Neurological Sciences**, v. 453, 2023 b. Disponível em: https://doi.org/10.1016/j.jns.2023.120812

WANG, M.; XU, J.; YANG, N.; ZHANG, T.; ZHU, H.; WANG, J. Insight Into the Metabolomic Characteristics of Post-Transplant Diabetes Mellitus by the Integrated LC-MS and GC-MS Approach- Preliminary Study. **Frontiers in Endocrinology**, v. 12, 2022 a. Disponível em: https://doi.org/10.3389/fendo.2021.807318

WANG, S.; YANG, L.; HU, H.; LV, L.; JI, Z.; ZHAO, Y.; ZHANG, H.; XU, M.; FANG, R.; ZHENG, L.; DING, C.; YANG, M..; XU, K.; LI, L. Characteristic gut microbiota and metabolic changes in patients with pulmonary tuberculosis. **Microbial Biotechnology**, v. 15, n. 1, p. 262–275, 2022 b. Disponível em: https://doi.org/10.1111/1751-7915.13761

WANG, Y.; HEMMELDER, M. H.; BOS, W. J. W.; SNOEP, J. D.; DE VRIES, A. P. J.; DEKKER, F. W.; MEULEMAN, Y. Mapping health-related quality of life after kidney transplantation by group comparisons: A systematic review. **Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association - European Renal Association**, v. 36, p. 2327-2339, 2021. Disponível em: https://doi.org/10.1093/ndt/gfab232

WASKOM, M. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021. Disponivel em: https://doi.org/10.21105/joss.03021

WISHART, D. S.; GUO, A.; OLER, E.; WANG, F.; ANJUM, A.; PETERS, H.; DIZON, R.; SAYEEDA, Z.; TIAN, S.; LEE, B. L.; BERJANSKII, M.; MAH, R.; YAMAMOTO, M.; JOVEL, J.; TORRES-CALZADA, C.; HIEBERT-GIESBRECHT, M.; LUI, V. W.;

VARSHAVI, D.; VARSHAVI, D.; ALLEN, D.; ARNDT, D.; KHETARPAL, N.; SIVAKUMARAN, A.; HARFORD, K.; SANFORD, S.; YEE, K.; CAO, X.; BUDINSKI, Z.; LIIGAND, J.; ZHANG, L.; ZHENG, J.; MANDAL, R.; KARU, N.; DAMBROVA, M.; SCHIÖTH, H. B.; GREINER, R.; GAUTAM, V. HMDB 5.0: the Human Metabolome Database for 2022. **Nucleic Acids Research**, v. 50, p. D622–D631, 2022. Disponível em: https://doi.org/10.1093/nar/gkab1062

WORLD HEALTH ORGANIZATION. Schistosomiasis: progress report 2001–2011 and strategic plan 2012–2020. Geneva: World Health Organization, 2013. 74 p. Disponível em: https://iris.who.int/handle/10665/78074 [acesso em: 12 jun. 2025].

WORLD HEALTH ORGANIZATION. WHO guideline on control and elimination of human schistosomiasis. Geneva: World Health Organization, 2022. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK578392/ [acesso em: 12 jun. 2025].

WORLEY, B.; POWERS, R. Simultaneous phase and scatter correction for NMR datasets. **Chemometrics and Intelligent Laboratory Systems**, v. 131, p. 1–6, 2014. Disponível em: https://doi.org/10.1016/j.chemolab.2013.11.005

XU, B.; LI, W.; ZHANG, Y.; CHEN, Y.; FENG, J.; SONG, X. Untargeted and spatial-resolved metabolomics characterize serum and tissue-specific metabolic reprogramming in acute kidney injury. **Heliyon**, v. 9, n. 11, 2023 a. Disponível em: https://doi.org/10.1016/j.heliyon.2023.e21171

XU, M.; XU, C.; CHEN, M.; XIAO, Z.; WANG, Y.; XU, Y.; XU, D. **Comparative analysis of commonly used bioinformatics software based on omics**. *[S. l.]*: Elsevier Inc., 2023 b. Disponível em: https://doi.org/10.1016/j.genrep.2023.101800

YANG, Q. J. *et al.* Serum and urine metabolomics study reveals a distinct diagnostic model for cancer cachexia. **Journal of Cachexia, Sarcopenia and Muscle**, v. 9, n. 1, p. 71–85, 2018. Disponível em: https://doi.org/10.1002/jcsm.12246

YANG, T.; HUI, R.; NOUWS, J.; SAULER, M.; ZENG, T.; WU, Q. **Untargeted metabolomics analysis of esophageal squamous cell cancer progression**. *[S. l.]*: BioMed Central Ltd, 2022. Disponível em: https://doi.org/10.1186/s12967-022-03311-z

YANG, T.; ZHOU, J.; FANG, L.; WANG, M.; DILINUER, M.; AINIWAER, A. Protection function of 18β-glycyrrhetinic acid on rats with high-altitude pulmonary hypertension based on 1H NMR metabonomics technology. **Analytical Biochemistry**, v. 631, 2021. Disponível em: https://doi.org/10.1016/j.ab.2021.114342

YANLAN, C. et al. A multi-omics approach based on 1H-NMR metabonomics combined with target protein analysis to reveal the mechanism of RIAISs on cervical carcinoma patients. **Aging** (Albany NY), [s.l.], v. 15, n. 6, p. 1878-1889, 27 set. 2022. DOI: 10.18632/aging.204305. Disponível em: https://doi.org/10.18632/aging.204305.

YUAN, W.; CHEN, Y.; ZHUANG, D.; ZENG, H.; LIN, X.; HONG, S.; LIN, F.; CHEN, X.; HUANG, S.; LIN, F. UHPLC-MS/MS-based central carbon metabolism unveils the

biomarkers related to colon cancer. **Cellular and Molecular Biology**, v. 69, n. 9, p. 167–171, 2023. Disponível em: https://doi.org/10.14715/cmb/2023.69.9.25

ZENG, S.; LIU, Z.; KE, B.; ZHANG, Y.; WANG, Q.; TAN, S. The non-invasive serum biomarkers contributes to indicate liver fibrosis staging and evaluate the progress of chronic hepatitis B. **BMC Infectious Diseases**, v. 24, n. 1, 2024. Disponível em: https://doi.org/10.1186/s12879-024-09465-z.

