

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-graduação em Física

Victor Hugo Silva De Paiva

**APRENDIZADO DE MÁQUINA APLICADO
À ELIPSOMETRIA ESPECTROSCÓPICA PARA
O DIAGNÓSTICO SOROLÓGICO DE MALÁRIA**

Belo Horizonte
2025

Victor Hugo Silva De Paiva

**APRENDIZADO DE MÁQUINA APLICADO À ELIPSOMETRIA
ESPECTROSCÓPICA PARA O DIAGNÓSTICO SOROLÓGICO DE MALÁRIA**

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Ciências.

Orientador: Prof. Dr. Juan Carlos González Pérez

Coorientador: Dr. Paulo Henrique Ribeiro Amaral

Belo Horizonte
2025

Dados Internacionais de Catalogação na Publicação (CIP)

P149a Paiva, Victor Hugo Silva de.

Aprendizado de máquina aplicado à elipsometria espectroscópica para o diagnóstico sorológico de malária / Victor Hugo Silva de Paiva. – 2025.
151 f. : il.

Orientador: Juan Carlos González Pérez.

Coorientador: Paulo Henrique Ribeiro Amaral.

Dissertação (mestrado) – Universidade Federal de Minas Gerais,
Departamento de Física.

Bibliografia: f. 139-143.

1. Matriz de Mueller. 2. Análise espectral. 3. Aprendizado do computador. 4. Malária. I. Título. II. González Pérez, Juan Carlos. III. Amaral, Paulo Henrique Ribeiro. IV. Universidade Federal de Minas Gerais, Departamento de Física.

CDU – 535.33 (043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS

ATA

ATA DA SESSÃO DE ARGUIÇÃO DA 750ª DISSERTAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA, defendida por **VICTOR HUGO SILVA DE PAIVA** orientado pelo professor Prof. Juan Carlos González Pérez e coorientado pelo Dr. Paulo Henrique Ribeiro Amaral (Departamento de Física/UFMG) para obtenção do grau de **MESTRE EM FÍSICA, área de concentração Física aplicada e interdisciplinar**. Às 09:00 horas de dez de fevereiro de dois mil e vinte e cinco reuniu-se, por videoconferência, a Comissão Examinadora composta pelos professores **Juan Carlos González Pérez** (Orientador - Departamento de Física/UFMG), **Dr. Paulo Henrique Ribeiro Amaral** (Coorientador - Departamento de Física/UFMG), **Luis Eugenio Fernandez Outon** (Departamento de Física/UFMG) **Raphael Rocha Wenceslau** (Departamento de Clínica e Cirurgia Veterinárias /UFMG), e **Eduardo Nery Duarte de Araújo** (Departamento de Física/UFV), para dar cumprimento ao Artigo 37 do Regimento Geral da UFMG, submetendo o **VICTOR HUGO SILVA DE PAIVA** à arguição de seu trabalho de dissertação, que recebeu o título de "**Aprendizado de Máquina Aplicado à Elipsometria Espectroscópica para o Diagnóstico Sorológico de Malária**". O candidato fez uma exposição oral de seu trabalho durante aproximadamente 50 minutos. Após esta, os membros da comissão prosseguiram com a sua arguição, e apresentaram seus pareceres individuais sobre o trabalho, concluindo pela aprovação do candidato.

Belo Horizonte, 10 de fevereiro de 2025.

Prof. Juan Carlos González Pérez
Orientador do estudante
Departamento de Física/UFMG

Dr. Paulo Henrique Ribeiro Amaral
Coorientador do estudante
Departamento de Física/UFMG

Prof. Luis Eugenio Fernandez Outon
Departamento de Física/UFMG

Prof. Raphael Rocha Wenceslau
Departamento de Clínica e Cirurgia Veterinárias /UFMG

Prof. Eduardo Nery Duarte de Araújo
Departamento de Física - UFV

Candidato: Victor Hugo Silva de Paiva

Documento assinado eletronicamente por **Victor Hugo Silva de Paiva, Usuário Externo**, em 11/02/2025, às 14:32, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Raphael Rocha Wenceslau, Professor do Magistério Superior**, em 11/02/2025, às 17:47, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Paulo Henrique Ribeiro Amaral, Usuário Externo**, em 11/02/2025, às 21:53, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Eduardo Nery Duarte de Araújo, Usuário Externo**, em 13/02/2025, às 13:27, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Juan Carlos Gonzalez Perez, Professor do Magistério Superior**, em 13/02/2025, às 20:38, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3953062** e o código CRC **011B6D90**.

Agradecimentos

Agradeço ao professor Dr. Juan Carlos González Pérez e ao Dr. Paulo Henrique Ribeiro Amaral, que foram, respectivamente, meu orientador e coorientador neste trabalho. Sou grato a ambos pela enorme contribuição ao meu trajeto durante o mestrado, pelo suporte constante e pelo valioso conhecimento que me transmitiram. Agradeço também por todos os conselhos e pela disponibilidade que sempre me ofereceram.

Saúdo e agradeço meus pais Maria José e Danilo De Paiva, que me apoiam em cada etapa de minhas jornadas. Sou grato ao meu professor e amigo Flamarion, ao meu primo Leandro e à Iandry, que sempre contribuíram de diversas formas para a minha trajetória acadêmica e crescimento pessoal, tal qual minha madrinha Maria das Graças.

Um agradecimento especial às pesquisadoras Dra. Luzia Helena Carvalho e Dra. Flora Satiko Kano, do grupo de Biologia Molecular e Imunologia da Malária do Instituto René Rachou, por fornecerem as amostras utilizadas neste trabalho. Agradeço também à Dra. Maria Ivonete Nogueira da Silva e à Monique Costa pelo trabalho dedicado nas medições elipsométricas das amostras. Agradeço ao CNPq pelo financiamento da minha pesquisa e estudos na Universidade Federal de Minas Gerais e pela oportunidade de contribuir para a ciência brasileira. Sou profundamente grato à UFMG e a todos os colegas, professores e funcionários do Departamento de Física que, de alguma forma, contribuíram para que eu alcançasse meus objetivos.

Por fim, saúdo a obra de Alan Moore e Dave Gibbons, que, por um acaso da vida, me ajudou a fazer a escolha que me trouxe até aqui.

"O enigma do aço sempre foi envolvido por mistérios. É preciso aprender esse caminho, pois não se pode confiar em ninguém no mundo: homem, mulher ou criatura. Apenas o aço é confiável!"
(Conan: o bárbaro, criado por: Robert E. Howard)

Resumo

A malária é uma doença grave que afeta principalmente regiões tropicais e menos desenvolvidas do planeta. Pacientes infectados precisam ser diagnosticados e tratados rapidamente para prevenir a progressão da doença para quadros potencialmente fatais. Contudo, diversos entraves técnicos e econômicos dificultam o acesso e a adoção dos métodos diagnósticos atualmente disponíveis. Este estudo apresenta um novo método de diagnóstico da malária que é simples, eficaz, de baixo custo, rápido e que não requer o uso de biomarcadores. Para isso, foi desenvolvido um código baseado em inteligência artificial, capaz de diferenciar indivíduos saudáveis de reativos para malária por meio do processamento de dados de espectroscopia da Matriz de Mueller em amostras de soro sanguíneo. Os modelos desenvolvidos demonstraram desempenhos com valores de AUC acima de 0.9 na classificação entre indivíduos negativos e reativos para malária, além de fornecerem resultados fisicamente consistentes na região do infravermelho próximo. Além disso, os modelos permitem distinguir pacientes provenientes de áreas endêmicas e não endêmicas, oferecendo uma abordagem promissora para diagnósticos em diferentes contextos epidemiológicos.

Palavras-chave: Matriz de Mueller, espectroscopia, aprendizado de máquina, diagnóstico, malária.

Abstract

Malaria is a serious disease that primarily affects tropical and less developed regions of the world. Infected patients need to be diagnosed and treated quickly to prevent the progression of the disease to potentially fatal stages. However, various technical and economic barriers hinder access to and adoption of currently available diagnostic methods. This study presents a new malaria diagnostic method that is simple, effective, low-cost, rapid, and does not require the use of biomarkers. To achieve this, an artificial intelligence-based code was developed, capable of differentiating healthy individuals from those reactive to malaria by processing Mueller Matrix spectroscopy data from blood serum samples. The developed models demonstrated performance with AUC values above 0.9 in classifying negative and malaria-reactive individuals, in addition to providing physically consistent results in the near-infrared region. Furthermore, the models can distinguish patients from endemic and non-endemic areas, offering a promising approach for diagnostics in different epidemiological contexts.

Keywords: Mueller matrix, spectroscopy, machine learning, diagnosis, malaria.

Lista de ilustrações

Figura 1 – Mapa de áreas de transmissão de malária no mundo	20
Figura 2 – Mapa de incidência parasitária anual da malária no Brasil	24
Figura 3 – Propagação de uma onda unidimensional como função da posição e tempo	26
Figura 4 – Variação da luz com a modificação da fase	27
Figura 5 – Propagação de uma onda eletromagnética a partir do campo elétrico .	28
Figura 6 – Onda eletromagnética propagando em um meio transparente e em um meio com absorção de luz	31
Figura 7 – Tipos de polarizações dielétricas	32
Figura 8 – Dielétrico entre um capacitor de placas paralelas	33
Figura 9 – Propagação de radiação de um dipolo em um dielétrico	34
Figura 10 – Gráfico da parte real e imaginária da constante dielétrica complexa em função da frequência angular da luz	36
Figura 11 – Reflexão e transmissão da luz para uma incidência oblíqua entre dois meios	38
Figura 12 – Reflexão das luzes polarizadas-p e -s	39
Figura 13 – Esquema de incidência, reflexão e transmissão para a luz em polarização- p e polarização-s	40
Figura 14 – Representação da polarização linear, à direita e elíptica	42
Figura 15 – Mudança de polarização por meio da variação da diferença de fase . . .	43
Figura 16 – Estrutura da calcita a partir do eixo óptico e propagação de luzes em uma amostra de calcita	45
Figura 17 – Prisma de Glan-Taylor usado como polarizador/analizador e sua estrutura	45
Figura 18 – Ação de um compensador	46
Figura 19 – Estrutura de um modulador fotoelástico	47
Figura 20 – Estrutura de um despolarizador	48
Figura 21 – Representação de um instrumento óptico usando matrizes de Jones . .	51
Figura 22 – Intensidade da luz na lei de Malus	52
Figura 23 – Representação gráfica dos parâmetros de Stokes pela intensidade da luz	53
Figura 24 – Interpretação dos estados de polarização por meio esfera de Poincaré .	55
Figura 25 – Representação da polarização elíptica em dois sistemas de coordenadas	56
Figura 26 – Representação de um ponto P na esfera de Poincaré	57
Figura 27 – Representações na esfera de Poincaré para diferentes polarizações . . .	58
Figura 28 – Seções da Matriz de Mueller	60
Figura 29 – Tipos de elipsômetros e seus elementos	64
Figura 30 – Modelo óptico constituído de ar, filme fino e substrato	65
Figura 31 – Esquema e princípios de medida para elipsometria	67

Figura 32 – Convergência dos erros empírico e esperado para dois espaços de busca diferentes	76
Figura 33 – Gráfico da curva de erros de treino (empírico) e teste (esperado) em função da capacidade do modelo.	78
Figura 34 – Representação da variância e viés no espaço de hipóteses	79
Figura 35 – Curvas dos erros por variância e viés em função da capacidade de um modelo.	79
Figura 36 – Gráficos para três tipos de ajustes de uma saída y em função de uma <i>feature</i> x	80
Figura 37 – Demonstração de uma curva ROC para uma $AUC = 0.83$	84
Figura 38 – Curva ROC de um classificador perfeito (AUC máxima).	85
Figura 39 – Esquema de otimização de um modelo com vinte tentativas com base na acurácia dos dados de teste.	86
Figura 40 – Divisão de uma base de dados em cinco <i>folds</i> de modo estratificado para duas classes.	88
Figura 41 – Exemplo de uma validação cruzada de cinco partições avaliada por acurácia.	88
Figura 42 – Comparativo de validações cruzadas para uma rede neural	89
Figura 43 – Esquema de funcionamento da <i>nested cross validation</i>	92
Figura 44 – Três diferentes situações para separar um conjunto de dados em duas dimensões com SVM	94
Figura 45 – Estimativa de como cada modelo SVM poderia se comportar em um grupo de dados de teste	95
Figura 46 – Exemplo da aplicação de um kernel	97
Figura 47 – Exemplo de ajuste a dados pelo algoritmo de regressão logística	99
Figura 48 – Diferentes tipos de redes neurais	100
Figura 49 – Esquematização de uma rede neural com <i>backpropagation</i>	105
Figura 50 – Separação de duas classes usando árvore de decisão	106
Figura 51 – Estrutura de uma árvore de decisão com folhas, galhos e nós.	107
Figura 52 – Duas árvores de decisão construídas a partir de dois atributos diferentes para o nó raiz	109
Figura 53 – Esquema simplificado do funcionamento de uma random forest para três árvores.	110
Figura 54 – Estrutura básica de funcionamento de um XGBoost	112
Figura 55 – Esquema do funcionamento do KNN em duas dimensões para duas classes	114
Figura 56 – Esquema das quatro etapas para avaliação de um novo diagnóstico clínico.	116
Figura 57 – Preparação de amostras de soro e medições de elipsometria espectroscópica de ângulo variável da matriz Mueller.	120
Figura 58 – Metodologia geral para a criação e avaliação de modelos	122

Figura 59 – Metodologia da validação cruzada estratificada para 10- <i>folds</i>	124
Figura 60 – Metodologia da <i>Nested-CV</i> utilizada para avaliar os modelos	126
Figura 61 – Representação de uma Matriz de Mueller Teórica que é isotrópica e despolarizante.	128
Figura 62 – Espectros da matriz de Mueller experimental para uma amostra de soro da base de dados para seis ângulos de incidência diferentes.	128
Figura 63 – Espectro médio de cada elemento da matriz de Mueller de amostras negativas para malária em um ângulo de incidência de 70°.	129
Figura 64 – Espectro médio de cada elemento da matriz de Mueller de amostras reativas para malária em um ângulo de incidência de 70°.	130
Figura 65 – Curvas ROC dos modelos da tabela 2.	131

Lista de tabelas

Tabela 1 – Grupos de amostras de malária	119
Tabela 2 – Resultados das métricas para os melhores modelos de cada grupo amostral usando a SKFCV	131
Tabela 3 – Resultados das métricas para os melhores modelos de cada grupo amostral usando a <i>Nested-CV</i>	134
Tabela 4 – Quantidade de amostras de COVID-19 e dengue	135
Tabela 5 – Especificidade dos modelos de malária contra COVID-19 e dengue . . .	135
Tabela 6 – Parâmetros e espaço de busca utilizados durante a otimização do modelo.	145
Tabela 7 – Hiperparâmetros e espaço de busca utilizados durante a otimização do modelo.	145
Tabela 8 – Resultados dos modelos não endêmicos	147
Tabela 9 – Resultados dos modelos reativos	148
Tabela 10 – Resultados dos modelos endêmicos	149
Tabela 11 – Resultados dos modelos negativos	150
Tabela 12 – Resultados dos modelos com todas as amostras	151

Sumário

	Notações	16
1	INTRODUÇÃO	17
2	MALÁRIA	20
2.1	Imunologia e tratamentos	21
2.2	Malária no Brasil	23
3	PRINCÍPIOS DE ÓPTICA	25
3.1	Propagação da luz	25
3.2	Tratamento ondulatório e corpuscular da luz	28
3.3	Índice de refração	29
3.4	Dielétricos	31
3.5	Reflexão e transmissão da luz	37
3.5.1	Refração da luz	37
3.5.2	Luz polarizada em -p e -s	38
4	FORMALISMO DE JONES E MUELLER	41
4.1	Polarização da luz	41
4.2	Elementos ópticos	43
4.2.1	Polarizador e analisador	44
4.2.2	Compensador e retardador	46
4.2.3	Modulador fotoelástico	46
4.2.4	Depolarizador	47
4.3	Cálculo de Jones	48
4.3.1	Vetores de Jones	48
4.3.2	Representação em matrizes de Jones dos elementos ópticos	50
4.3.3	Representação de medidas ópticas por matrizes de Jones	50
4.4	Cálculo de Mueller	52
4.4.1	Parâmetros de Stokes	52
4.5	Matriz de Mueller	58
4.5.1	Seções da matriz de Mueller	60
4.5.2	Medições na matriz de Mueller	62
5	ELIPSOMETRIA ESPECTROSCÓPICA	65
5.1	Medições de elipsometria	66
5.2	Considerações gerais sobre elipsometria	69