ZHANG, H.; FU, Q.; LIU, J.; LI, J.; DENG, R.; WU, C.; NIE, W.; CHEN, X.; LIU, L.; WANG, C. Risk factors and outcomes of prolonged recovery from delayed graft function after deceased kidney transplantation. **Renal Failure**, v. 42, n. 1, p. 792–798, 2020. Disponível em: https://doi.org/10.1080/0886022X.2020.1803084

ZHANG, H.; WANG, L.; YIN, D.; ZHOU, Q.; LV, L.; DONG, Z.; SHI, Y. Integration of proteomic and metabolomic characterization in atrial fibrillation-induced heart failure. **BMC Genomics**, v. 23, n. 1, 2022. Disponível em: https://doi.org/10.1186/s12864-022-09044-z

ZHANG, J.; YU, X.; XIE, Z.; WANG, R.; LI, H.; TANG, Z. F.; NA, N. A bibliometric and knowledge-map analysis of antibody-mediated rejection in kidney transplantation. **Renal Failure**, v. 45, n. 2, 2023. Disponível em: https://doi.org/10.1080/0886022X.2023.2257804

ZHENG, G.; PRICE, W. S. Solvent signal suppression in NMR, **Progress in Nuclear Magnetic Resonance Spectroscopy,** v. 56, p. 267-288, 2010. Disponível em: https://doi.org/10.1016/j.pnmrs.2010.01.001

ZHENG, G.; YUE, X.; YI, W.; JIA, R. Establishment, interpretation and application of logistic regression models for predicting thermal sensation of elderly people. **Energy and Buildings**, v. 315, 2024. Disponível em: https://doi.org/10.1016/j.enbuild.2024.114318

ZHENG, H.; ZHENG, P.; ZHAO, L.; JIA, J.; TANG, S.; XU, P.; XIE, P.; GAO, H. Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. **Clinica Chimica Acta**, v. 464, p. 223–227, 2017. Disponível em: https://doi.org/10.1016/j.cca.2016.11.039

ZHONG, Y.; SHI, J.; ZHENG, Z.; NAWAZ, M. A.; CHEN, C.; CHENG, F.; KONG, Q.; BIE, Z.; HUANG, Y. NMR-based fruit metabonomic analysis of watermelon grafted onto different rootstocks under two potassium levels. **Scientia Horticulturae**, v. 258, 2019. Disponível em: https://doi.org/10.1016/j.scienta.2019.108793

ZHOU, C.; LI, J.; GUO, C.; ZHOU, Z.; YANG, Z.; ZHANG, Y. et al. Alterations in gut microbiome and metabolite profile of patients with Schistosoma japonicum infection. **Parasites & Vectors**, v. 16, n. 1, 346, 2023. Disponível em: https://doi.org/10.1186/s13071-023-05932-9

ZHU, X.; CHEN, L.; WU, J.; TANG, H.; WANG, Y. Salmonella typhimurium infection reduces Schistosoma japonicum worm burden in mice. **Scientific Reports**, v. 7, n. 1, 1349, 2017. Disponível em: https://doi.org/10.1038/s41598-017-01437-5

APÊNDICE A – PARÂMETROS DOS MODELOS

Tabela A 1 – Parâmetros utilizados na otimização dos algoritmos pelo método GridSearchCV().

Algoritmo	*Parâmetros*	*Valores otimizados*
SVM	C	[0.001, 0.01, 0.1, 1, 10, 100]
	kernel	["linear", "rbf", "poly"]
	gamma	["scale", "auto", 0.001, 0.01, 0.1, 1]
LDA	Encolhimento	[sim; não]
	Solucionador	["svd", "lsqr", "eigen"]
LR	C	[0.01, 0.1, 1, 10, 100]
DT	máx. profundidade	[1, 2, 3, 4, 5, 6, 8, 10]
	nº min. de amostras por folha	[1, 2, 3, 4, 5, 6]

Mais detalhes sobre os parâmetros e os algoritmos podem ser acessados na referência PEDREGOSA et al., 2011a.

Tabela A 2 – Parâmetros de cada modelo de classificação após a otimização.

Conjunto de dados	*Algoritmos*	*Seletor*	*Parâmetros*
LRA	SVM	-	C: 1, kernel:"linear"
	LDA	-	shrinkage: "auto", solver: "lsqr"
	LR	-	C: 1, solver:'lbfgs'
NL	SVM	-	C: 1, gamma: 0.1, kernel: 'rbf'
	LDA	-	solver: 'svd'
	LR	-	C: 1, solver:'lbfgs'
	SVM	SFM-LR	C: 1, gamma: 0.1, kernel: 'rbf'
	LDA	SFM-LR	shrinkage:"auto", solver:"lsqr"
	LR	SFM-LR	C: 100, solver: 'lbfgs'
	SVM	SFS-LR	C: 10, gamma: 0.1, kernel: 'rbf'
	LDA	SFS-LR	shrinkage:"auto", solver:"lsqr"
	LR	SFS-LR	C: 1, solver:'lbfgs'
	SVM	SFM-RF	C: 1, gamma: 0.1, kernel: 'rbf'
	LDA	SFM-RF	shrinkage:"auto", solver:"lsqr"
	LR	SFM-RF	C: 0.01, solver: 'lbfgs'
	SVM	SFS-RF	C: 1, gamma: 0,1, kernel: 'rbf'
	LDA	SFS-RF	shrinkage:"auto", solver:"lsqr
	LR	SFS-RF	C: 1, solver:'lbfgs'
	SVM	GA	C: 1.0, kernel: 'rbf'
	LDA	GA	solver: 'svd'
	LR	GA	C: 1.0, solver: 'lbfgs'
FPP	SVM		C: 0.01, kernel: "linear"
	LDA		solver: 'svd'
	LR		C: 0.1, solver: 'lbfgs'
	DT		criterion: 'gini', max_depth: 5, min_samples_leaf: 3

APÊNDICE B – GRÁFICOS DE PESOS DA PCA: LRA.

Figura B1 – Gráficos de pesos (*loadings*). a) PC1; b) PC2.

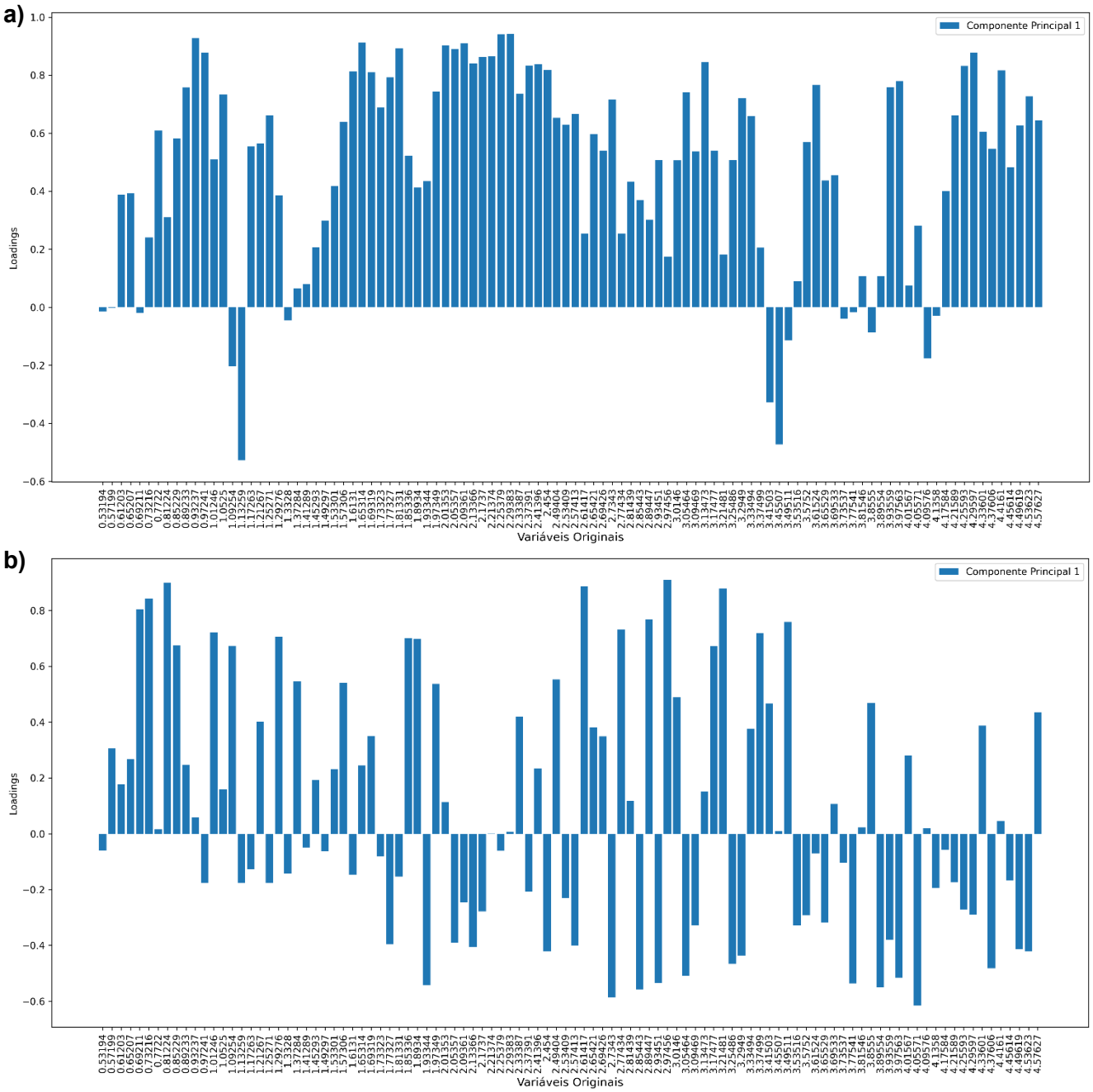
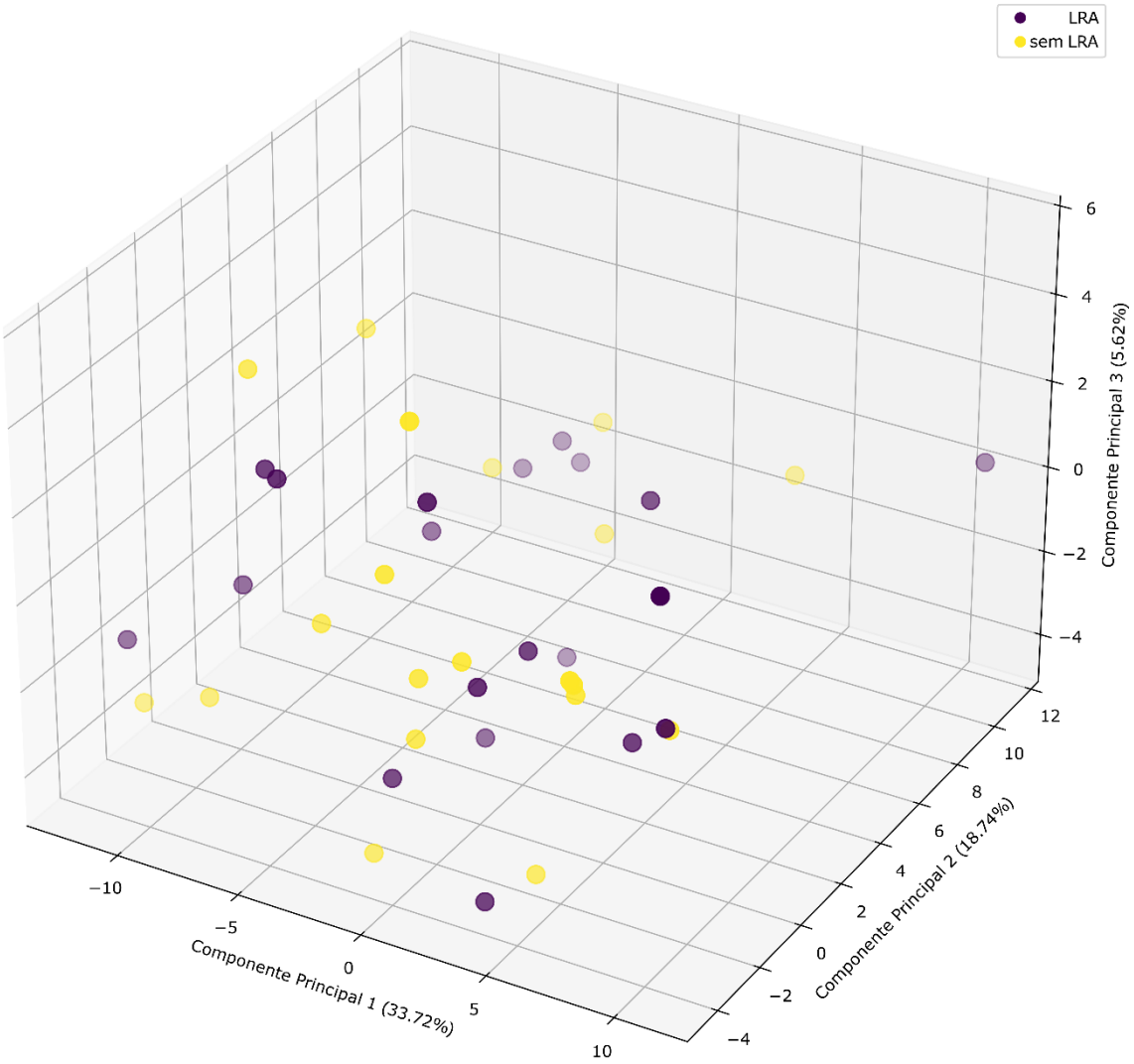


Figura B2 - Gráfico de escores 3D.