6	INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA	70
6.1	Princípios de aprendizagem de máquina	71
6.1.1	Treino e teste do modelo: o aprendizado é possível?	74
6.1.2	Capacidade dos modelo	76
6.1.3	Métricas de desempenho	81
6.1.4	Otimização de um modelo	85
6.2	Métodos de validação	87
6.2.1	Validação cruzada estratificada em k-partições	87
6.2.2	Validação cruzada <i>Nested</i>	90
6.3	Algoritmos	92
6.3.1	Support Vector Machine (SVM)	93
6.3.2	Regressão Logística (LR)	97
6.3.3	Perceptron e Rede Neural (<i>Neural Network</i>)	99
6.3.4	Árvore de decisão e Extreme Gradient Boosting	105
6.3.5	<i>K-Nearest Neighbors</i> (KNN)	112
7	DIAGNÓSTICOS CLÍNICOS	115
7.1	Análise e validação de testes clínicos	115
7.2	Testes clínicos por aprendizado de máquina	117
8	METODOLOGIA	118
8.1	Amostras de soro sanguíneo	118
8.2	Medidas espectrais de soro sanguíneo	119
8.3	Construção do Modelo de aprendizado de máquina	121
9	RESULTADOS E DISCUSSÃO	127
9.1	Análise espectral do soro sanguíneo	127
9.2	Modelos de diagnóstico por aprendizado de máquina	130
9.3	Especificidade dos modelos contra outras doenças	134
10	CONCLUSÃO	137
	REFERÊNCIAS	139
	APÊNDICES	144
	APÊNDICE A – TABELAS DE ESPAÇOS DE BUSCA PARA PARÂMETROS E HIPERPARÂMETROS	145

APÊNDICE B – TABELAS DE RESULTADOS DE TODOS MO- DELOS DOS GRUPOS AMOSTRAIS	146
---	-----

Notações

Aqui apresentamos um conjunto de notações úteis (esquerda) e suas definições (direita) usadas ao longo do texto. As notações que forem separadas por vírgulas são aquelas que podem ser entendidas como equivalentes, pois há momentos em que é necessário usar uma notação alternativa para evitar ambiguidades.

n Número escalar real

$\vec{V}, \vec{v}$ Vetor

$\vec{V}_i, \vec{V}^{(i)}$ Vetor de instância i

$\hat{v}$ Vetor unitário

$\tilde{M}$ Matriz

$\overleftrightarrow{T}$ Tensor

$v_{mi}, v_m^{(i)}$ Componente m de um vetor de instância i

$\tilde{M}_{ij}$ Elemento da linha i e coluna j da matriz $\tilde{M}$

$\tilde{M}_{:j}$ Coluna j da matriz $\tilde{M}$

$\overleftrightarrow{T}_{::k}$ Partição k do tensor $\overleftrightarrow{T}$

$\overleftrightarrow{T}_{ijk}$ Elemento da linha i , coluna j e partição k do tensor $\overleftrightarrow{T}$

$\{\phi_i\}$ Conjunto ϕ com i elementos

$\vec{V} \cdot \vec{v}, \vec{V} \vec{v}$ Produto escalar entre dois vetores

$\tilde{M} \cdot \vec{v}, \tilde{M} \vec{v}$ Operação de uma matriz sobre um vetor

$\dot{f}(x)$ Primeira derivada de uma função

1 Introdução

Ao longo da cultura e história da humanidade, os inventores sempre idealizaram criar máquinas e seres artificiais com capacidade de pensamento, e esse desejo acompanha a humanidade pelo menos desde a época da Grécia antiga [1]. As figuras míticas de Pigmalião, Dédalo e Hefesto podem ser interpretadas como esses inventores lendários, e Galateia, Talos e Pandora podem ser consideradas como vida artificial [2–4]. Essa ambição pelo artificial perpassou por décadas até que se manifestou novamente em um encontro entre a literatura de horror de Mary Shelley com a eletricidade de Galvani e Volta na obra *Frankenstein: O Prometeu moderno*. Essa icônica obra aborda o sucesso do Dr. Victor Frankenstein na criação de vida artificial para reanimar um cadáver a partir do princípio de eletricidade animal [5, 6]. Embora o Dr. Frankenstein e sua criatura sejam ficção, a relação entre a Física e as criações artificiais da humanidade são uma realidade no contexto de inteligência artificial.

O aprendizado de máquina e as inteligências artificiais têm apresentado desempenhos de sucesso na resolução de problemas de Física [7]. Essas ferramentas são capazes de modelar processos físicos do mundo real e possuem uma variedade de construções para resolver problemas desafiadores de maneiras não convencionais [7]. Porém, a resolução de problemas por aprendizado de máquina demanda uma grande quantidade de dados, pois os algoritmos precisam aprender e detectar padrões para realizar tarefas e solucionar problemas. Todo esse processo tende a ser mais eficiente quando se tem disponível uma grande quantidade de dados, então a resolução de problemas físicos usando inteligência artificial demanda de técnicas que sejam capazes de fornecer grandes volumes de informações para alimentar os modelos de aprendizado de máquina.

A elipsometria espectroscópica é uma técnica óptica sem contato, não destrutiva e amplamente utilizada para medir as constantes ópticas de materiais [7], e consiste na incidência de radiação eletromagnética (aqui denominada como luz) polarizada em uma amostra e na análise espectral da mudança de polarização da parte da luz transmitida ou especularmente refletida. Devido à natureza espectral dessa técnica, é possível coletar o espectro de uma amostra para uma variedade de comprimentos de onda e em vários ângulos de incidência diferentes. Dessa forma, essa técnica é capaz de coletar grandes quantidades de dados, e isso faz com que ela seja uma fonte de dados promissora para alimentar um algoritmo de aprendizado de máquina. Além disso, a elipsometria é capaz de caracterizar opticamente vários tipos de amostras, incluindo aquelas de origem biológica. Essa descrição óptica pode ser feita para amostras biológicas de indivíduos que estejam contaminados por alguma doença. Com isso, podemos usar os espectros de amostras infectadas para montar um modelo de inteligência artificial que seja capaz de identificar

amostra contaminadas por alguma doença específica.

A malária é uma doença potencialmente fatal, que demanda rapidamente de diagnóstico e tratamento para minimizar suas consequências. Apesar de o diagnóstico microscópico ser o padrão-ouro, a detecção e combate de uma quantidade significativa de infecções submicroscópicas é crucial para progredir no controle e eliminação da doença. Os atrasos no diagnóstico podem ocorrer em regiões não endêmicas por não ser uma enfermidade clinicamente suspeita para aquela área. Portanto, há uma demanda urgente para a elaboração de novos métodos de diagnóstico de malária que sejam rápidos, eficazes e de baixo custo. Logo, a malária é uma enfermidade que apresenta um contexto adequado para ser usada em um processo de criação de um modelo de diagnóstico baseado na aprendizagem de máquina, e há estudos e revisões sobre como o uso da espectroscopia e aprendizado de máquina são ferramentas promissoras para monitorar e diagnosticar a malária [8].

O objetivo deste trabalho é construir e avaliar modelos de aprendizado de máquina capazes de classificar, eficientemente, pacientes com e sem anticorpos específicos para malária a partir de dados de espectroscopia da Matriz de Mueller de amostras de soro sanguíneo e sem a utilização de biomarcadores. Usando técnicas de regularização e de tratamento de dados, os modelos de inteligência artificial elaborados devem ser capazes de classificar os pacientes como reativos (os que tem anticorpos específicos contra infecção por malária) ou negativos para a malária, levando sempre em mente que todo o processo deve ser feito com baixo custo, de maneira rápida e sem biomarcadores. Objetiva-se também criar modelos que permitam classificar se as amostras são de região endêmica ou não endêmica para incidência de malária.

Por se tratar de um trabalho de física aplicada, ele foi estruturado de forma que há uma apresentação teórica para cada tópico crucial da física aplicada ao aprendizado de máquina para diagnósticos médicos. Os resultados e discussões dessa aplicação são colocados logo após a exposição dos capítulos teóricos. Há ainda informações e conteúdos que estão elucidados nos apêndices deste trabalho para que não ocorra uma quebra muito brusca na leitura do texto ao enunciar informações extensas.

Uma abordagem geral da malária está presente no capítulo dois deste trabalho. Nele, encontraremos uma contextualização da doença no Brasil a partir de uma análise das áreas de ocorrência da malária e quais as consequências sociais e econômicas para país. Os aspectos ligados às características imunológicas, diagnóstico e tratamento da doença são discutidos para compreender como a infecção opera e quais os procedimentos vigentes para combatê-la. Dentro disso, dados sobre índices de infecções e mortalidade serão utilizados como elementos motivadores na proposta de um método alternativo de diagnóstico para essa protozoose.

Os capítulos três, quatro e cinco deste trabalho abordam a teoria que envolve a

técnica de elipsometria espectroscópica. Para tal exposição, serão apresentados os princípios e fundamentos da óptica tal que o objetivo é perscrutar conceitos dessa área para expor e demonstrar com clareza o cálculo de Jones, parâmetros de Stokes e as Matrizes e cálculo de Mueller. Embora muitos dos conceitos anteriores a esses tópicos alvo sejam amplamente conhecidos, eles serão abordados de forma direcionada ao entendimento do cálculo de Jones e de Mueller e a relação deles com a elipsometria espectroscópica. Este capítulo é fundamental para compreender a natureza dos fenômenos, processos e medidas experimentais que serão abordados e usados posteriormente na construção dos modelos de inteligência artificial.

O arcabouço necessário referente ao aprendizado de máquina e inteligência artificial é apresentado no capítulo seis. Nessa seção será apresentada uma contextualização sobre o aprendizado de máquina, exposição de algoritmos canônicos, técnicas e métricas de validação eficientes na análise de modelos. Cada um desses assuntos abordados por si só já podem render uma quantidade vasta de conteúdo a ser explorada, então a escrita deste capítulo é focada em uma apresentação geral, intuitiva e por vezes algébrica dos conceitos e tópicos apresentados.

O capítulo sete foi reservado para discorrer sobre a teoria dos diagnósticos clínicos e temas relacionados com a produção de novos modelos de diagnósticos. Nesse ponto, é crucial a abordagem sobre os vieses e problemas a serem contornados ao propor novos processos de diagnosticar doenças. As etapas, procedimentos e comportamentos esperados para a implementação de um novo modelo de diagnóstico clínico serão, ao fim do capítulo, relacionados às estruturas e processos que envolvem o aprendizado de máquina em aplicações de diagnósticos.

Nos capítulos oito e nove têm-se, respectivamente, a exposição da metodologia para a criação dos modelos de aprendizado de máquina e os resultados obtidos a partir deles. Os detalhes das medidas de elipsometria, das amostras utilizadas, espaços de busca e dos algoritmos estão apresentados no capítulo 8 e no apêndice A. Os principais resultados experimentais e discussões podem ser encontrados no capítulo 9, e os valores das métricas para todos modelos construídos estão presentes nas tabelas do apêndice B.

2 Malária

Entre as protozooses mais conhecidas, a Malária é um destaque, encontrando-se atualmente presente em praticamente todas as regiões tropicais e subtropicais do globo [9] conforme é ilustrado na Figura 1, e estima-se que pode afetar entre 300 e 500 milhões de pessoas globalmente nessas regiões [10]. Apesar dos avanços alcançados pela Organização Mundial de Saúde no combate à malária desde o final da década de 70, esta enfermidade permanece como um dos problemas mais graves de saúde pública em escala global, podendo ocasionar cerca de 2.5 milhões de óbitos anualmente, a maioria entre crianças [9]. O impacto da malária vai além das questões de saúde pública, pois afeta também o desenvolvimento econômico e social das comunidades afetadas [10].

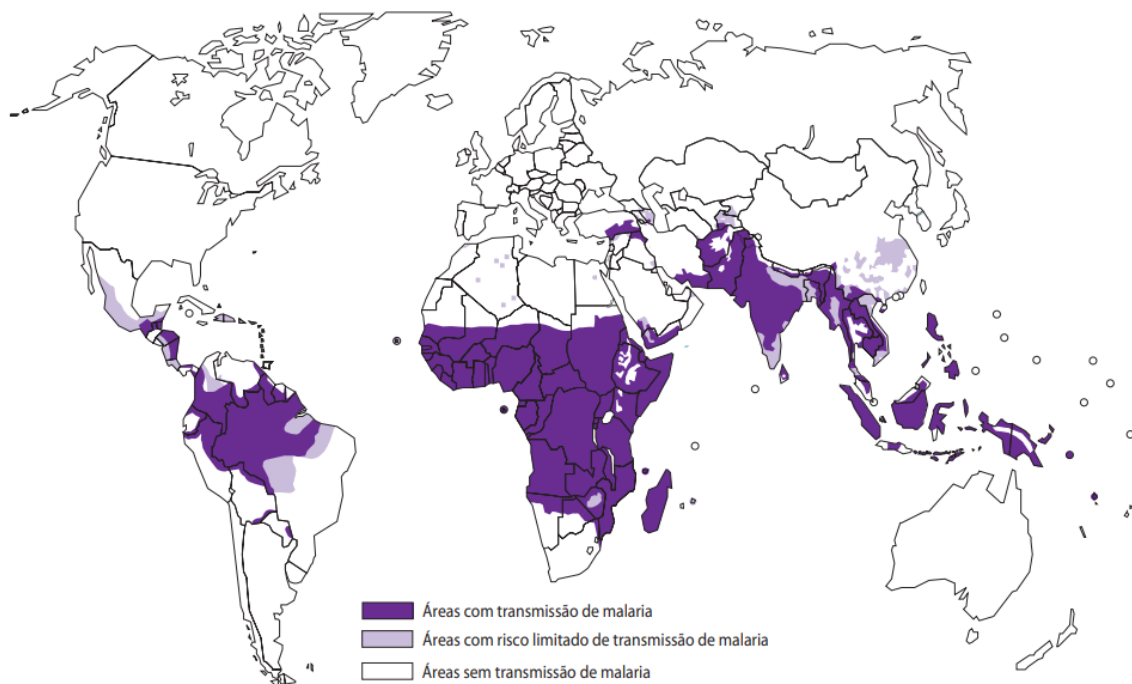


Figura 1 – Mapa de áreas de transmissão de malária no mundo elaborado pela OMS em 2007. Adaptado de [11].

A malária é provocada por protozoários pertencentes ao gênero *Plasmodium*, transmitidos ao ser humano por fêmeas de mosquitos do gênero *Anopheles* que estão infectadas, sendo apenas quatro das aproximadamente 100 espécies desses protozoários responsáveis por infectar seres humanos: *P. falciparum*, *P. vivax*, *P. ovale* e *P. malariae* [9]. O *P. falciparum* é a espécie mais virulenta, responsável pelas formas graves da doença, enquanto o *P. vivax* é mais prevalente no Brasil [10]. O período de incubação da malária varia

conforme a espécie de *Plasmodium*: de 8 a 12 dias para *P. falciparum*, 13 a 17 dias para *P. vivax* e 28 a 30 dias para *P. malariae* [10]. Esse período de incubação pode ser mais longo em indivíduos considerando alguns fatores: exposição prévia à infecção, uso profilático de antimaláricos ou que trataram parcialmente a infecção [10]. Uma característica importante da doença é que, clinicamente, a malária pode ser confundida com infecções virais comuns, atrasando o diagnóstico [10].

A malária é uma infecção de características sistêmicas capaz de provocar a ruptura de hemácias e alterações em diversos órgãos, tal que as manifestações clínicas e patológicas surgem exclusivamente do ciclo eritrocítico assexuado, que destrói os eritrócitos e libera parasitas e metabólitos na circulação, desencadeando respostas inflamatórias e alterações morfológicas e funcionais no hospedeiro [10]. Essa infecção pelo parasita da malária provoca modificações significativas nas hemácias (glóbulos vermelhos), que são fundamentais para a patogênese da doença. Essas alterações incluem mudanças na estrutura e função das hemácias, tornando-as mais rígidas e aderentes [12]. A doença caracteriza-se por proporcionar febres intermitentes, que, dependendo da espécie de plasmódio, ocorrem a cada 2 ou 3 dias, além de dores de cabeça, dores no corpo, anemia, icterícia e inchaço do fígado e baço, e, no caso da malária cerebral, o parasita pode comprometer progressivamente o sistema nervoso central [9]. Durante a infecção, os episódios de calafrios e febre alta podem durar de 3 a 4 horas, acompanhados de mal-estar, náuseas, cefaléias e dores nas articulações, e, após alguns dias, o quadro volta a se repetir por semanas até que o indivíduo, não tratado, cure espontaneamente ou chegue a óbito pelas complicações renais, pulmonares e cerebrais mencionadas anteriormente [10]. Então, clinicamente, a malária pode variar de uma doença febril branda até formas graves e fatais com casos severos relacionados a coma cerebral e falência múltipla de órgãos [10], e o tratamento adequado para a doença é necessário.

2.1 Imunologia e tratamentos

A resposta imunológica à malária de indivíduos infectados envolve mecanismos inatos e adquiridos, e aqueles indivíduos que são expostos repetidamente à doença podem desenvolver uma imunidade parcial capaz de reduzir a gravidade dos sintomas sem eliminar completamente o parasita. Porém, a diversidade e capacidade de variação antigênica do *Plasmodium*, especialmente do *P. falciparum*, permite que o parasita escape da resposta imunológica, que pode dificultar o desenvolvimento de vacinas eficazes [10].