APÊNDICE C – TESTE DE KOLMOGOROV-SMIRNOV: LRA

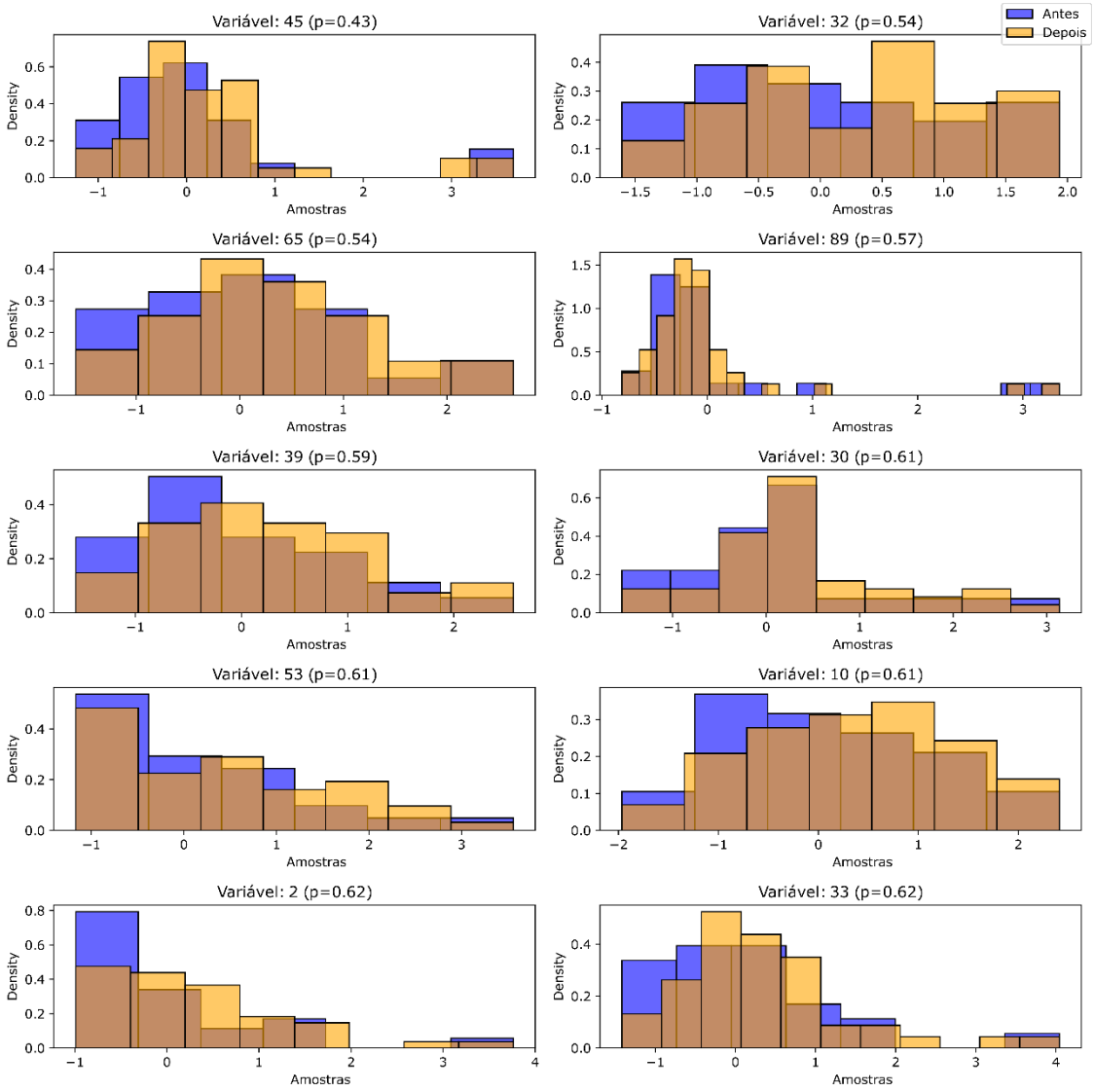
Tabela C1 - Resultados do teste de Kolmogorov-Smirnov (KS).

Índice	Variável	D (KS)	p_valor	Índice	Variável	D (KS)	p_valor
0	0.53194	0.095318	0.992405	51	2.57413	0.115385	0.954587
1	0.57199	0.073579	0.999832	52	2.61417	0.135452	0.869359
2	0.61203	0.173913	0.620789	53	2.65421	0.175585	0.609575
3	0.65207	0.120401	0.936742	54	2.69426	0.14214	0.830821
4	0.69211	0.105351	0.980507	55	2.7343	0.115385	0.954587
5	0.73216	0.070234	0.999934	56	2.77434	0.130435	0.896505
6	0.7722	0.118729	0.94357	57	2.81439	0.128763	0.90547
7	0.81224	0.128763	0.90547	58	2.85443	0.091973	0.995148
8	0.85229	0.168896	0.653915	59	2.89447	0.117057	0.949879
9	0.89233	0.168896	0.653915	60	2.93451	0.143813	0.824894
10	0.93237	0.175585	0.609575	61	2.97456	0.113712	0.960102
11	0.97241	0.14214	0.830821	62	3.0146	0.117057	0.949879
12	1.01246	0.130435	0.896505	63	3.05464	0.170569	0.646525
13	1.0525	0.108696	0.973425	64	3.09469	0.147157	0.80083
14	1.09254	0.14214	0.830821	65	3.13473	0.185619	0.539818
15	1.13259	0.083612	0.998741	66	3.17477	0.132107	0.888197
16	1.17263	0.152174	0.770043	67	3.21481	0.167224	0.668392
17	1.21267	0.118729	0.94357	68	3.25486	0.125418	0.917873
18	1.25271	0.152174	0.770043	69	3.2949	0.117057	0.949879
19	1.29276	0.150502	0.783088	70	3.33494	0.133779	0.879598
20	1.3328	0.153846	0.759172	71	3.37499	0.152174	0.770043
21	1.37284	0.152174	0.770043	72	3.41503	0.107023	0.977148
22	1.41289	0.150502	0.783088	73	3.45507	0.118729	0.94357
23	1.45293	0.108696	0.973425	74	3.49511	0.100334	0.987428
24	1.49297	0.118729	0.94357	75	3.53516	0.147157	0.80083
25	1.53301	0.145485	0.813272	76	3.5752	0.162207	0.702762
26	1.57306	0.128763	0.90547	77	3.61524	0.147157	0.80083
27	1.6131	0.157191	0.736464	78	3.65529	0.118729	0.94357
28	1.65314	0.157191	0.736464	79	3.69533	0.14214	0.830821
29	1.69319	0.167224	0.668392	80	3.73537	0.168896	0.653915
30	1.73323	0.175585	0.609575	81	3.77541	0.098662	0.988559
31	1.77327	0.138796	0.853481	82	3.81546	0.130435	0.896505
32	1.81331	0.185619	0.539818	83	3.8555	0.100334	0.987428
33	1.85336	0.173913	0.620789	84	3.89554	0.153846	0.759172
34	1.8934	0.145485	0.813272	85	3.93559	0.133779	0.879598
35	1.93344	0.113712	0.960102	86	3.97563	0.148829	0.794511
36	1.97349	0.158863	0.723471	87	4.01567	0.09699	0.990622
37	2.01353	0.16388	0.688755	88	4.05571	0.12709	0.913739
38	2.05357	0.145485	0.813272	89	4.09576	0.180602	0.574729
39	2.09361	0.17893	0.587849	90	4.1358	0.100334	0.987428
40	2.13366	0.132107	0.888197	91	4.17584	0.167224	0.668392

41	2.1737	0.158863	0.723471	92	4.21589	0.14214	0.830821
42	2.21374	0.132107	0.888197	93	4.25593	0.170569	0.646525
43	2.25379	0.167224	0.668392	94	4.29597	0.153846	0.759172
44	2.29383	0.160535	0.716412	95	4.33601	0.147157	0.80083
45	2.33387	0.202341	0.434622	96	4.37606	0.137124	0.858953
46	2.37391	0.145485	0.813272	97	4.4161	0.168896	0.653915
47	2.41396	0.167224	0.668392	98	4.45614	0.135452	0.869359
48	2.454	0.128763	0.90547	99	4.49619	0.147157	0.80083
49	2.49404	0.148829	0.794511	100	4.53623	0.133779	0.879598
50	2.53409	0.145485	0.813272	101	4.57627	0.123746	0.925838

p-valor > 0,05 sugere que as distribuições das variáveis antes e depois do SMOTE são estatisticamente semelhantes.

Figura C1 - Histogramas das variáveis com as maiores diferenças nas distribuições antes e depois do SMOTE, selecionadas pelas menores p-valores do teste de KS.



APÊNDICE D – GRÁFICOS DE ESCORES E PESOS DA PCA: NL

Figura D1- Gráfico de escores da PCA nos dados de NL. a) PC1 vs. PC2; b) PC1 vs. PC3; c) PC2 vs. PC3.

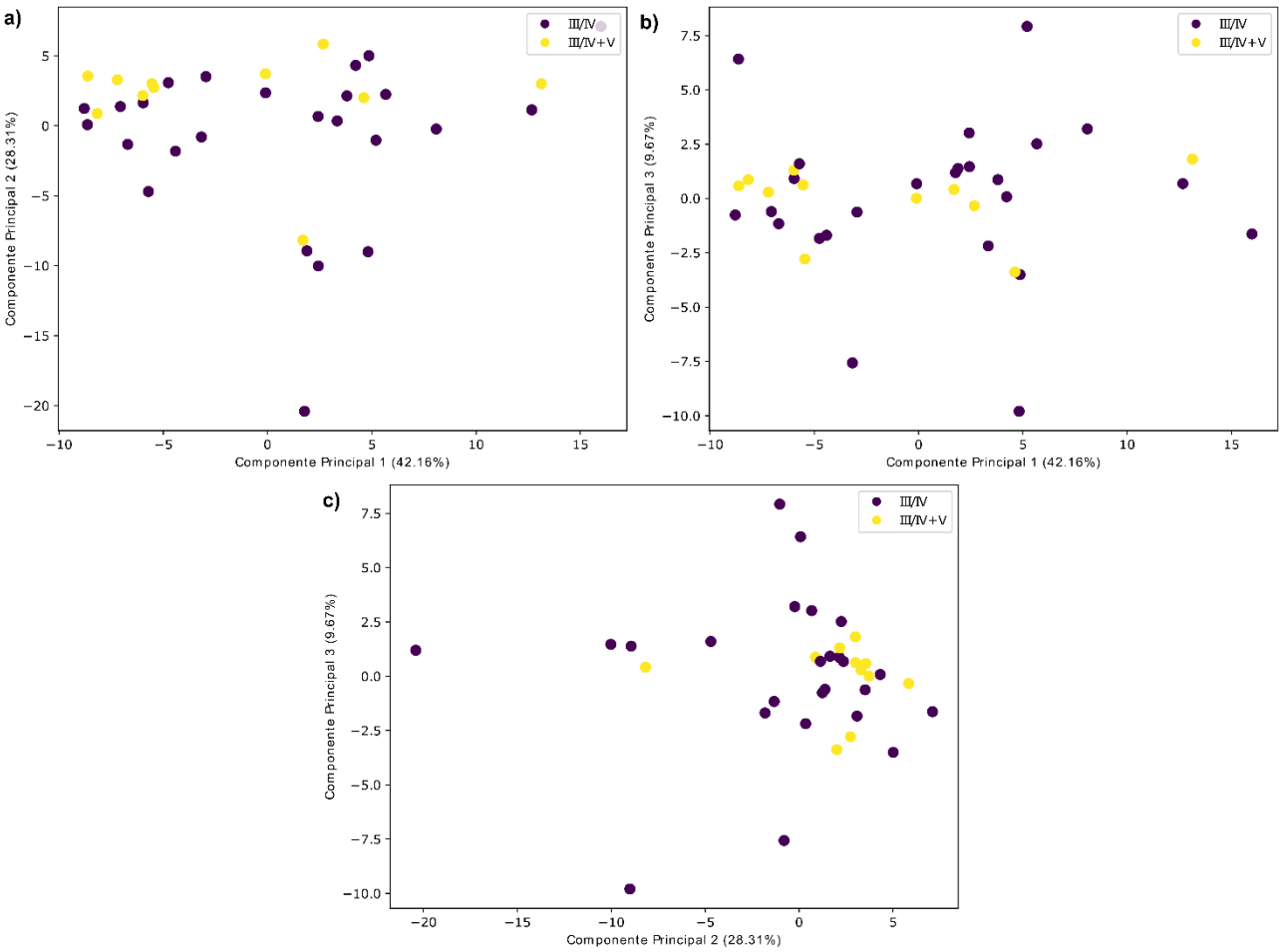
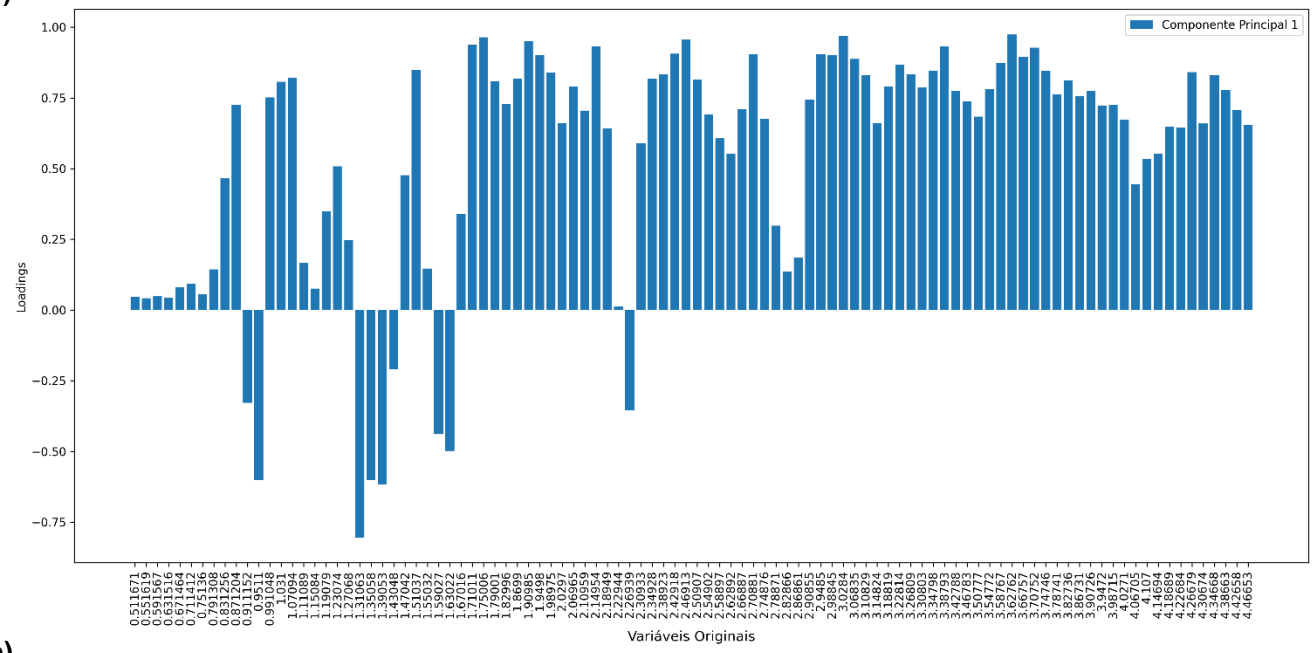
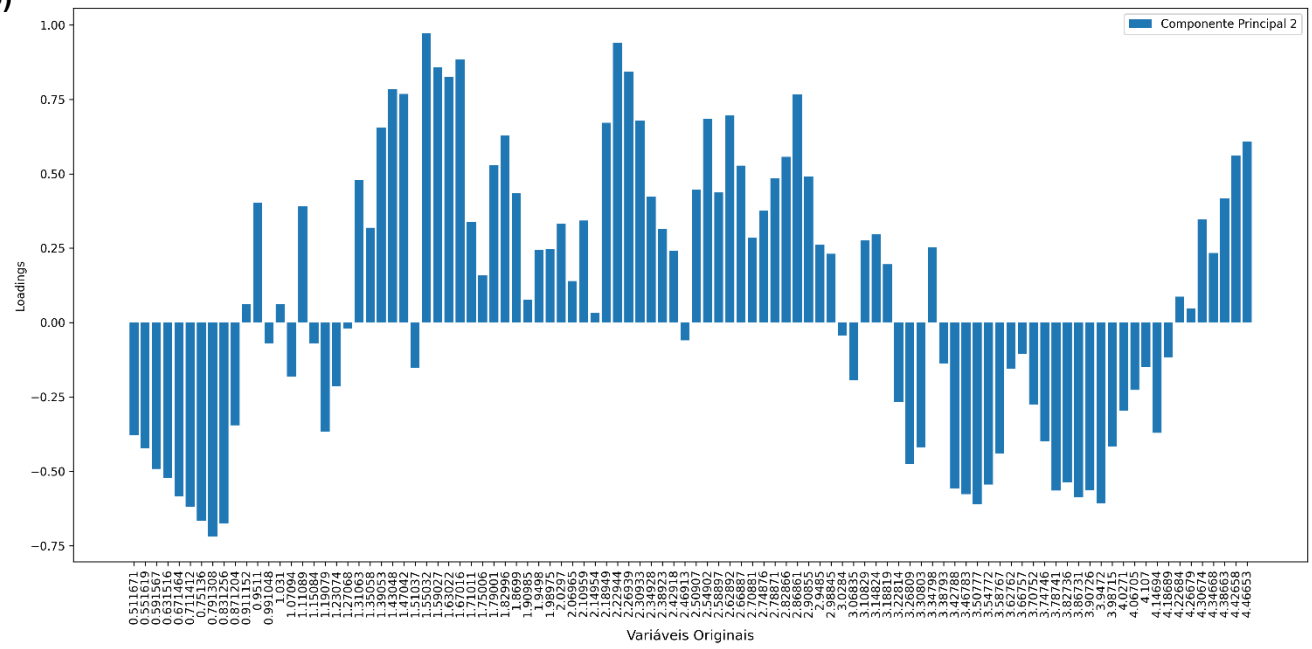