A microscopia se mantém como o padrão-ouro para detecção de malária nos últimos 100 anos [13, 14], mas há outros meios para o diagnóstico da malária, e vamos avaliar dois dos principais métodos sorológicos: Imunofluorescência Indireta (IFI) e Teste Imunoenzimático ELISA, em que ambos podem ser usados para indivíduos primo-infectados

(infectados por malária pela primeira vez) ou em casos mistos (infecções causadas por mais de um parasita) [15]. O IFI é um método de referência para detecção dos anticorpos IgG e IgM contra *Plasmodium falciparum* e *Plasmodium vivax*, enquanto o ELISA é uma alternativa ao IFI para detecção de anticorpos em situações de estudos epidemiológicos e diagnósticos individuais [15]. Ambos os testes podem apresentar alta sensibilidade e especificidade, mas com diferenças em relação ao tipo de infecção. Para indivíduos primo-infectados pelo *P. falciparum*, têm-se pesquisas que reportam uma sensibilidade de 71% para ambos os testes, mas para infecções mistas ou pacientes com infecção ativa, os índices alcançaram 100% [15]. Para a especificidade o IFI alcançou um resultado de 100%, e o ELISA atingiu 95% com relação aos indivíduos não contaminados por malária [15]. Com o tempo, as técnicas e modos de diagnósticos tendem a se tornarem mais eficientes, mas esses resultados apresentados mostram que pelo menos dois dos métodos de diagnóstico de malária já possuem bons resultados na tarefa de diferenciar indivíduos infectados dos não infectados a partir do estudo realizado [15], mas com uma sensibilidade baixa para indivíduos primo-infectados.

Após o diagnóstico, ações para tratar o indivíduo infectado e combater a malária precisam ser tomadas. O tratamento da malária causada por *Plasmodium vivax*, *P. ovale* e *P. malariae*, conhecidos como agentes de malárias benignas, deve incluir o uso de cloroquina, que é eficaz contra as formas sanguíneas e os gametócitos dessas espécies [10]. Porém, a cloroquina não apresenta efeito sobre as formas hepáticas (hipnozoítas) do parasita, e por isso é necessário associá-la à primaquina, que é um esquizonticida tecidual direcionado aos hipnozoítas, para garantir a eliminação completa do ciclo parasitário e com baixa toxicidade para os indivíduos em tratamento [10].

Para infecções causadas por *Plasmodium falciparum*, recomenda-se o uso de uma combinação de quinina e tetraciclina (doxiciclina) e, alternativamente, em casos de intolerância ou resposta inadequada ao tratamento, a mefloquina, que é um esquizonticida sanguíneo de dose única, pode ser administrada [10]. Apesar ter uma eficácia comprovada, o combate à malária é dificultado pela disseminação global de cepas resistentes de *P. falciparum* a medicamentos como cloroquina e sulfadoxina-pirimetamina, e essa resistência, que é concentrada especialmente na América do Sul e no sudeste da Ásia, limita a eficácia da cloroquina em regiões endêmicas para *P. falciparum*, onde a mefloquina é a profilaxia mais indicada [10].

Uma forma promissora de tratamento é o uso de derivados da substância artemisinina extraída da planta *Artemisia annua*, que proporciona alívio dos sintomas em menos tempo que cloroquina, mefloquina ou quinina [10]. Nos casos graves de infecções por *P. falciparum*, os derivados de artemisinina são uma boa primeira escolha para efetuar o tratamento, seguidos pela administração de outro antimalárico, como a mefloquina, após o paciente atingir um estado clínico estável [10].

A busca por vacinas eficazes contra a malária concentra-se em diversas formas evolutivas do parasita, incluindo esporozoítos, formas hepáticas, eritrocíticas assexuadas e gametócitos [10]. No entanto, as vacinas antiesporozoíticas demonstraram uma imunidade limitada, enquanto aquelas voltadas para as formas eritrocíticas ainda carecem de estudos que comprovem sua eficácia a longo prazo. Além disso, vacinas antigametócitos estão sendo investigadas com o objetivo de dificultar a evolução do parasita no mosquito vetor, embora essa estratégia não proporcione um benefício direto aos indivíduos vacinados [10].

2.2 Malária no Brasil

No Brasil, 99.9% da transmissão da malária está concentrada na região da Amazônia, e a incidência dessa doença no país, no ano de 2007, é apresentada na Figura 2 [11]. Ao analisar o histórico da malária, observamos uma diminuição na frequência de casos de 2010 a 2016 segundo os dados do Ministério da Saúde, mas em 2017 ocorreu um aumento significativo de 52.7% nos casos autóctones (no local de diagnóstico) em comparação com o ano anterior, totalizando 189515 casos [16]. Já em 2021, foram registrados 139112 casos autóctones, significando uma redução de quase 3% em relação ao ano de 2020 [16].

Quanto às mortes por malária, houve um aumento significativo no número de casos no Brasil em 2018. Nesse ano, o número de óbitos aumentou em 64.7%, passando de 34 em 2017 para 56 em 2018. No ano seguinte, em 2019, houve uma redução de 33.9% com 37 óbitos registrados [16]. Em 2020, houve um novo aumento, totalizando 51 óbitos, representando um incremento de 37.8% em comparação ao ano anterior [16]. O número de óbitos seguiu aumentando, pois em 2021, de acordo com dados preliminares, foram registrados 58 óbitos por malária no país, o que representa um aumento de 13.7% em comparação com 2020 [16]. A letalidade por malária na região amazônica é baixa, situando-se em 0.04% enquanto nas demais regiões do país, a letalidade chega a ser maior, posto que o falecimento nessas áreas ocorre principalmente para pessoas que chegam infectadas de outros estados da região amazônica ou de outros países. Nesse caso, essas vítimas da doença não recebem diagnóstico e tratamento oportuno muitas vezes devido à dificuldade na suspeição de uma doença relativamente rara nessas áreas não endêmicas de malária, e também à falta de informação por parte dos viajantes sobre os riscos associados à doença [16]. Isso demonstra como o diagnóstico precoce se faz necessário e que as áreas não endêmicas de malária necessitam de métodos eficazes e acessíveis para detectar infectados por malária.

Devido aos males na saúde e economia que a malária pode causar, é preciso tomar ações que sejam eficientes para mitigar os seus efeitos. As estratégias de controle da malária incluem o diagnóstico precoce, tratamento adequado, uso de inseticidas e medidas de proteção individual, como mosquiteiros impregnados e inseticidas [10]. Apesar disso, os

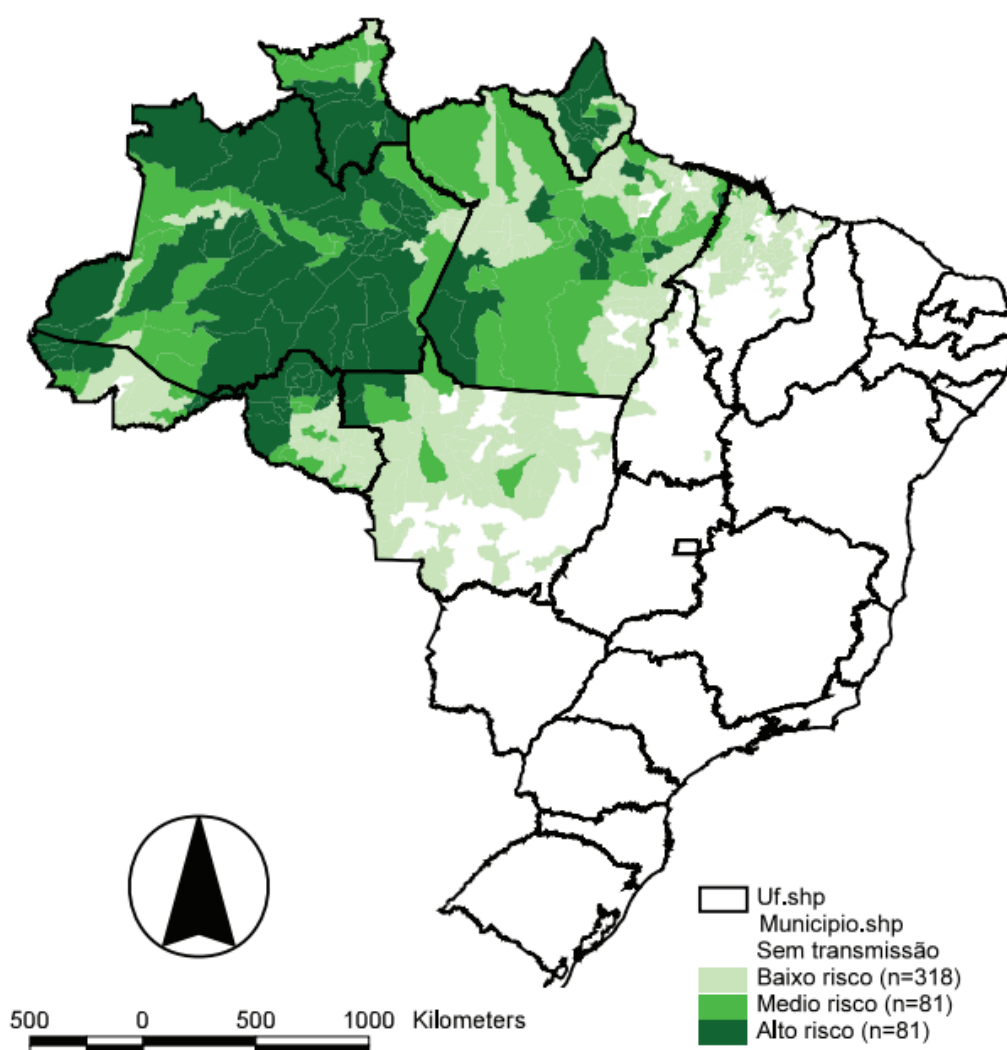


Figura 2 – Mapa de incidência parasitária anual (IPA) da malária no Brasil. Adaptado de [11].

esforços nesse controle enfrentam problemas como: a falta de infraestrutura da saúde em áreas endêmicas, o alto índice de portadores da doença que são assintomáticos e o impacto de fatores ecológicos e sociais [10]. A educação sanitária e o envolvimento da comunidade são fundamentais para o sucesso das estratégias de controle da doença [10].

3 Princípios de óptica

Como este trabalho tem foco na aplicação da física ao diagnóstico da malária, é necessário apresentar ferramentas físicas que permitam alcançar esse objetivo. Como visto anteriormente, infecções por malária provocam alterações físicas e químicas no soro sanguíneo [17]. Dessa forma, é de grande interesse buscar uma medida física capaz de detectar essas alterações. Para isso, direcionaremos a atenção para a constante dielétrica e a função dielétrica, que são grandezas sensíveis às características e componentes da matéria. Portanto, elas são parâmetros físicos adequados para detectar as modificações que podem ocorrer no soro sanguíneo em decorrência da infecção por malária. Como a constante dielétrica está relacionada à interação da luz com a matéria, torna-se necessário abordar alguns fenômenos e propriedades da luz para analisarmos a função dielétrica de maneira apropriada.

3.1 Propagação da luz

A natureza ondulatória da luz demanda uma teoria que seja capaz de tratar ondas eletromagnéticas de forma que considere todos os fenômenos associados a uma onda, como características de reflexão, refração, transmissão, frequência, amplitude. Essas propriedades da luz podem estar associadas a grandezas fisicamente mensuráveis capazes de caracterizar problemas de ordem complexa desde que ocorra um tratamento adequado. A luz manifesta dois comportamentos típicos: o corpuscular, proposto por Einstein em 1905, e o ondulatório. A partir do tratamento da luz como uma onda, foi mostrado por Maxwell que ela também tem propriedades eletromagnéticas de forma então que classificamos a luz como uma onda eletromagnética propagante [18]. Esse comportamento ondulatório da luz é possível ser descrito por meio de funções senoidais de tal modo que a caracterização da onda será feita a partir de parâmetros espaciais e temporais. Uma maneira direta de ilustrar isso é tomando uma onda de amplitude A que propaga com velocidade constante s em apenas uma dimensão para um dado eixo x e com um comprimento de onda λ . A partir disso, essa onda pode ser caracterizada matematicamente como [18]

$$\phi = A \sin(K(x - st)). \quad (3.1)$$

Para essa expressão, o termo K , que é chamado de número de onda, indica a quantidade de ondas senoidais presentes em um intervalo de 0 até 2π . O número de onda é definido por

$$K = \frac{2\pi}{\lambda}. \quad (3.2)$$

O período τ dessa onda é encontrado a partir da velocidade de propagação

$$\tau = \frac{\lambda}{s}, \quad (3.3)$$

e com isso a frequência v e frequência angular ω podem ser definidas como [18]

$$v = \frac{1}{\tau}, \quad (3.4)$$

$$\omega = \frac{2\pi}{\tau} = 2\pi v. \quad (3.5)$$

As equações (3.2), (3.3) e (3.5) podem ser usadas em (3.1) para reescrever a equação de onda em uma dimensão para uma forma mais usual [18]:

$$\phi = A \sin(Kx - \omega t), \quad (3.6)$$

em que é comum chamar o termo $(Kx - \omega t)$ de fase da onda, pois partindo dele os parâmetros espaciais e temporais são tomados para determinar o valor da onda. É possível compilar essas informações para compreender melhor a propagação de uma onda em relação a algumas variáveis essenciais. A figura 3 mostra uma onda ϕ propagando com velocidade s ao longo do eixo x , com um período τ e comprimento de onda λ . Os dois eixos além de x representam o tempo t e o valor da onda ϕ em função dos parâmetros espaciais e temporais. Combinando 3.2-3.5, encontramos a expressão para a velocidade de propagação s como [18]

$$s = \frac{\lambda}{\tau} = \lambda v = \frac{\omega}{K}. \quad (3.7)$$

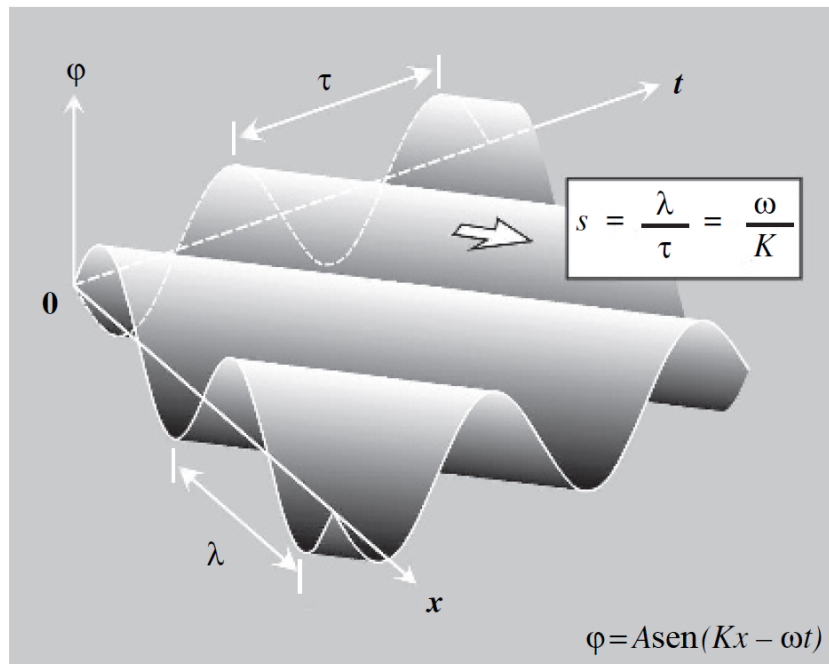


Figura 3 – Propagação de uma onda unidimensional como função da posição x e tempo t . Para a onda, temos a velocidade de propagação s , período τ e comprimento de onda λ . Adaptado de [18].

Partindo de suas características ondulatórias, a luz é capaz de sofrer o fenômeno de reflexão quando atinge um meio, que faz com que a onda possa sofrer alterações em sua fase. A fim de mensurar melhor essa mudança e expressar esse fenômeno, é salutar introduzir uma variável δ que represente a fase inicial dessa luz incidente no meio. Para tal, podemos modificar a equação de onda (3.6) introduzindo a fase inicial:

$$\phi = A \sin(Kx - \omega t + \delta), \quad (3.8)$$

Essa expressão pode ser reescrita por meio de funções exponenciais e número complexos. Há uma forte correlação entre onda, constantes ópticas e números complexos, e a partir da forma exponencial de Euler podemos propor a reformulação da equação da onda (3.8) para [18]:

$$\phi = A e^{i(\omega t - Kx + \delta)}. \quad (3.9)$$

Partindo dessa expressão, é possível coletar o valor absoluto da amplitude de forma prática, tal que

$$|\phi| = |A e^{i(\omega t - Kx + \delta)}| = |A|. \quad (3.10)$$

Comparando (3.8) e (3.9), temos uma mudança no sinal da fase de $(Kx - \omega t)$ para $(\omega t - Kx)$. Porém, essa mudança não afeta a propagação da onda, pois $\sin(-x) = \cos(x + \frac{\pi}{2})$, embora a fase inicial δ precise ser ajustada para preservar corretamente a propagação da onda [18].

Exemplificando como a fase altera a propagação de uma onda, temos a figura 4, que ilustra uma onda da forma $E = E_0 \cos(\omega t - Kx + \delta)$ em um tempo inicial $t = 0$. Nessa situação, a onda tem uma fase inicial $\delta = 0$ e é modificada para $\delta = \frac{\pi}{2}$. Se considerarmos que ocorreu uma mudança de fase no valor de δ , a distância entre os picos das duas ondas da figura 4 será de $\frac{\delta}{K}$. Portanto, se essas duas ondas forem postas para propagarem na mesma direção, a diferença de fase afeta os valores da onda resultante, que é o resultado esperado.