Figura D2 - Gráficos de pesos (loadings). a) PC1; b) PC2.

a)



b)



APÊNDICE E – TESTE DE KOLMOGOROV-SMIRNOV: NL

Tabela E1 - Resultados do teste de Kolmogorov-Smirnov (KS).

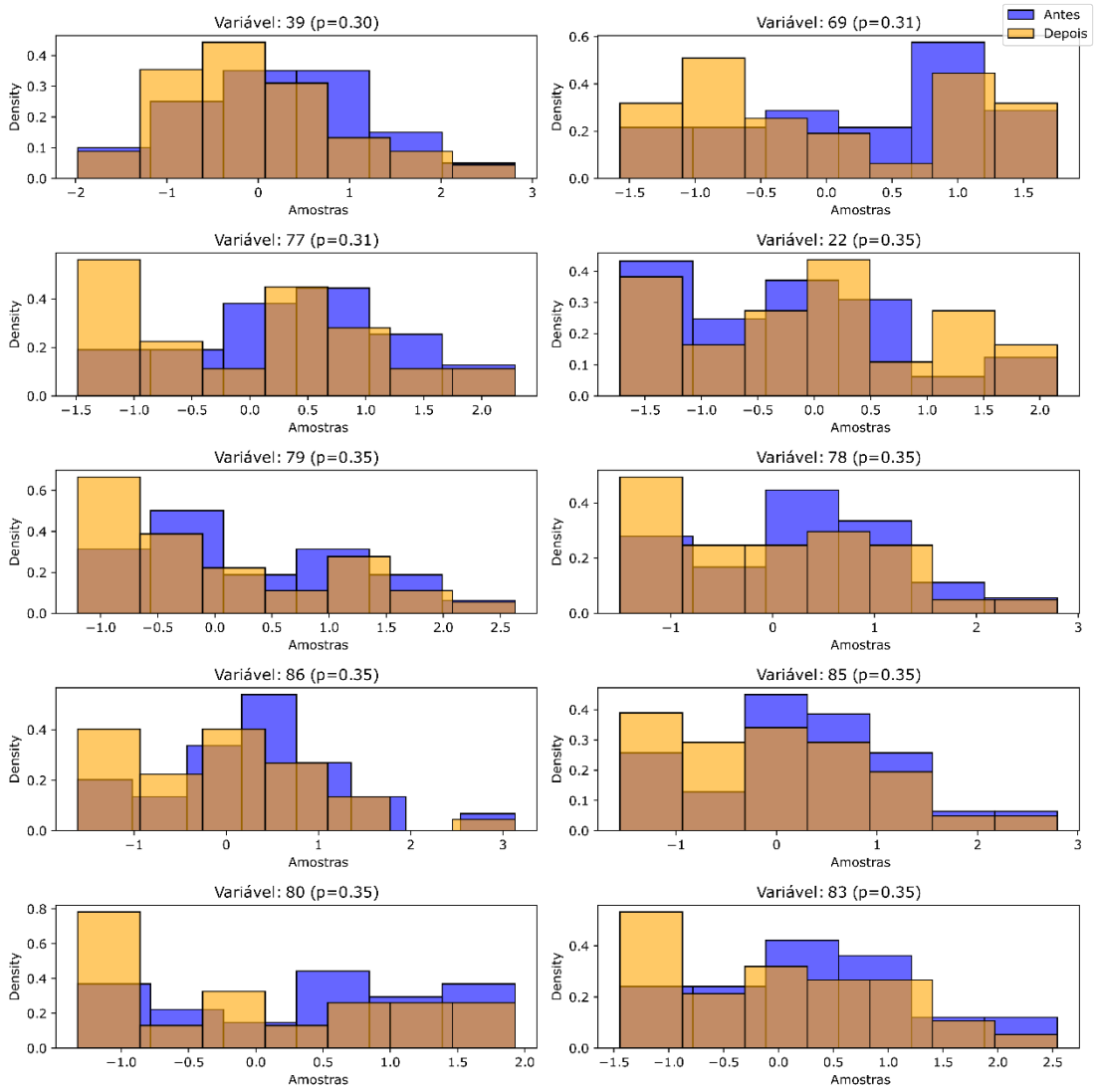
Índice	Variável	D (KS)	p_valor	Índice	Variável	D (KS)	p_valor
0	0.511671	0.084848	0.999364	51	2.54902	0.135758	0.912112
1	0.551619	0.077576	0.999868	52	2.58897	0.135758	0.912112
2	0.591567	0.117576	0.969994	53	2.62892	0.135758	0.912112
3	0.631516	0.107879	0.985928	54	2.66887	0.105455	0.989028
4	0.671464	0.077576	0.999868	55	2.70881	0.146667	0.861036
5	0.711412	0.087273	0.999002	56	2.74876	0.146667	0.861036
6	0.75136	0.077576	0.999868	57	2.78871	0.087273	0.999002
7	0.791308	0.087273	0.999002	58	2.82866	0.163636	0.770751
8	0.831256	0.166061	0.752737	59	2.86861	0.164848	0.761848
9	0.871204	0.224242	0.40273	60	2.90855	0.126061	0.946338
10	0.911152	0.059394	1.00000	61	2.9485	0.174545	0.700867
11	0.9511	0.164848	0.761848	62	2.98845	0.174545	0.700867
12	0.991048	0.115152	0.975347	63	3.0284	0.203636	0.521032
13	1.031	0.145455	0.868193	64	3.06835	0.214545	0.454436
14	1.07094	0.174545	0.700867	65	3.10829	0.174545	0.700867
15	1.11089	0.115152	0.975347	66	3.14824	0.174545	0.700867
16	1.15084	0.106667	0.987527	67	3.18819	0.164848	0.761848
17	1.19079	0.084848	0.999364	68	3.22814	0.193939	0.580503
18	1.23074	0.144242	0.875207	69	3.26809	0.243636	0.309024
19	1.27068	0.066667	0.999993	70	3.30803	0.133333	0.923151
20	1.31063	0.184242	0.638462	71	3.34798	0.124848	0.950586
21	1.35058	0.206061	0.501553	72	3.38793	0.155152	0.817097
22	1.39053	0.235152	0.346206	73	3.42788	0.224242	0.40273
23	1.43048	0.174545	0.700867	74	3.46783	0.224242	0.40273
24	1.47042	0.087273	0.999002	75	3.50777	0.224242	0.40273
25	1.51037	0.153939	0.82525	76	3.54772	0.224242	0.40273
26	1.55032	0.106667	0.987527	77	3.58767	0.243636	0.309024
27	1.59027	0.184242	0.638462	78	3.62762	0.233939	0.354308
28	1.63022	0.214545	0.454436	79	3.66757	0.233939	0.354308
29	1.67016	0.146667	0.861036	80	3.70752	0.233939	0.354308
30	1.71011	0.145455	0.868193	81	3.74746	0.224242	0.40273
31	1.75006	0.203636	0.521032	82	3.78741	0.224242	0.40273
32	1.79001	0.126061	0.946338	83	3.82736	0.233939	0.354308
33	1.82996	0.135758	0.912112	84	3.86731	0.224242	0.40273
34	1.8699	0.185455	0.628445	85	3.90726	0.233939	0.354308
35	1.90985	0.174545	0.700867	86	3.9472	0.233939	0.354308
36	1.9498	0.145455	0.868193	87	3.98715	0.193939	0.580503
37	1.98975	0.164848	0.761848	88	4.0271	0.204848	0.511486
38	2.0297	0.225455	0.393754	89	4.06705	0.113939	0.977817
39	2.06965	0.244848	0.302126	90	4.107	0.116364	0.972693
40	2.10959	0.116364	0.972693	91	4.14694	0.145455	0.868193