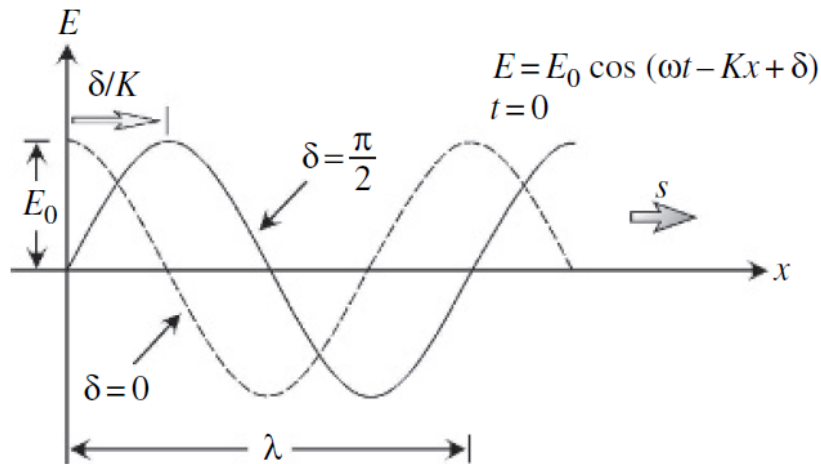


Figura 4 – Variação da luz com a modificação da fase δ . Adaptado de [18].

3.2 Tratamento ondulatório e corpuscular da luz

Conforme o estudo da eletricidade e do magnetismo se desenvolviam, surgiam correlações experimentais entre esses fenômenos tal que em 1849 Maxwell apresenta que a luz é uma onda que pode ser descrita por uma teoria ondulatória eletromagnética [18]. Essas ondas são chamadas de transversais e possuem dois campos oscilantes em sua propagação, que são os campos elétrico ($\vec{E}$) e magnético ($\vec{B}$), e a figura 5 ilustra a propagação da luz com velocidade c na direção $\hat{x}$ a partir da oscilação de $\vec{E}$ e $\vec{B}$. Esses dois campos são perpendiculares entre si e a onda eletromagnética resultante da oscilação desses campos propaga em uma direção que é perpendicular à $\vec{E}$ e $\vec{B}$. Com isso, temos que $\vec{E} \cdot \vec{B} = 0$ na direção de propagação da luz, e essa condição também pode ser generalizada de tal forma que sempre que $\vec{E} = 0$, então $\vec{B} = 0$ [18]. Sem entrar em mérito das discussões de casualidade envolvendo $\vec{E}$ e $\vec{B}$, a propagação perpendicular e simultânea desses campos é coerente com as Leis de Indução de Faraday e de Ampère.

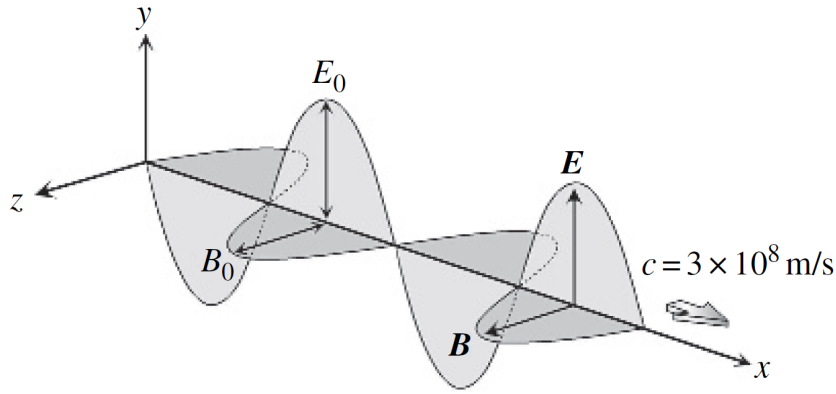


Figura 5 – Propagação de uma onda eletromagnética a partir do campo elétrico $\vec{E}$ e indução magnética $\vec{B}$. Adaptado de [18].

Embora a figura ilustre um valor de $c = 3 \times 10^8 \frac{\text{m}}{\text{s}}$, o valor mais adequado para situações com alta demanda de precisão é $c = 2.99792 \times 10^8 \frac{\text{m}}{\text{s}}$, que é a velocidade de propagação da luz no vácuo e ela independe do comprimento de onda da luz escolhida.

Partindo desse comportamento ondulatório da luz, podemos utilizar a equação (3.9) para representar os campos elétrico e magnético nas formas

$$E = E_0 e^{i(\omega t - kx + \delta)}, \quad (3.11)$$

$$B = B_0 e^{i(\omega t - kx + \delta)}. \quad (3.12)$$

Evocando agora o aspecto corpuscular da luz, temos então a proposta de Einstein em tratar a luz como um conjunto de partículas, que são chamadas de fótons (γ), munidas de energia. Dessa forma, podemos considerar que a luz é emitida ou absorvida na forma

de fótons [18], e a partir do momento que consideramos esse tratamento de partícula da luz, a energia E_γ de um fóton pode ser expressa como

$$E_\gamma = hv = \hbar\omega, \quad (3.13)$$

tal que h é a constante de Planck e $\hbar = \frac{h}{2\pi}$. Partindo disso, é conveniente usar a energia do elétron em unidade de elétron volts [eV] e em função do comprimento de onda λ :

$$E_\gamma = \frac{1.23984 \times 10^{-6}}{\lambda} \text{eV}. \quad (3.14)$$

3.3 Índice de refração

Quando a luz atravessa um meio e começa a propagar em outro, ocorre o fenômeno de refração, em que a luz altera sua velocidade de propagação para um valor s e também muda o seu caminho. Como meios diferentes de propagação podem resultar em valores distintos de velocidade para luz, é útil medir o efeito da refração a partir da constante chamada de índice de refração, que é expressa por

$$n = \frac{c}{s}, \quad (3.15)$$

tal que s é a velocidade de propagação da luz no meio e c é a velocidade da luz no vácuo. Por essa expressão, é notável que um alto índice de refração de um meio implica que a luz nele tenderá a ter baixa velocidade.

Se o meio analisado não absorve a luz que propaga em seu interior, então o número de onda K pode ser reescrito usando 3.7 e 3.15 como [18]

$$K = \frac{\omega n}{c} = \frac{2\pi n}{\lambda}, \quad (3.16)$$

tal que λ é o comprimento de onda no vácuo da luz refratada. Por consequência, temos que, para um mesmo meio, luzes com diferentes comprimentos de onda irão propagar nele com velocidades distintas, pois a velocidade s pode ser escrita em função de λ ao relacionar as expressões 3.15 e 3.16:

$$\lambda = \frac{2\pi nc}{Ks}. \quad (3.17)$$

O campo elétrico de uma onda eletromagnética que propaga em um meio transparente pode ser reescrito substituindo 3.16 em 3.11:

$$E = E_{t_0} e^{i(\omega t - Kx + \delta)} = E_{t_0} e^{i(\omega t - \frac{2\pi n}{\lambda} x + \delta)}, \quad (3.18)$$

tal que E_{t_0} é análogo a E_0 , mas em um meio transparente [18], e a figura 6(a) representa essa onda ao trocar de meio de propagação. Nessa situação, ao considerar que a luz pode

ser refratada e refletida, temos então $E_0 > E_{t_0}$. Para os meios transparentes, o fenômeno da refração pode ser completamente descrito usando o índice de refração.

Para um meio com absorção da luz, ilustrado pela figura 6(b), temos uma situação em que o fenômeno da refração não pode ser descrito completamente usando apenas o índice de refração n . Nesse caso, é necessário inserir um novo elemento físico para modelar essa situação. Para tal tarefa, vamos usar o coeficiente de extinção k (não confundir com o número de onda K), que em conjunto com n resultam no índice de refração complexo N :

$$N = n - ik. \quad (3.19)$$

Substituindo n da expressão 3.18 pelo índice de refração complexo N , teremos a equação para uma onda eletromagnética que se propaga em um meio absorvedor de luz [18]:

$$E = E_{t_0} e^{i(\omega t - \frac{2\pi N}{\lambda} x + \delta)} = E_{t_0} e^{-\frac{2\pi k x}{\lambda}} e^{i(\omega t - \frac{2\pi n}{\lambda} x + \delta)}. \quad (3.20)$$

Observando a figura 6(b) em conjunto com a expressão 3.20, uma das consequências para um campo eletromagnético propagando em um meio absorvedor é que a amplitude da luz decai ao longo da direção e sentido de propagação x . Além disso, a absorção de luz é dada pelo fator real $e^{-\frac{2\pi k x}{\lambda}}$ e propaga com comprimento de onda igual a $\frac{\lambda}{n}$.

A intensidade I de uma onda eletromagnética pode ser calculada usando [18]

$$I = |E|^2 = EE^*. \quad (3.21)$$

Se usarmos a expressão (3.20) na equação (3.21) e usando a expressão (3.10), temos a intensidade da onda eletromagnética ao longo do meio [18]:

$$I = |E_{t_0} e^{-\frac{2\pi k x}{\lambda}}|^2 = |E_{t_0}|^2 e^{-\frac{4\pi k x}{\lambda}}. \quad (3.22)$$

Essa intensidade no meio é descrita pela lei empírica de Beer dada por

$$I = I_0 e^{-\alpha d}, \quad (3.23)$$

em que α é o coeficiente do meio e d é a distância a partir da superfície. Combinando as equações (3.22) e (3.23), temos [18]:

$$\alpha = \frac{4\pi k}{\lambda}, \quad (3.24)$$

em que, se o meio não absorve luz, temos $\alpha = k = 0$. Para um meio absorvedor, a intensidade da luz diminui enquanto propaga, tal que quanto maior a distância d percorrida maior será a perda. Se $\frac{I}{I_0} = \frac{1}{e} \approx 0.37$ na equação 3.23, temos que $\alpha d = 1$, e essa profundidade é chamada de profundidade de penetração (*penetration depth*) [18]:

$$d_d = \frac{1}{\alpha}. \quad (3.25)$$

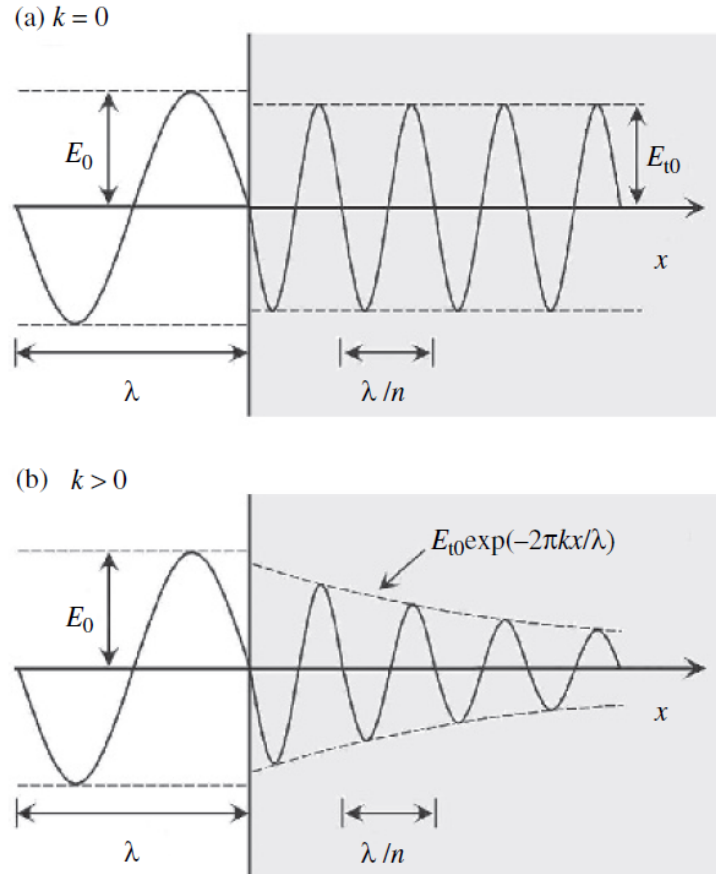


Figura 6 – Onda eletromagnética propagando em (a) um meio transparente ($k = 0$) e (b) um meio com absorção de luz ($k > 0$). Adaptado de [18].

3.4 Dielétricos

Ao aplicar um campo elétrico em um material, as consequências no alvo dependerão do tipo de interação de seus componentes. Em materiais que não possuem um conjunto de cargas livres em sua composição devido a, por exemplo, fortes ligações químicas entre os componentes, ocorre que a presença do campo elétrico não é suficiente para provocar um grande fluxo de cargas livres, mas pode ser o bastante por modificar de forma sutil a distribuição espacial das cargas positivas e negativas dos átomos ou moléculas do material [18]. Essa pequena mudança espacial pode ser suficiente para criar regiões positivas e negativas no interior do alvo, gerando uma polarização dielétrica. Logo, materiais que apresentem essas características perante um campo elétrico são chamados de dielétricos.

A figura 7 ilustra exemplos típicos de polarização em dielétricos. No caso da figura 7(d), temos duas cargas opostas de módulo $|q|$ separadas por uma distância l , proporcionando um momento de dipolo definido por

$$\vec{\mu} = q\vec{l}. \quad (3.26)$$

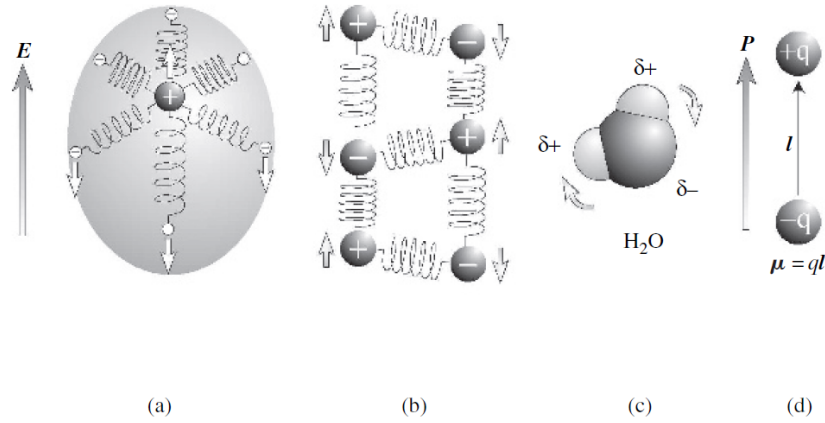


Figura 7 – Tipos de polarizações dielétricas: polarização elétrica (a), polarização atômica (b), polarização orientada (c) e momento de dipolo (d). Adaptado de [18].

Partindo da equação (3.26), é útil associar o momento de dipolo com o volume de um corpo, pois podemos fazer uso dessa grandeza para objetos em escala macroscópica. Se usarmos a soma dos momentos de dipolo μ_i de um objeto por unidade de volume, teremos a polarização dielétrica associada a ele [18]:

$$\vec{P} = \sum_i \mu_i. \quad (3.27)$$

Pela figura 7 (d), observamos que polarização dielétrica é um vetor que tem direção e sentido que partem da carga negativa até a carga positiva, em contraste com o que ocorre para o caso dos campos elétricos [18].

O estudo dos dielétricos requer o uso de uma propriedade associada às características desses materiais. Vamos propor essa propriedade por meio de um exemplo empírico em que, conforme a figura 8 [18], um dielétrico é posto entre duas placas paralelas de um capacitor, que opera com um campo elétrico alternado. Isso faz com que apareça uma densidade de cargas superficiais no capacitor denotada por $\vec{D}$, conhecida também por deslocamento elétrico. Se o meio entre as placas do capacitor for o vácuo, a relação experimental entre o campo elétrico $\vec{E}$ e o deslocamento elétrico $\vec{D}$ é dada por:

$$\vec{E} = \frac{\vec{D}}{\epsilon_0}. \quad (3.28)$$

Nesse caso, o campo elétrico é proporcional ao deslocamento elétrico, e a constante ϵ_0 que faz valer a equação (3.28) é a permissividade do vácuo.

O resultado da equação (3.28) pode ser generalizado para outros dielétricos, desde que algumas considerações sejam feitas. Se a polarização $\vec{P}$ em um dielétrico for proporcional ao campo elétrico $\vec{E}$ que a produz, desde que $\vec{E}$ não seja muito grande, estaremos abordando dielétricos lineares [19]. Para esse tipo de dielétrico, se o meio entre as placas

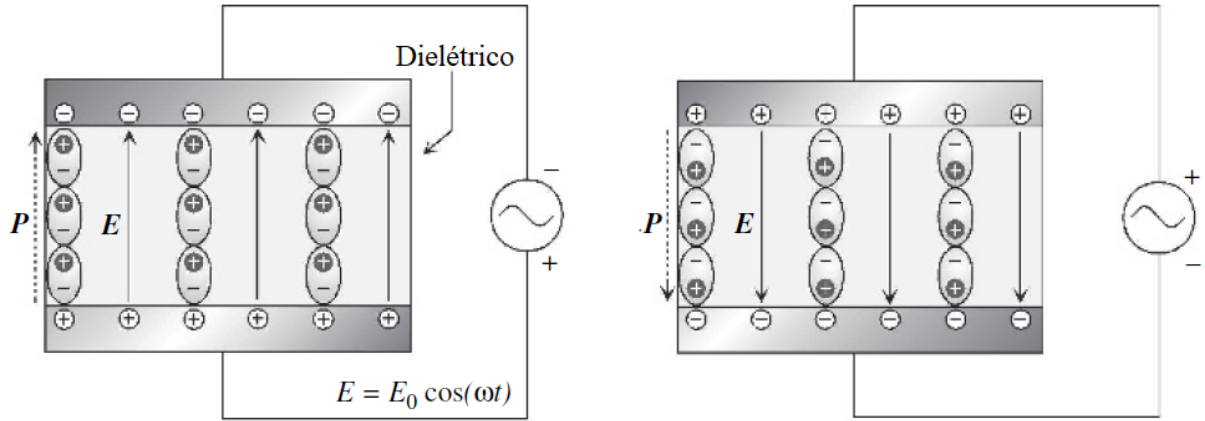


Figura 8 – Dielétrico entre um capacitor de placas paralelas. Adaptado de [18].

tiver uma permissividade ϵ_p , então temos:

$$\vec{E} = \frac{\vec{D}}{\epsilon_p}. \quad (3.29)$$

Nessa situação, considerando a polarização $\vec{P}$ provocada no material pelo campo $\vec{E}$, a expressão (3.29) pode ser reescrita considerando a polarização [18]:

$$\vec{E} = \frac{\vec{D} - \vec{P}}{\epsilon_0}. \quad (3.30)$$

Dada a importância da permissividade, podemos usar a forma conveniente de uma permissividade relativa:

$$\epsilon = \frac{\epsilon_p}{\epsilon_0}. \quad (3.31)$$

Por convenção, ϵ será chamada de constante dielétrica e podemos escrevê-la como [18]:

$$\epsilon = 1 + \frac{P}{\epsilon_0 E} = 1 + \chi, \quad (3.32)$$

considerando que χ é susceptibilidade elétrica do meio e as quantidades P e E são o módulo da polarização e do campo elétrico respectivamente. Analisando a expressão (3.32), é notável que o aumento da polarização dielétrica ou da suscetibilidade aumentam o valor de ϵ [18].

Um fenômeno valioso de ser analisado é a maneira como a propagação da luz no interior de um dielétrico afeta a polarização desse material. O campo elétrico de ondas eletromagnéticas oscila no tempo e espaço de maneira senoidal. Como a polarização dielétrica está diretamente ligada às características do campo elétrico que a gera, é possível concluir que a polarização no interior do material também irá sofrer as consequências desse campo elétrico oscilante [18].

A figura 9 mostra a interação do campo eletromagnético com os dipolos do meio em que a luz está propagando. O campo elétrico oscilante é aplicado na direção que une os dois polos. Quando o campo é positivo (figura 9 (a-b)), os dois polos se afastam e assim permanecem até que o campo decaia para zero e os polos retornem para a posição de equilíbrio (figura 9 (e)). Em seguida, o campo agora tem valor negativo e provoca um novo afastamento dos polos (figura 9 (f-g)) e, posteriormente, o campo retorna a zero, fazendo com que os polos retornem a uma posição de equilíbrio. Esse processo de movimento dos polos provoca uma aceleração neles, fazendo com que ondas eletromagnéticas sejam emitidas pelos dipolos com a mesma frequência das ondas incidentes no material [18].

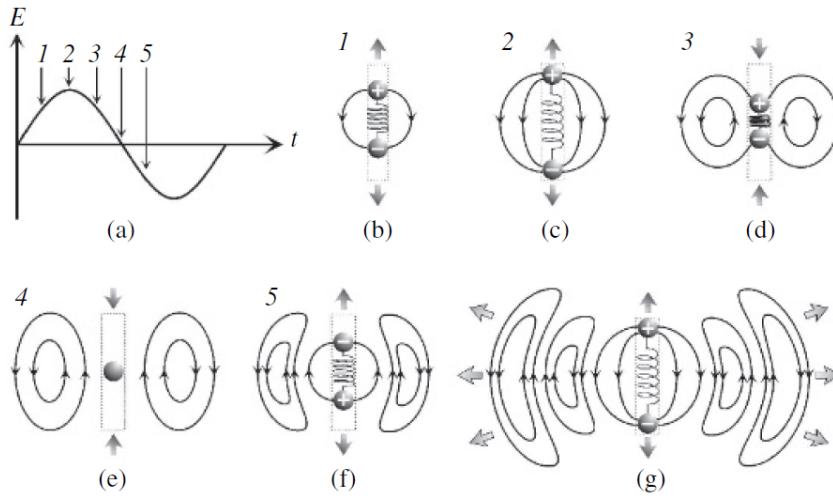


Figura 9 – Propagação de radiação de um dipolo em um dielétrico: em (a) o campo elétrico externo oscilante no tempo é aplicado no dielétrico e em (b-f) temos o comportamento do material quando os pontos 1-5 do campo elétrico (a) são aplicados. Adaptado de [18].

Considerando que essa luz se propaga por uma rede de átomos, as ondas eletromagnéticas emitidas vão se sobrepor de forma construtiva e propagam pelo material como uma única onda. Como os comprimentos de onda da luz são muito maiores do que as distâncias entre cada átomo de uma rede, é sempre possível achar átomos que estejam a uma distância de λ um do outro, fazendo com que a propagação de onda mencionada anteriormente seja possível [18].

Meios com alto índice de refração tendem a ter uma velocidade baixa de propagação de luz, e podemos relacionar agora o índice de refração com a constante dielétrica para explicar esse fenômeno. Para valores altos de ϵ , há uma tendência maior do meio gerar dipolos elétricos durante a propagação de uma onda eletromagnética. A geração desses dipolos causa pequenos atrasos na propagação da luz, e isso reduz a velocidade da onda no meio. A partir das equações de Maxwell para condutores, o índice de refração complexo pode ser escrito a partir de ϵ por [18]

$$N^2 = \epsilon, \quad (3.33)$$

em que, nessa equação, a constante dielétrica é um número complexo da forma

$$\epsilon = \epsilon_1 - i\epsilon_2. \quad (3.34)$$

Então, partindo das equações (3.19), (3.33) e (3.34), temos

$$\epsilon_1 = n^2 - k^2, \quad (3.35)$$

$$\epsilon_2 = 2nk. \quad (3.36)$$

Experimentalmente, os dielétricos podem apresentar comportamentos muito complicados, e a montagem de um modelo pode ser tarefa árdua. A constante dielétrica do meio varia e ela depende da resposta que o dielétrico apresenta diante de uma onda eletromagnética. A resposta que um dielétrico dá em função da frequência angular ou energia de fóton de uma fonte de luz é chamada de função dielétrica ou dispersão dielétrica [18]. Ao longo desta escrita, será usado o termo perfil dielétrico para se referir às características gerais ligadas a um material dielétrico.

O comportamento de um dielétrico pode ser observado ao analisar a parte real e complexa da constante dielétrica em função da frequência angular, e isso está ilustrado nos gráficos da figura 10, que está em escala logarítmica no eixo de ω . Para frequências baixas, ϵ_1 apresenta um valor constante de ϵ_s , que é um valor que considera as polarizações atômicas e elétricas [18]. A luz que propaga pelo meio será absorvida desde que o material tenha polos que oscilem na mesma frequência que essa onda eletromagnética, e isso será citado a partir deste momento como uma absorção por frequência ressonante. A absorção ligada à polarização atômica é geralmente observada na região infravermelha, enquanto a absorção para polarização eletrônica ocorre na região UV/visível [18]. Para frequências ressonantes, as absorções podem ser verificadas nos picos da figura 10, posto que o meio absorve a luz apenas para valores específicos de ω .

Pela função dielétrica de ϵ_1 da figura 10, é possível verificar que, conforme a polarização atômica está confinada a ocorrer dentro da faixa do infravermelho, o aumento da frequência angular da luz faz com que essa polarização acabe após transgredir a faixa do infravermelho. Dessa forma, ϵ_1 terá agora o valor da constante dielétrica ϵ_∞ de altas frequências [18]. A partir disso, se as frequências aumentam ainda mais, as condições para a polarização elétrica também não serão possíveis, e não restará mais polarização no material, resultando que $\epsilon_1 = 1$ [18].

Quando existem elétrons livres e buracos em um meio, a absorção de luz ocorre por intermédio desses portadores livres, conforme é possível verificar no gráfico da função dielétrica para ϵ_2 da figura 10. Em metais ou semicondutores que possuem alta concentração de portadores, observamos a função dielétrica indicada pelas linhas pontilhadas da figura 10. A absorção pelos portadores livres aumenta para frequências angulares mais baixas,

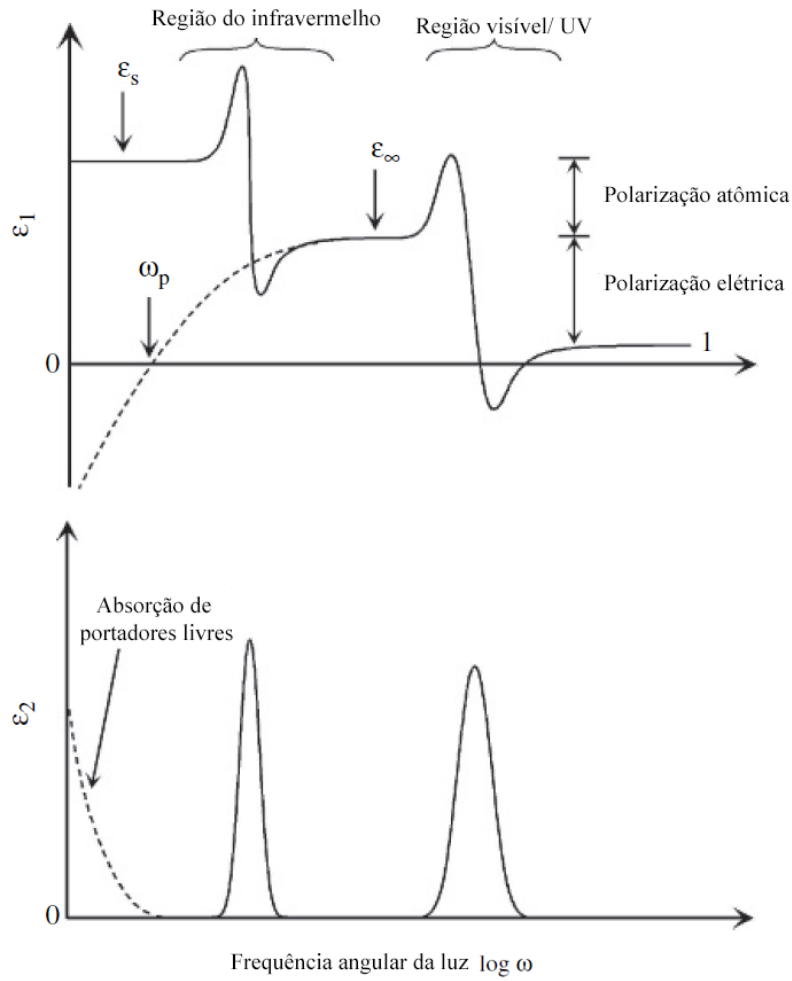


Figura 10 – Gráfico da parte real e imaginária da constante dielétrica complexa em função da frequência angular da luz $\log(\omega)$. Adaptado de [18].

e os valores de ϵ_1 se tornam negativos em frequências inferiores à frequência angular do plasma ω_p [18].

A partir dessas considerações sobre a constante dielétrica e a função dielétrica, podemos concluir que essas grandezas físicas são ferramentas úteis para a obtenção de informações sobre os materiais. Como a constante dielétrica está relacionada à interação da luz com a matéria e é sensível aos átomos, arranjos eletrônicos e aos mecanismos de polarização, cada material apresenta uma função dielétrica distinta. Dessa forma, ao retomar o objetivo de diagnóstico da malária por meio do soro sanguíneo, é possível inferir que a constante dielétrica constitui uma medida física adequada para diferenciar amostras negativas de reativas para malária. Isso se deve ao fato de que essa grandeza é sensível às alterações físicas e químicas que ocorrem no soro sanguíneo em decorrência da infecção pelo parasita da malária.

3.5 Reflexão e transmissão da luz

A reflexão e a transmissão da luz são fenômenos determinados a partir dos índices de refração complexos (ou constantes dielétricas complexas) dos meios de propagação [18]. O entendimento e fundamentação desses fenômenos é de extrema importância em técnicas ópticas para análise de espectros.

3.5.1 Refração da luz

Quando a luz se move para um meio com incidência oblíqua, a direção de propagação dela geralmente muda devido à refração que a luz sofre. Esse fenômeno pode ser interpretado a partir da variação da velocidade da luz em uma interface de dois meios [18]. A figura 11 mostra a luz propagando em um meio de índice de refração n_i e incidindo obliquamente em uma interface com um meio de índice n_t . A luz incidente é uma onda plana que tem a mesma fase no plano, e a fase constante do campo elétrico é indicada pelas linhas [18]. Quando essa luz propaga pelo meio, os átomos dele emitem luz, e a propagação luminosa é mantida. Em (a), a radiação por dipolo elétrico ocorre sequencialmente a partir do átomo localizado no lado esquerdo [18] e, conforme as outras parcelas da interface também recebem frentes de ondas, os outros átomos também vão auxiliar na propagação no interior do material. Se em (a) uma onda incidente vai do ponto B para D durante um tempo t , a onda transmitida avança do ponto A para E, e a onda refletida se move até o ponto C [18]. Definindo θ_i , θ_r e θ_t como sendo os ângulos de incidência, reflexão e transmissão, respectivamente, a partir de (a), temos

$$\frac{\overline{BD}}{\overline{BD}} = \frac{\overline{AC}}{\overline{AC}} = \frac{\overline{AE}}{\overline{AE}} = \frac{1}{\overline{AD}}. \quad (3.37)$$

Como a velocidade da luz em um meio é dada por $s = \frac{c}{n}$ e a luz percorre uma distância $st = \frac{ct}{n}$ durante um tempo t , podemos inferir que [18]

$$\overline{BD} = \frac{ct}{n_i}, \quad (3.38a)$$

$$\overline{AC} = \frac{ct}{n_r}, \quad (3.38b)$$

$$\overline{AE} = \frac{ct}{n_t}, \quad (3.38c)$$

tal que n_i , n_r e n_t são, respectivamente, os índices de refração para o incidência, reflexão e transmissão. Nesse caso, como $n_i = n_r$, então $\overline{BD} = \overline{AC}$, e a consequência disso na expressão (3.37) é a canônica lei da reflexão [18]

$$\theta_i = \theta_r. \quad (3.39)$$

Ao combinarmos as equações (3.37) e (3.38), temos a lei de Snell:

$$n_i \sin \theta_i = n_t \sin \theta_t. \quad (3.40)$$

Pela figura 11(b), se assumirmos que há um círculo na interface que a luz incide, obtemos $L_i = L_r$ e $\frac{n_t}{n_i} = \frac{L_i}{L_t}$ da lei de reflexão e da lei de Snell, respectivamente [18]. A lei de Snell também pode ser aplicada usando o índice de refração complexo, tal que θ_i e θ_t são agora números complexos. Porém, este tratamento torna mais difícil de visualizar o processo de propagação da luz [18]. Neste caso, a lei de Snell é expressa por

$$N_i \sin \theta_i = N_t \sin \theta_t. \quad (3.41)$$

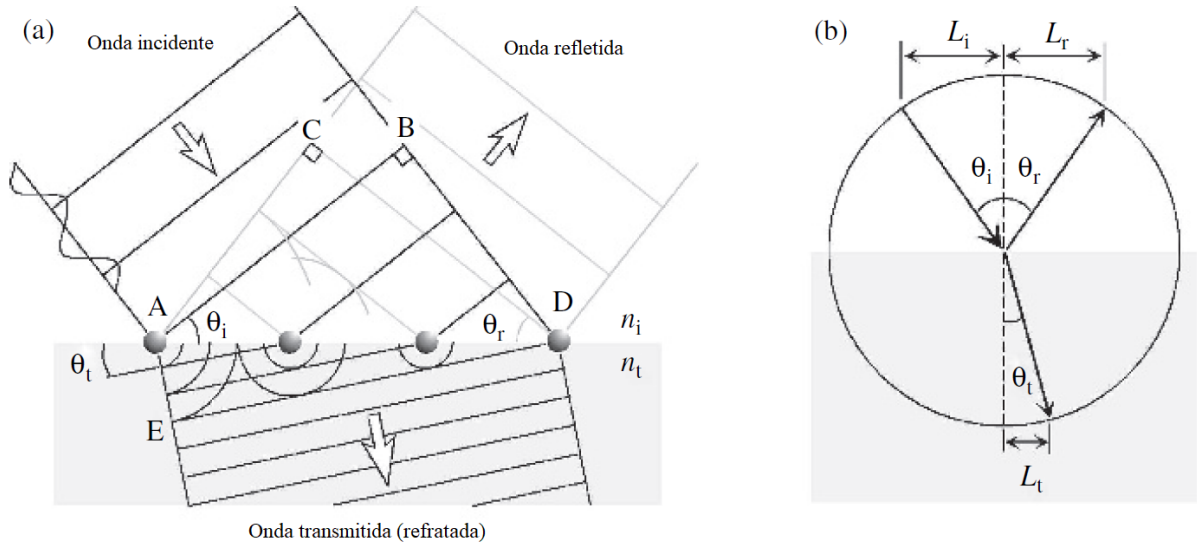


Figura 11 – Reflexão e transmissão da luz para uma incidência oblíqua entre dois meios em que $n_i < n_t$. Adaptado de [18].