41	2.14954	0.164848	0.761848
42	2.18949	0.095758	0.996071
43	2.22944	0.175758	0.691155
44	2.26939	0.166061	0.752737
45	2.30933	0.088485	0.998769
46	2.34928	0.135758	0.912112
47	2.38923	0.164848	0.761848
48	2.42918	0.156364	0.808776
49	2.46913	0.193939	0.580503
50	2.50907	0.124848	0.950586

92	4.18689	0.206061	0.501553
93	4.22684	0.116364	0.972693
94	4.26679	0.195152	0.570594
95	4.30674	0.156364	0.808776
96	4.34668	0.156364	0.808776
97	4.38663	0.13697	0.906355
98	4.42658	0.146667	0.861036
99	4.46653	0.117576	0.969994

p-valor > 0,05 sugere que as distribuições das variáveis antes e depois do SMOTE são estatisticamente semelhantes.

Figura E1 - Histogramas das variáveis com as maiores diferenças nas distribuições antes e depois do SMOTE, selecionadas pelas menores p-valores do teste de KS.



APÊNDICE F – NOTA DE IMPRENSA

QUIMIOMETRIA APLICADA À AVALIAÇÃO CLÍNICA DE LESÃO RENAL AGUDA, NEFRITE LÚPICA E ESQUISTOSSOMOSE

(Tese de Doutorado)

Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco

Doutorando: Antonia Regina dos Santos Gois (Bolsista Capes)

Orientador: Prof. Dr. Ricardo Oliveira da Silva

A abordagem metabonômica visa investigar alterações nos perfis de metabólitos associadas a condições fisiológicas ou patológicas, como doenças. A informação contida em biofluidos (sangue, urina, fezes, entre outros) pode ser utilizada para auxiliar no diagnóstico precoce, estadiamento e prognóstico de diversas enfermidades, tornando a metabonômica uma ferramenta minimamente invasiva aplicável na tomada de decisões clínicas. Para extrair essas informações, são empregadas técnicas analíticas, seguidas de algoritmos de aprendizado de máquina, que consistem em métodos matemáticos capazes de processar grandes conjuntos de dados e identificar padrões relevantes.

Nesta tese de doutorado, foram desenvolvidos modelos metabonômicos para diagnosticar Lesão Renal Aguda (LRA) e monitorar Nefrite Lúpica (NL) proliferativa e, de forma paralela, modelos quimiométricos para monitorar a Fibrose Periportal (FPP). O conjunto de dados LRA foi coletado na literatura e o primeiro a ser investigado, baseado em 40 amostras de bebês prematuro o modelo de diagnóstico apresentou exatidão de 86%, sensibilidade de 71,40% e especificidade de 100%. Para o estadiamento de NL proliferativa com e sem lesão membranosa, as amostras foram coletadas no Hospital das Clínicas (HC) da UFPE, analisadas e passaram para etapa de modelagem, na qual, o modelo de estadiamento apresentou exatidão de 92,3%, sensibilidade de 100% e especificidade de 85,7%. Para o estadiamento de FPP, dados de exame de rotina de 184 pacientes, do HC e de Jaboatão dos Guararapes, foram investigados, no qual o modelo construído apresentou exatidão de 86%, sensibilidade de 75% e especificidade de 96%.

A investigação em torno de diferentes algoritmos de aprendizado de máquina, associados com etapas de pré-processamento resultaram em modelos com bons desempenhos, mesmo enfrentando problemas com o número de amostras reduzido nos dados de LRA e NL. Os resultados mostraram o potencial dos modelos metabonômicos e de aprendizado de máquina como uma alternativa minimamente invasiva que, além de reduzir a necessidade de procedimentos invasivos, como biópsias renais, fornece dados sobre alterações metabólicas específicas das doenças.