3.5.2 Luz polarizada em -p e -s

Estamos representando a propagação da luz usando o campo elétrico que compõe uma onda eletromagnética. O caráter vetorial do campo elétrico permite que ele possa ser representado por uma soma de vetores que resulte nesse mesmo vetor de campo elétrico. Dessa forma, podemos decompor um campo elétrico em componentes vetoriais que sejam convenientes para um determinado modelo ou situação experimental. Para a análise de uma onda eletromagnética incidente sobre um plano, é comum decompor a luz nas componentes p e s , tal que p é uma componente paralela ao plano de incidência e s é perpendicular a ele, conforme está ilustrado na figura 12. A figura 13 mostra uma luz incidindo e sendo refletida e transmitida por um plano a partir dos campos elétrico $\vec{E}$ e indução magnética $\vec{B}$. O valor do campo elétrico de incidência, reflexão e transmissão são, respectivamente, $\vec{E}_i = \vec{E}_{ip} + \vec{E}_{is}$, $\vec{E}_r = \vec{E}_{rp} + \vec{E}_{rs}$ e $\vec{E}_t = \vec{E}_{tp} + \vec{E}_{ts}$, tal que os índices i , r e t indicam incidência, reflexão e transmissão respectivamente, e p e s são as direções de polarização. Em (a) temos as componentes elétricas para a polarização- p , e se considerarmos a continuidade que as

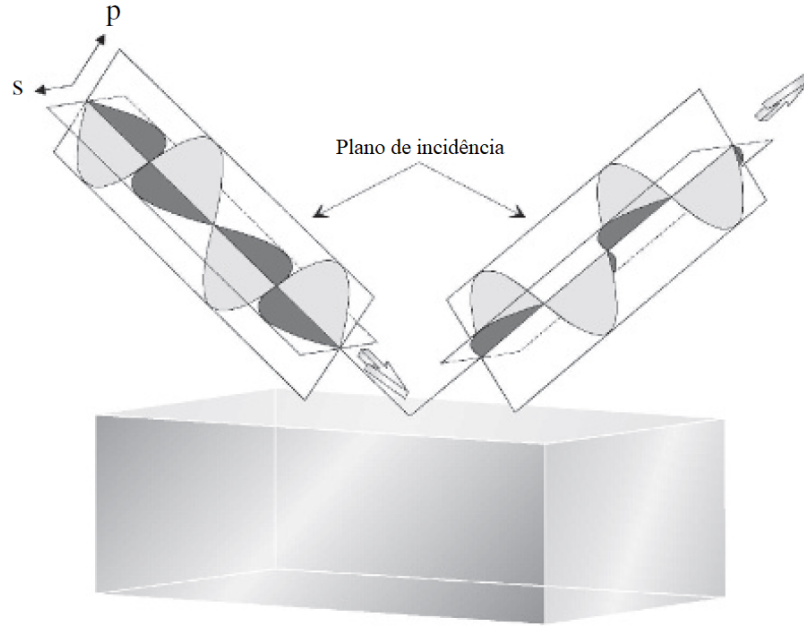


Figura 12 – Reflexão das luzes polarizadas-p e -s. Adaptado de [18].

componentes paralelas devem obedecer na interface dos meios e usando $\theta_i = \theta_r$, temos então o coeficiente de reflexão de amplitude r_p e de transmissão t_p para a polarização-p [18]:

$$r_p = \frac{E_{rp}}{E_{ip}} = \frac{n_t \cos \theta_i - n_i \cos \theta_t}{n_t \cos \theta_i + n_i \cos \theta_t}, \quad (3.42)$$

$$t_p = \frac{E_{tp}}{E_{ip}} = \frac{2n_i \cos \theta_i}{n_t \cos \theta_i + n_i \cos \theta_t}. \quad (3.43)$$

De modo similar para a situação (b), que apresenta as componentes para a polarização-s, temos o coeficiente de reflexão de amplitude r_s e de transmissão t_s [18]:

$$r_s = \frac{E_{rs}}{E_{is}} = \frac{n_i \cos \theta_i - n_t \cos \theta_t}{n_i \cos \theta_i + n_t \cos \theta_t}, \quad (3.44)$$

$$t_s = \frac{E_{ts}}{E_{is}} = \frac{2n_i \cos \theta_i}{n_i \cos \theta_i + n_t \cos \theta_t}. \quad (3.45)$$

As expressões (3.42) e (3.44) são quantidades usadas para caracterização de materiais usando espectroscopia, pois essas razões de amplitude dependem das características da superfície em que a luz incide. No capítulo sobre elipsometria, veremos que uma importante medida experimental dessa técnica de espectroscopia é a razão $\frac{r_p}{r_s}$. Então, considerando que há um contexto experimental que envolve essas razões de amplitude, podemos pensar em boas práticas experimentais para medi-las. Um procedimento experimental adequado é usar um ângulo de incidência de forma que r_p e r_s tenham uma diferença entre si que seja maximizada. Empiricamente, podemos usar um ângulo de incidência tal que $E_{rp} = 0$, que significa que uma luz que incide por esse ângulo terá sua componente de polarização-p totalmente anulada após sofrer uma reflexão, resultando em $r_p = 0$. O ângulo que é capaz

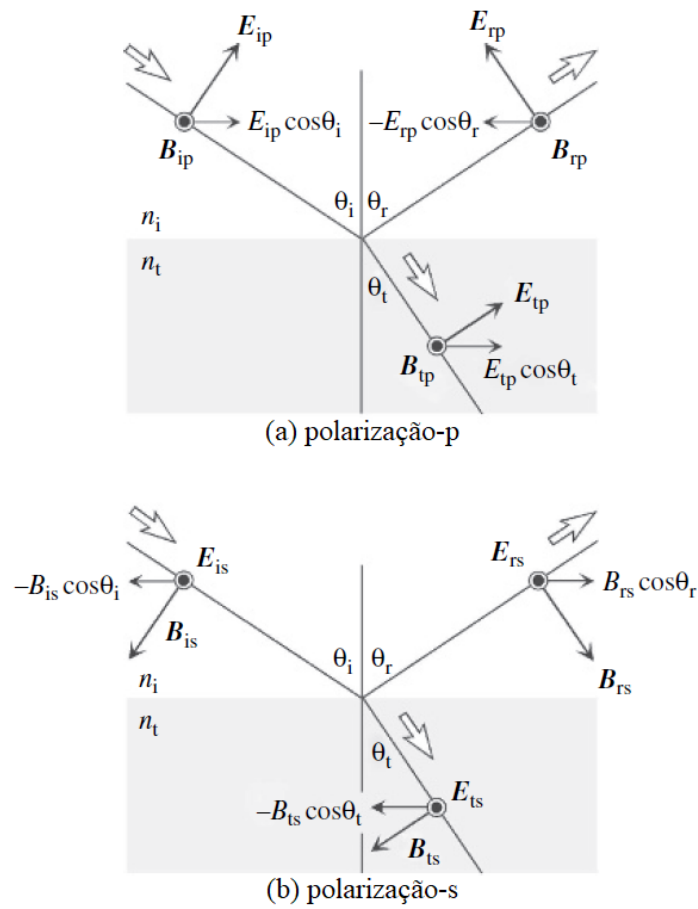


Figura 13 – Esquema detalhado de incidência, reflexão e transmissão para a luz em (a) polarização-p e (b) polarização-s. Adaptado de [18].

disso é chamado de ângulo de Brewster, que pode maximizar a diferença entre r_p e r_s , promovendo uma medida espectroscópica mais sensível para a razão dos coeficientes de reflexão de amplitude $\frac{r_p}{r_s}$ [18].

4 Formalismo de Jones e Mueller

Como estamos interessados na constante dielétrica para caracterizar amostras de soro sanguíneo no diagnóstico de malária, é preciso ter recursos para representar a interação da luz com a matéria de maneira eficiente, pois a constante dielétrica é uma medida física relacionada a essa interação. Partindo desse contexto, como as propriedades de ondas eletromagnéticas são importantes nas técnicas de espectroscopia, há um formalismo físico e matemático para caracterizar os elementos, medições e fenômenos de interesse. A representação da luz polarizada e não polarizada é feita por meio do formalismo dos vetores de Jones e de Stokes. Esses vetores podem retratar as luzes incidentes, refletidas e transmitidas, tal que é possível convertê-los uns nos outros usando operadores matriciais do cálculo de Jones e de Mueller.

É fundamental, no uso da técnica de elipsometria espectroscópica, saber tratar a luz polarizada e representar matematicamente, de forma eficiente, a luz que incide, reflete e refrata por um material. Para tal, serão construídos os fundamentos para representação da luz polarizada e não polarizada, buscando sempre descrever as principais características da luz para o uso de técnicas espectroscópicas. Então, o objetivo dessa seção é apresentar o cálculo de Jones e Muller como ferramentas para representar a luz, medidas e elementos ópticos.

4.1 Polarização da luz

A luz polarizada é uma onda eletromagnética que está orientada em direções específicas ou em padrões circulares notáveis, e se uma onda luminosa oscila seus campos em direções aleatórias, classificamos essa onda como sendo não polarizada [18]. Se considerarmos um conjunto de ondas propagando na mesma direção, elas vão se sobrepor de forma que a fase entre elas pode criar um padrão de oscilação. Então, essa luz resultante pode ter uma polarização diferente das ondas que a constituem, e isso dependerá da fase de cada onda luminosa envolvida no processo de sobreposição. Logo, a luz pode ser polarizada de maneiras variadas.

Para um sistema de três eixos x, y, z , uma luz que propaga na direção $\hat{z}$ pode ter sua polarização representada pela projeção do campo elétrico nos eixos x e y . Nesse caso, a sobreposição dos campos nesses eixos resulta [18]:

$$E(z, t) = E_x(z, t) + E_y(z, t) = E_{x0}e^{i(\omega t - Kz + \delta_x)}\hat{x} + E_{y0}e^{i(\omega t - Kz + \delta_y)}\hat{y}. \quad (4.1)$$

Na descrição dos estados de polarização, o valor absoluto das fases δ_x e δ_y não são necessários, pois o que importa nesse caso é a diferença entre essas fases $(\delta_x - \delta_y)$ ou $(\delta_y - \delta_x)$

[18]. Isso é um resultado importante e fundamental para a elipsometria espectroscópica, porque as medidas experimentais dessa técnica levam em consideração a diferença de fase, e não os seus valores absolutos.

A figura 14 mostra três tipos de polarização para uma luz propagando na direção $\hat{z}$ em que as amplitudes nos eixos x e y são iguais ($E_{x_0} = E_{y_0}$), com diferença de fase ($\delta_y - \delta_x$) e $K = 1$. Na situação (a) temos ($\delta_y - \delta_x = 0$), então os campos elétricos da onda sempre oscilam a 45° no plano xy , caracterizando uma polarização linear. O caso (b) apresenta uma diferença de fase de $\delta_y - \delta_x = \frac{\pi}{2}$ entre $\vec{E}_x$ e $\vec{E}_y$, e essa polarização é chamada de circular. Nesse caso, sempre que $|\vec{E}_x| = 0$ temos $|\vec{E}_y| = E_{y_0}$, e o oposto também é válido. Para a configuração (c), há uma diferença de fase ($\delta_y - \delta_x = \frac{\pi}{4}$), e essa situação é chamada de polarização elíptica [18].

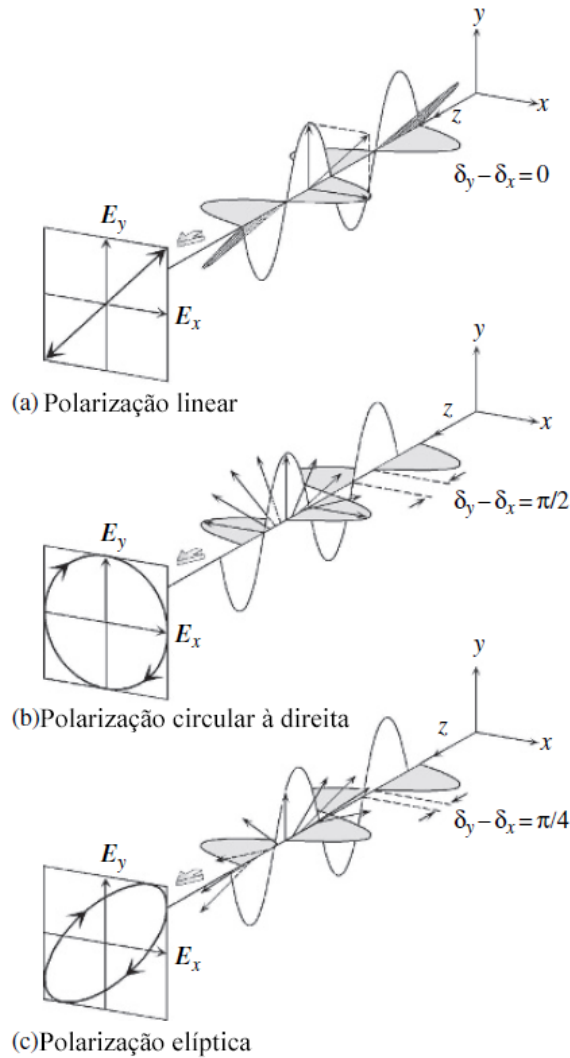


Figura 14 – Representação da (a) polarização linear, (b) polarização circular à direita e (c) polarização elíptica. Adaptado de [18].

Então, a diferença de fase entre ondas é de grande importância para determinar o

tipo de polarização que a luz apresenta. Na figura 15, observamos os tipos de polarização que surgem ao mudar a diferença de fase de 0 até 2π . É importante notar que não ocorre apenas uma mudança no tipo de polarização (linear, circular e elíptica), mas também ocorre uma mudança na orientação e sentido dessas polarizações. Para $(\delta_y - \delta_x) = 0$ e $(\delta_y - \delta_x) = \pi$ temos, respectivamente, polarizações lineares a 45° e -45° . Para $(\delta_y - \delta_x) = \frac{\pi}{4}$ e $(\delta_y - \delta_x) = \frac{3\pi}{4}$ a polarização é elíptica à direita, e para $(\delta_y - \delta_x) = \frac{5\pi}{4}$ e $(\delta_y - \delta_x) = \frac{7\pi}{4}$ é para a esquerda. Nas situações $(\delta_y - \delta_x) = \frac{\pi}{2}$ e $(\delta_y - \delta_x) = \frac{3\pi}{2}$ as polarizações são, respectivamente, circulares para a direita e esquerda. A partir de agora, as polarizações

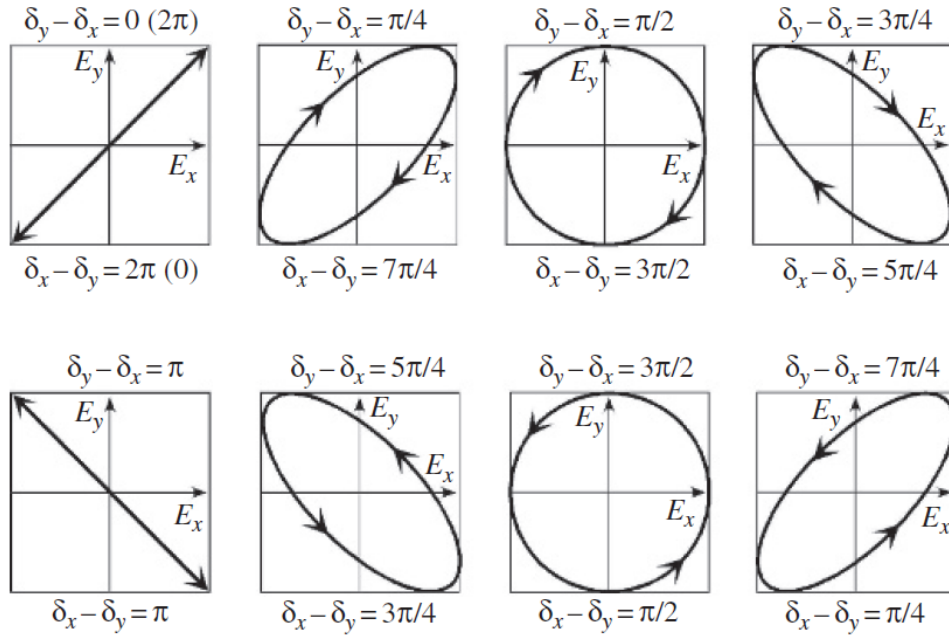


Figura 15 – Mudança de polarização por meio da variação da diferença de fase $(\delta_y - \delta_x)$. Adaptado de [18].

lineares $(x, y, +45^\circ, -45^\circ)$ e polarizações circulares (direita, esquerda) serão referidas como polarizações canônicas, pois serão frequentemente utilizadas. Isso porém não significa que esses sejam os únicos estados de polarizações possíveis de serem obtidos.

4.2 Elementos ópticos

Por se tratar de uma técnica experimental, a elipsometria espectroscópica demanda de meios capazes de processar a luz para alguma finalidade desejada. Para que isso ocorra, é usado um conjunto de elementos ópticos: polarizador/ analisador, compensador, modulador fotoelástico e despolarizador. Todos esses elementos integram um instrumento óptico para determinar o estado de polarização da luz.

4.2.1 Polarizador e analisador

Polarizador e analisador representam o mesmo elemento óptico, mas desempenham papéis diferentes. O polarizador é usado para fazer com que uma luz não polarizada seja convertida para uma polarização linear, e é geralmente colocado de frente para a fonte de luz a ser polarizada. No caso do analisador, ele é colocado próximo a um detector de luz para determinar o estado de polarização da luz incidente por meio da intensidade luminosa que chega até ele [18]. Logo, para uma fonte em que a intensidade é conhecida, é possível determinar o estado de polarização da luz usando um analisador.

De forma geral, o polarizador e o analisador são feitos de prismas de $CaCO_3$ (calcita), e a figura 16(a) mostra a estrutura desse cristal com os raios ordinário ($\vec{E}_o$) e extraordinário ($\vec{E}_e$), que oscilam, respectivamente, paralelamente e perpendicularmente ao plano desta escrita. A maneira com que a calcita é formada permite uma concentração de elétrons um pouco maior na parcela CO_3 . Conforme foi discutido anteriormente nos fundamentos de óptica, a constante dielétrica e o índice de refração de um material, para a faixa do visível e do UV, são determinados a partir da polarização elétrica do meio. Então, é razoável inferir que a calcita terá regiões com índice de refração e constante dielétrica diferentes, pois a alta polarização dielétrica resulta em um aumento do índice de refração [18]. Nesse caso, $\vec{E}_o$ e $\vec{E}_e$ propagam em porções do meio que têm, respectivamente, os índices de refração valendo n_o e n_e , tal que $n_o > n_e$. Devido à distribuição de elétrons de maneira não uniforme na calcita, podemos considerar que esse é um meio anisotrópico com relação a essa distribuição de cargas, e essa anisotropia está no plano perpendicular a esta escrita, fazendo com que o eixo óptico esteja na direção de $\vec{E}_e$. Essa densidade não uniforme de elétrons faz com que a luz propague com uma velocidade maior (eixo rápido) na direção do raio extraordinário e menor (eixo lento) para o raio ordinário. Então, nas direções de $\vec{E}_o$ e $\vec{E}_e$ os comprimentos de onda são $\frac{\lambda}{n_o}$ e $\frac{\lambda}{n_e}$ respectivamente. Vale pontuar que nem sempre o eixo rápido e o lento correspondem às direções dos raios extraordinários e ordinários respectivamente, pois há situações, como no caso do quartzo, em que a direção do raio ordinário é a mesma do eixo rápido, e, para este material, $n_o < n_e$ [18].

A figura 17 mostra um prisma de Glan-Taylor [18], que é formado por dois prismas capazes de extrair luz polarizada de uma fonte luminosa não polarizada. Na figura 17(a) temos a atuação desse prisma na luz polarizada tal que, de toda a luz incidente, apenas a parcela paralela à direção de $\vec{E}_x$ atravessou o prisma, e chamamos de eixo de transmissão aquele que permite esse fenômeno. A figura 17(b) mostra uma visão de cima da figura 17(a) com os dois prismas e um camada de ar entre eles, formando o prisma de Glan-Taylor. Podemos ver que os eixos ópticos dos dois prismas são paralelos entre si e repousam no plano desta escrita. Para uma luz não polarizada que propaga da esquerda para a direita em um meio de índice de refração n_i , é possível que essa onda luminosa seja completamente refletida na interface prisma/ar se usarmos o ângulo crítico $\text{sen}\theta_c = \frac{1}{n_i}$ [18].

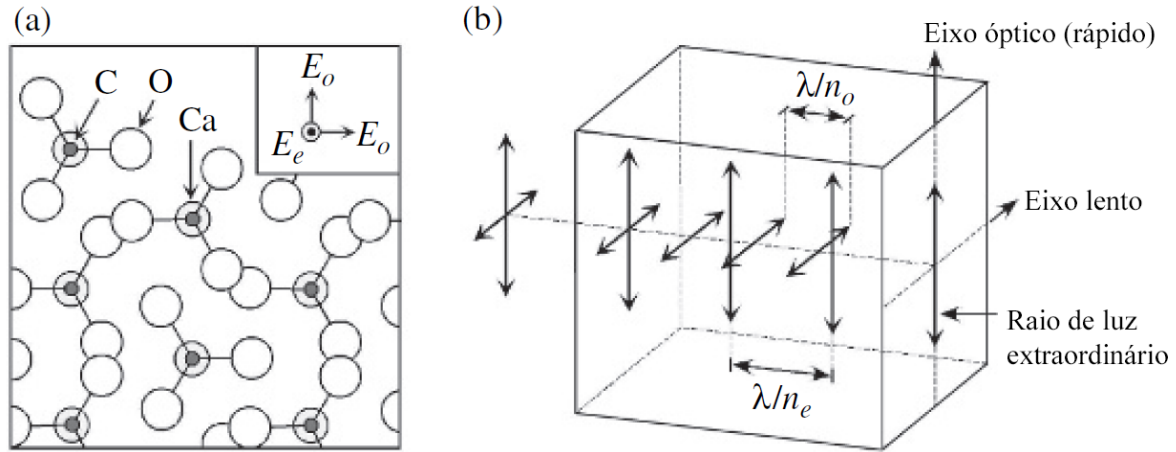


Figura 16 – (a) Estrutura da calcita a partir do eixo óptico e (b) propagação de luzes em uma amostra de calcita. Adaptado de [18].

Porém, como esses prismas são feitos de calcita, esse material apresenta um índice de refração que depende do eixo em que a luz se propaga. Dessa forma, como para calcita vale $n_o > n_e$, o ângulo crítico para o eixo do raio ordinário é menor do que o do raio extraordinário. Logo, é possível escolher um ângulo θ que reflita totalmente o raio ordinário na interface prisma/ar, mas que permita que o raio extraordinário atravesse todo o prisma de Glan-Taylor. Dessa forma, o raio extraordinário será a única parcela de luz a atravessar o prisma, gerando uma luz polarizada ao fim do processo. A condição para que isso ocorra é dada por $\frac{1}{n_o} < \sin\theta < \frac{1}{n_e}$.

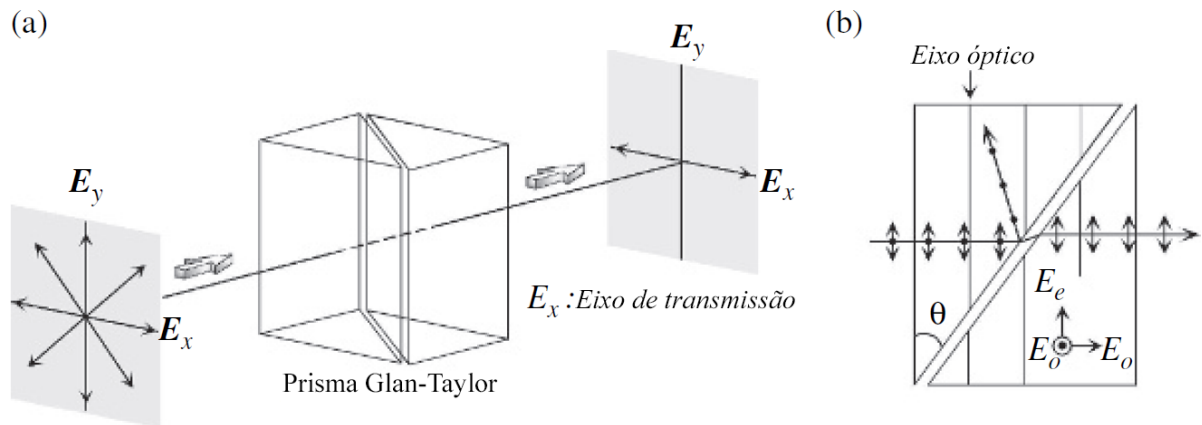


Figura 17 – (a) Prisma de Glan-Taylor usado como polarizador/analizador e (b) sua estrutura. Adaptado de [18].

4.2.2 Compensador e retardador

O compensador (retardador) é usado para converter uma luz linearmente polarizada em circularmente polarizada e é um elemento colocado atrás de um polarizador ou na frente de um analisador [18]. Para fazer essa mudança de polarização, o compensador deve ser feito de um material que apresente anisotropia para o índice de refração, pois essa característica faz com que a luz incidente nesse elemento óptico tenha velocidades diferentes para eixos de propagação distintos. Com isso, uma diferença de fase δ é gerada entre os raios de luz e é calculada usando os índices de refração extraordinário e ordinário e a espessura d do compensador:

$$\delta = \frac{2\pi|n_e - n_o|d}{\lambda} \quad (4.2)$$

A figura 18 ilustra o funcionamento de um compensador ao transformar uma luz linearmente polarizada a 45° em circularmente polarizada para a esquerda. Como $\vec{E}_x$ e $\vec{E}_y$ estão alinhados com os eixos rápido e lento respectivamente, uma diferença de fase é gerada entre esses componentes do campo elétrico. Nesse exemplo, o compensador tem uma anisotropia que permite que essa diferença de fase seja de $\delta = \frac{\pi}{2}$.

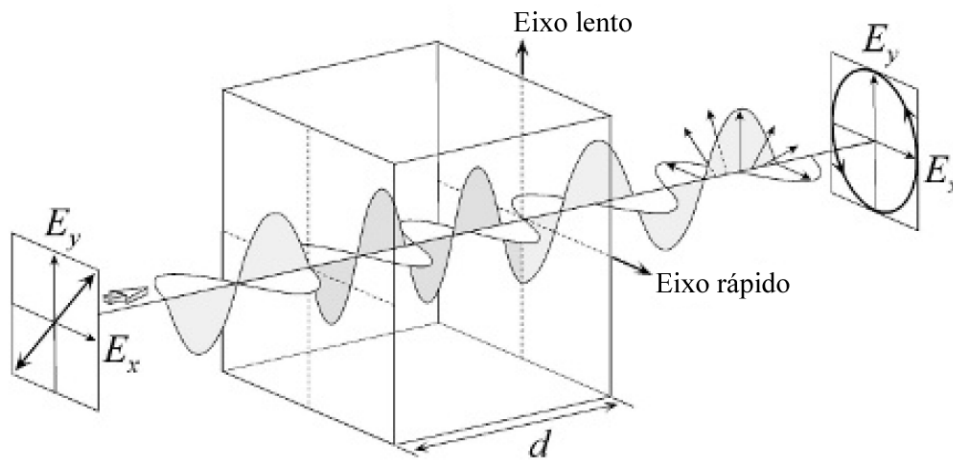


Figura 18 – Ação de um compensador ao transformar uma luz linearmente polarizada em circularmente polarizada. Adaptado de [18].

4.2.3 Modulador fotoelástico

A fotoelasticidade é um fenômeno que ocorre em materiais opticamente isotrópicos, e seu principal efeito é causar um grau de anisotropia nesses materiais. Ao aplicar tensão mecânica em um corpo, é possível alterar a densidade de elétrons no interior do material, gerando anisotropia com relação à distribuição dessas cargas.

A figura 19 mostra dois exemplos de moduladores fotoelásticos sendo o modulador da figura 19(b) uma modificação do modulador da figura 19(a). Os dois modelos de

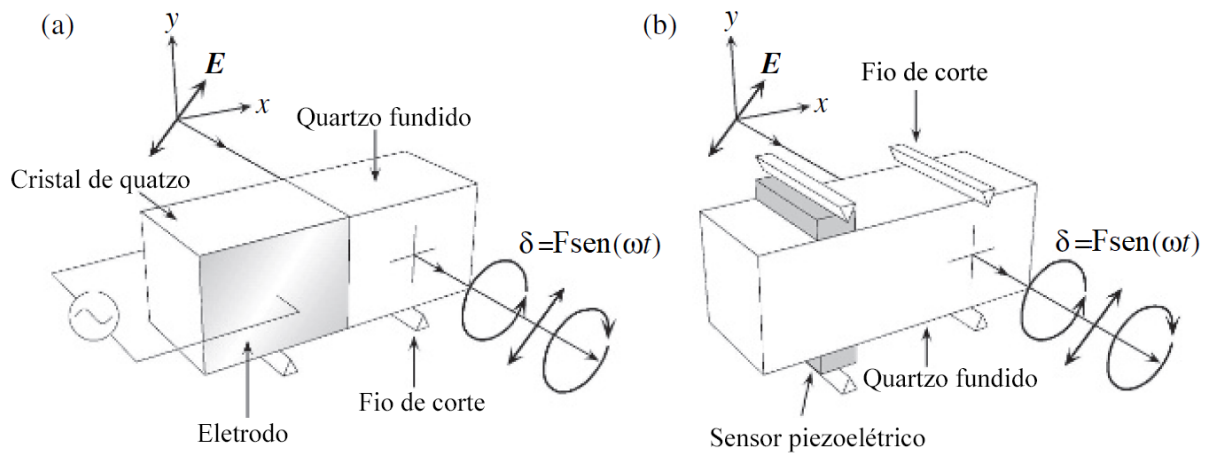


Figura 19 – Estrutura de um modulador fotoelástico em dois momentos: (a) representa os primeiros moduladores desenvolvidos e (b) é uma modificação mais moderna. Adaptado de [18].

moduladores são sensíveis à variação de temperatura, então o uso desses dispositivos demanda um ajuste rigoroso da temperatura para ter um controle preciso da diferença de fase δ . O modulador da figura 19(a) é montado usando um cristal de quartzo acoplado em quartzo fundido, e essa estrutura é apoiada sobre um fio de corte. O cristal de quartzo faz o papel de um sensor piezoelétrico [18]. Ao aplicar no cristal de quartzo um campo elétrico com frequência de 50kHz, que é a frequência ressonante desse material, uma tensão periódica é aplicada no quartzo fundido [18]. Para uma luz polarizada a 45° que entra no modulador fotoelástico, a diferença de fase relativa δ entre as componentes x e y do campo será gerada pelo efeito de fotoelástico. O valor de δ no tempo é expresso por

$$\delta = F \sin(\omega t) \quad (4.3)$$

em que $\omega = 2\pi v$ e $v = 50kHz$ [18]. Na expressão (4.3), a amplitude da fase F é proporcional ao fator $\frac{V}{\lambda}$, tal que V é a tensão elétrica aplicada no material e λ é o comprimento de onda do campo elétrico incidente [18]. Para a figura 19(b) a montagem é diferente e mais simples do que a anterior. Nessa abordagem alternativa, temos uma peça inteira de quartzo fundido colocada entre duas peças de sensores piezoelétricos. Um dos sensores é usado para aplicar uma tensão mecânica no quartzo fundido e o outro sensor é utilizado para monitorar esse valor aplicado usando a diferença de fase δ , que é expressa pela equação (4.3) para uma frequência de ressonância de 50kHz.

4.2.4 Despolarizador

O despolarizador é um elemento óptico usado para converter uma luz polarizada em não polarizada. Uma outra utilidade desse elemento é corrigir fontes não polarizadas que possam ter um leve grau de polarização para algumas componentes, eliminando

essa pequena parcela polarizada. A figura 20 mostra a estrutura de um despolarizador feito de um cristal birrefringente. Como esse cristal apresenta diferentes espessuras e é birrefringente, os raios de luz propagantes por diferentes espessuras do cristal terão fases distintas entre si. O resultado disso é que os raios de luz transmitidos pelo material terão diversas fases, ocasionando em uma luz não polarizada.

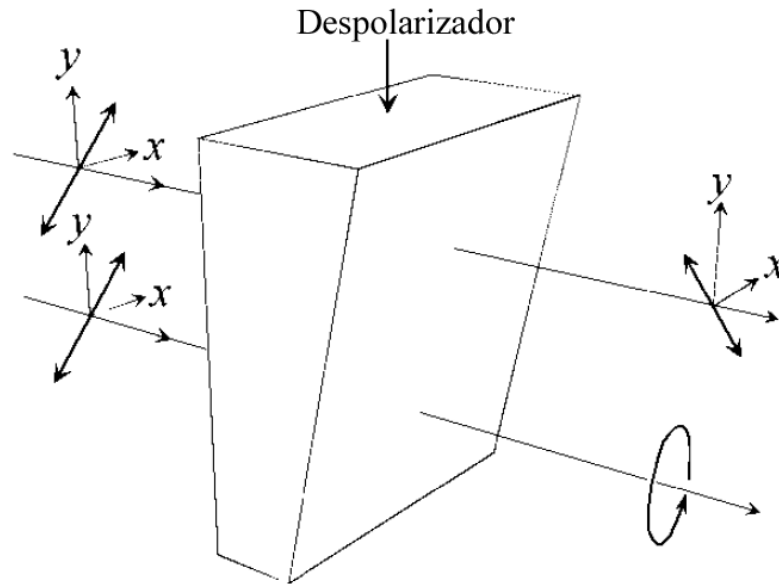


Figura 20 – Estrutura de um despolarizador. Adaptado de [18].

4.3 Cálculo de Jones

Podemos representar a luz polarizada e suas alterações ao escrevê-la como um vetor, em que as componentes desse vetor carregam todas as informações suficientes para expressar a onda eletromagnética desejada. Devido a essa forma vetorial, podemos então operar essa luz utilizando matrizes, que representam algum tipo de fenômeno, medição ou transformação aplicada a essa luz. Essas ferramentas matemáticas expostas são chamadas de vetores e matrizes de Jones, que fazem parte do cálculo de Jones para a luz polarizada.