ANEXO 1

1. PRIMEIRA PÁGINA DO ARTIGO PUBLICADO REFERENTE AO CAPÍTULO 3: Revisão sobre ensaios metabolômicos aplicados em estudos sobre esquistossomose.

Rodrigues et al. BMC Infectious Diseases (2025) 25:211
https://doi.org/10.1186/s12879-025-10606-1

BMC Infectious Diseases

SYSTEMATIC REVIEW Open Access

Metabolomics assays applied to schistosomiasis studies: a scoping review

Milena Lima Rodrigues[1], Antonia Regina dos Santos Gois[1], Ana Lúcia Coutinho Domingues[1,3], Ricardo Oliveira Silva[2] and Edmundo Pessoa Lopes[1,3*]

Abstract

Background Metabolomics is an analytical approach utilized to explore the metabolic profiles of biological systems. This process typically involves the application of techniques such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). In the case of schistosomiasis, metabolomics has been employed to identify potential diagnostic biomarkers, examine the host's metabolic response, and explore more effective therapeutic strategies. The objective of this scoping review is to assess the scope and characteristics of metabolomic research on schistosomiasis conducted over the past decade.

Methods To identify relevant original publications, a systematic search was conducted in the PubMed and Web of Science databases using the following search terms: ("Metabolomics" OR "Metabolomic" OR "Metabonomics" OR "Metabonomic") AND ("Schistosomiasis" OR "*Schistosoma*"). These terms were applied to the titles and abstracts of the publications, with a focus on the period from January 2014 to December 2024.

Results The initial search yielded 48 articles. However, after a thorough evaluation of the abstracts, 14 articles were selected based on the established inclusion criteria. The selection process is visually depicted in the PRISMA flowchart. The majority of the studies included in this review were conducted in China (7 articles) and Brazil (3 articles). Approximately two-thirds of the studies utilized animal models, with serum serving as biofluid in 66% of the studies. The findings of this scoping review suggest that chromatographic techniques coupled with mass spectrometry are predominantly used in metabolomic research on schistosomiasis, accounting for 75% of the studies. The identified metabolites are associated with metabolic pathways related to glycolysis, the TCA cycle, and amino acid metabolism, as well as demonstrating alterations resulting from intestinal dysbiosis observed during the infection. As exemplified by succinate and citrate, which are present in the alterations of energy pathways in *Schistosoma mansoni* and *Schistosoma japonicum* species. The serum levels of these metabolites are modified, reflecting the host's metabolic and immunological responses induced by the infections.

Conclusions These studies successfully elucidated the metabolic pathways and key metabolites involved in schistosomiasis. The findings are significant for the future identification of diagnostic biomarkers and the development of novel antiparasitic agents targeting *Schistosoma* species.

Clinical trial Not Applicable.

*Correspondence:
Edmundo Pessoa Lopes
epalopes@uol.com.br

Full list of author information is available at the end of the article

ANEXO 2

2. PRIMEIRA PÁGINA DO ARTIGO PUBLICADO EM ESTUDO PARALELO A TESE: Triagem Metabolômica de Amostras Fecais como Método Alternativo aos Ensaios Clínicos Iniciais de Alergia à Proteína do Leite de Vaca em Lactentes.

Brazilian Journal of Analytical Chemistry
Pre-publication – Accepted after peer review
doi: 10.30744/brjac.2179-3425.AR-187-2024

BrJAC

ARTICLE

Metabolomic Screening of Fecal Samples as an Alternative Method to Initial Clinical Trials of Allergy to Cow's Milk Protein on Infants

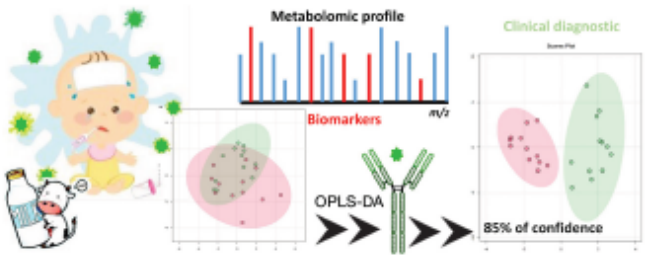
Kátia Suzane P. da Silva[1], Antonia Regina S. Gois[2], Tássia Brena B. C. da Costa[2], Wenes Ramos da Silva[3], Alberto Wisniewski Jr[3]*, Ricardo Oliveira Silva[2], Sarah Cristina F. Vieira[4], Jackeline Motta-Franco[4], Ricardo Queiroz Gurgel[1,4]

[1]*Programa de Pós-Graduação em Biologia Parasitária, Universidade Federal de Sergipe Av. Marcelo Déda Chagas, s/n, 49107-230 São Cristóvão, SE, Brazil*

[2]*Departamento de Química Fundamental, Universidade Federal de Pernambuco Av. Prof. Moraes Rego, 1235, Cidade Universitária, 50670-901 Recife, PE, Brazil*

[3]*Departamento de Química, Universidade Federal de Sergipe, Av. Marcelo Déda Chagas, s/n, 49107-230 São Cristóvão, SE, Brazil*

[4]*Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal de Sergipe, Rua Cláudio Batista s/n, Hospital Universitário, 49060-100 Aracaju, SE, Brazil*



Cow's Milk Protein Allergy (CMPA) is one of the most recurrent pediatric conditions of food allergies, which occur in early childhood. The diagnosis is not easy and demands oral food challenge tests (OFC). 24 metabolomic profiles of fecal samples from infants suspected of CMPA were assessed while searching for a possible intestinal metabolite that can be used as a biomarker of CMPA. The children were previously diagnosed with an open OFC. Feces samples were extracted and directly analyzed by ultra-high resolution mass spectrometry (HRMS) using an FT-Orbitrap mass spectrometer. Metabolomic profiles were initially treated by principal component analysis (PCA) which was not efficient in distinguishing the samples to propose a diagnosis. Then, the metabolomic profile of a specific *m/z* range obtained in the negative mode of analysis was successfully subjected to orthogonal partial least squares discriminant analysis (OPLS-DA) which separated the two groups with and without CMPA. The model fit was $R^2 = 0.88$, with a predictive capability of $Q^2 = 0.52$ in the test with 2000 permutations and significance p-value <0.05. In this preliminary study, the model obtained using the metabolomic profile showed significant validation values, indicating the potential to distinguish between the two groups of interest, suggesting its use as a possible diagnostic tool for CMPA patients.

Keywords: allergy, cow's milk, metabolomics, microbiota

Cite: Silva, K. S. P.; Gois, A. R. S.; Costa, T. B. B. C.; Silva, W. R.; Wisniewski Jr, A.; Silva, R. O.; Vieira, S. C. F.; Motta-Franco, J.; Gurgel, R. Q. Metabolomic Screening of Fecal Samples as an Alternative Method to Initial Clinical Trials of Allergy to Cow's Milk Protein on Infants. *Braz. J. Anal. Chem.* (Forthcoming). http://dx.doi.org/10.30744/brjac.2179-3425.AR-187-2024

Submitted December 16, 2024; Resubmitted February 17, 2025; Accepted February 21, 2025; Available online March, 2025.