4.3.1 Vetores de Jones

O estado de polarização da luz pode ser representado ao decompor uma onda eletromagnética em dois eixos perpendiculares x e y e descrever como os campos nessas componentes variam no tempo e espaço. O vetor de Jones é usado para fazer essa representação da luz polarizada. Para uma onda eletromagnética geral propagando em uma

direção $\hat{z}$ ao longo do tempo t , o vetor de Jones associado a ela é:

$$\vec{E}(z, t) = \begin{bmatrix} E_{x0} e^{i(\omega t - Kz + \delta_x)} \\ E_{y0} e^{i(\omega t - Kz + \delta_y)} \end{bmatrix} = e^{i(\omega t - Kz)} \begin{bmatrix} E_{x0} e^{i\delta_x} \\ E_{y0} e^{i\delta_y} \end{bmatrix}. \quad (4.4)$$

A equação (4.4) pode ser escrita de maneira mais compacta considerando a diferença de fase $(\delta_x - \delta_y)$ e omitindo $e^{i(\omega t - Kz)}$ [18]:

$$\vec{E}(z, t) \begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} E_{x0} e^{i(\delta_x - \delta_y)} \\ E_{y0} \end{bmatrix} = \begin{bmatrix} |E_x| e^{i(\delta_x - \delta_y)} \\ |E_y| \end{bmatrix} \quad (4.5)$$

Partindo dessas definições e da equação (3.21), podemos então escrever a intensidade da luz como:

$$I = I_x + I_y = E_{x0}^2 + E_{y0}^2 = |E_x|^2 + |E_y|^2 = E_x E_x^* + E_y E_y^*. \quad (4.6)$$

Um dos interesses da elipsometria está na mudança relativa da amplitude da luz e de sua fase, de forma que essas variações são suficiente para o uso da técnica [18]. Uma das consequências disso é a normalização dos valores de intensidade da luz, pois dada uma intensidade luminosa total no valor de I , todos os valores de intensidades medidos serão divididos por I , fazendo com que a intensidade total se torne $I = 1$. Essa é uma maneira conveniente, pois permite escrever os vetores de Jones de forma mais polida e sem perda de informações necessárias.

Como anteriormente foi apresentada uma seleção de estados possíveis para a polarização da luz, agora podemos usar o formalismo do vetor de Jones para representar esses estados. Uma polarização linear no eixo x e no eixo y pode ser representada, respectivamente, por [18]:

$$\vec{E}_{linear,x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \vec{E}_{linear,y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \quad (4.7)$$

A luz polarizada a 45° é escrita como: [18]

$$\vec{E}_{+45^\circ} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (4.8)$$

Para a luz polarizada circularmente para a direita ou para a esquerda, temos: respectivamente [18]

$$\vec{E}_D = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}, \quad \vec{E}_E = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}. \quad (4.9)$$

Usando as polarizações circulares podemos montar uma polarização elíptica se tivermos uma diferença de fase $(\delta_x - \delta_y) = \frac{7\pi}{4}$. A representação em vetor de Jones de uma polarização elíptica para essa diferença de fase é [18]:

$$\vec{E}_{el} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0.707 + i0.707 \\ i \end{bmatrix}. \quad (4.10)$$

4.3.2 Representação em matrizes de Jones dos elementos ópticos

Os elementos ópticos vistos anteriormente podem alterar a luz, significando que essa mudança deve também ser feita no vetor de Jones dessa onda. A modificação desse vetor deve ser feita por uma matriz, e portanto cada elemento óptico terá uma matriz de Jones associada. O processo matemático geral que transforma uma luz incidente $\vec{\gamma}$ em uma luz $\vec{\gamma}'$ por atuação de um elemento óptico, representado pela matriz $\tilde{Q}$, é dado por:

$$\tilde{Q}\vec{\gamma} = \vec{\gamma}'. \quad (4.11)$$

Esse cálculo matricial é eficiente, pois caso seja preciso usar mais de um elemento óptico para atuar em uma luz, basta aplicar as matrizes dos elementos em sequência ao lado esquerdo do vetor de Jones.

A matriz de Jones para um polarizador $\tilde{P}$ (analisador $\tilde{A}$) com o eixo de transmissão paralelo ao eixo x é dada por:

$$\tilde{P} = \tilde{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \quad (4.12)$$

Usando a equação (4.11), se um polarizador do tipo da expressão (4.12) é usado em uma luz com polarização em 45° , a luz que emerge do polarizador é dada por [18]

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (4.13)$$

Um compensador com eixo rápido paralelo ao eixo x pode ser representado por:

$$\tilde{C} = \begin{bmatrix} 1 & 0 \\ 0 & e^{(-i\delta)} \end{bmatrix}. \quad (4.14)$$

A atuação de um compensador com $\delta = \frac{\pi}{2}$ em uma luz com polarização em 45° é dada por:

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & e^{(-i\frac{\pi}{2})} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}. \quad (4.15)$$

Essa transformação mostra que a luz polarizada em 45° se torna circularmente polarizada para a esquerda ao interagir com esse compensador de $\delta = \frac{\pi}{2}$ [18].

Por fim, a matriz para o modulador fotoelástico é dada por:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & e^{(i\delta)} \end{bmatrix}. \quad (4.16)$$

4.3.3 Representação de medidas ópticas por matrizes de Jones

Considerando a interação de uma fonte luminosa com elementos ópticos, a luz que resulta desse processo pode ser medida e interpretada. A figura 21 mostra uma fonte

luminosa não polarizada incidindo em um conjunto de elementos, que juntos formam um instrumento óptico capaz de converter a luz para um estado de polarização linear. Abaixo de cada elemento há a representação correspondente em matrizes de Jones. O eixo de transmissão do polarizador é rotacionado por um ângulo α no plano xy , e o eixo de transmissão do analisador é paralelo ao eixo x . Logo, com relação à luz emitida pela fonte de luz, apenas a porção luminosa transmitida pelo polarizador (E_P) é considerada. Se escolhermos as coordenadas $x'y'$ de modo que o eixo de transmissão do polarizador seja paralelo ao eixo x' , este eixo é rotacionado por um ângulo α em relação ao eixo x . Neste caso, para realizar o cálculo matricial, é necessária uma rotação de coordenadas de $x'y'$ para xy . Essa rotação de coordenadas presente na figura 21 é expressa por $\tilde{R}(-\alpha)$, uma vez que a rotação das coordenadas é no sentido horário [18]. Se operarmos todas essas

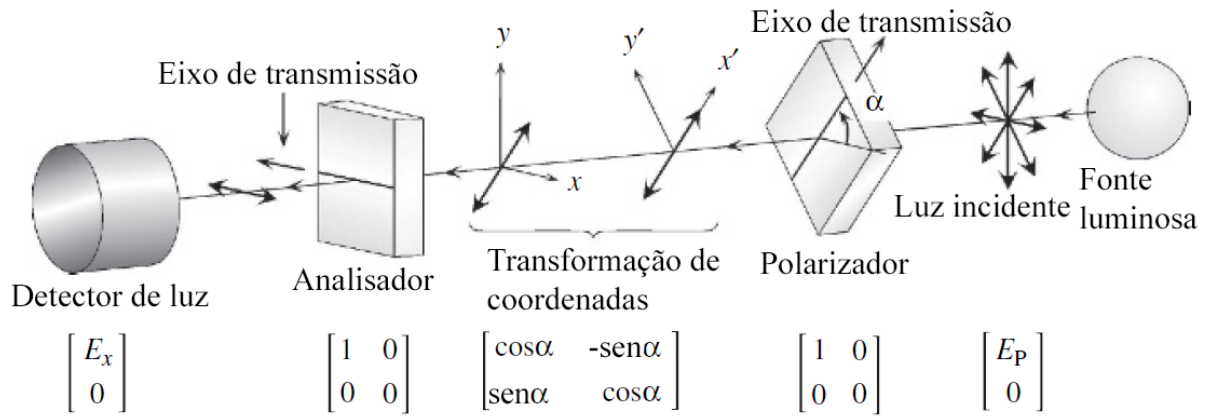


Figura 21 – Representação de um instrumento óptico usando matrizes de Jones. Adaptado de [18].

matrizes na luz que sai da fonte luminosa, as componentes E_x e E_y da luz que saem do instrumento óptico serão [18]:

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E_P \\ 0 \end{bmatrix}. \quad (4.17)$$

Pela figura 21, vemos que apenas a componente do eixo x restou, como era esperado pela argumentação fornecida anteriormente. Usando a equação (3.21), a intensidade da luz medida pelo detector seria:

$$I = |E_x|^2 = |E_P|^2 \cos^2 \alpha, \quad (4.18)$$

que é a equação da lei de Malus [18], mostrando que a intensidade da luz varia ao estabelecer um ângulo de rotação α entre dois polarizadores. No caso da figura 21, temos um polarizador que gira por um ângulo α em relação ao eixo x e um segundo analisador (segundo polarizador) que repousa em um ângulo zero em relação mesmo eixo. Podemos ver um exemplo da lei de Malus na figura 22, que mostra o gráfico da intensidade normalizada

de uma luz linearmente polarizada em função de um ângulo α de um polarizador. Quando o eixo de transmissão é paralelo à direção de polarização da luz ($\alpha = 0^\circ$), a onda luminosa é transmitida sem perda de intensidade. Para o eixo de transmissão sendo perpendicular à direção de polarização da luz ($\alpha = \frac{\pi}{2}$), a intensidade luminosa será totalmente perdida durante o processo.

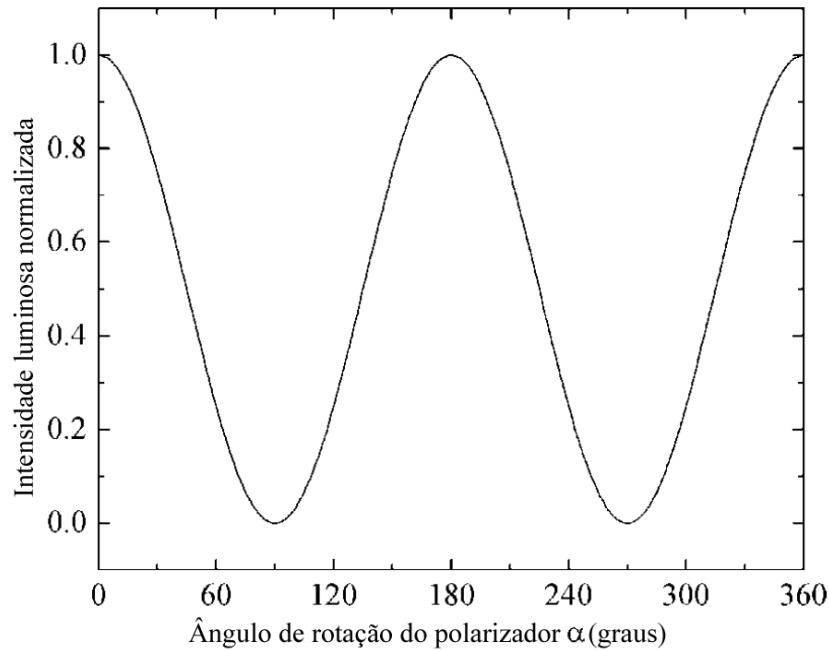


Figura 22 – Intensidade da luz na lei de Malus em função do ângulo α de um polarizador. Adaptado de [18].

4.4 Cálculo de Mueller

O formalismo de Jones apresentado anteriormente é adequado para o tratamento de luzes com um estado polarizado bem definido. Porém, como também há um interesse em manipular a luz de fontes não polarizadas ou com sobreposição de estados de polarização, é preciso apresentar um formalismo que seja capaz de representar bem os fenômenos que envolvem a luz com essas características. Para isso, apresentaremos o cálculo de Mueller e os parâmetros de Stokes, que são as ferramentas matemáticas necessárias para fazer esse tratamento de fontes luminosas não polarizadas e também para representar os elementos ópticos e amostras que vão interagir com essas luzes.

4.4.1 Parâmetros de Stokes

Os parâmetros de Stokes são ferramentas úteis para descrever a luz não polarizada em forma de vetores. Esses parâmetros podem ser apresentados de várias maneiras, mas

aqui eles serão utilizados em função da intensidade luminosa da luz não polarizada. Um vetor luminoso $\vec{S}$ de uma fonte pode ser caracterizado por quatro parâmetros como:

$$\vec{S} = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} I_x + I_y \\ I_x - I_y \\ I_{+45^\circ} - I_{-45^\circ} \\ I_R - I_S \end{bmatrix}. \quad (4.19)$$

Na equação (4.19), temos o vetor de Stokes $\vec{S}$ formado a partir do conjunto $\{S_n\}$, que são os parâmetros de Stokes. S_0 representa a intensidade total da luz, S_1 caracteriza a subtração da intensidade da polarização na direção y pela intensidade da direção x . O parâmetro S_2 reproduz a subtração da intensidade da polarização em $+45^\circ$ pela intensidade em -45° , enquanto S_3 descreve a subtração da intensidade da polarização para direita pela intensidade para esquerda. A figura 23 apresenta uma representação gráfica dos parâmetros de Stokes baseados na intensidade da luz.

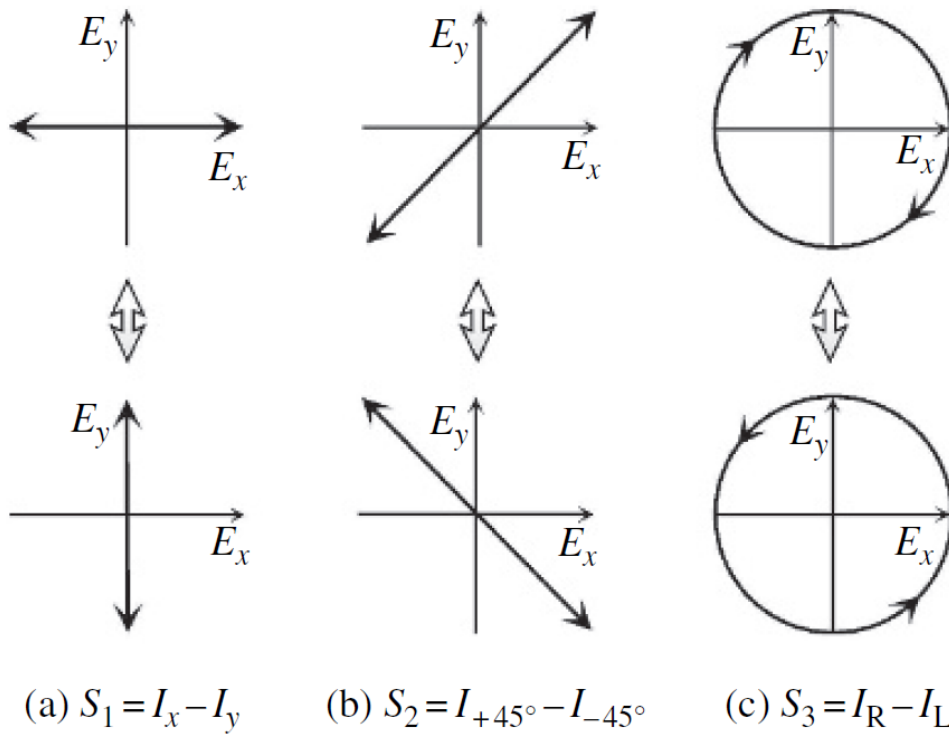


Figura 23 – Representação gráfica dos parâmetros de Stokes pela intensidade da luz: parâmetros para (a) polarizações em x e y , (b) $+45^\circ$ e -45° e (c) direita e esquerda. Adaptado de [18].

Os parâmetros de Stokes também podem ser representados usando os valores de campos elétricos. Essa representação é particularmente útil, pois os campos têm uma grande importância nas medidas de elipsometria. Podemos descrever S_0 e S_1 desse modo como [18]:

$$S_0 = I_x + I_y = E_{x0}^2 + E_{y0}^2 = E_x E_x^* + E_y E_y^* \quad (4.20)$$

$$S_1 = I_x - I_y = E_{x0}^2 - E_{y0}^2 = E_x E_x^* - E_y E_y^*. \quad (4.21)$$

Para expressar S_2 em termos dos campos elétricos, devemos utilizar a fase relativa $\Delta = \delta_x - \delta_y$, e então temos:

$$S_2 = 2E_{x0}E_{y0}\cos\Delta. \quad (4.22)$$

De forma similar, podemos escrever S_3 como [18]:

$$S_3 = -2E_{x0}E_{y0}\sin\Delta. \quad (4.23)$$

Esses estados de polarização apresentados podem ser visualizados a partir de uma interpretação geométrica chamada de Esfera de Poincaré, que está ilustrada na figura 24. Ao tomar o conjunto $\{S_n\}$ dos quatro parâmetros de Stokes, podemos usar o parâmetro S_0 para definir o raio da esfera e os três parâmetros restantes serão utilizados como três eixos para formar um sistema de coordenadas, em que o centro da esfera se encontra na origem desse sistema. Se tomarmos uma analogia entre a esfera de Poincaré com o globo terrestre, os eixos que estão no plano da linha do equador representam as polarizações lineares, enquanto os polos descrevem as polarizações circulares [18]. No eixo S_1 temos uma polarização linear em x para $S_1 > 0$ e uma polarização em y quando $S_1 < 0$. Para o eixo S_2 a polarização é linear em $+45^\circ$ e -45° para $S_2 > 0$ e $S_2 < 0$ respectivamente. No polo norte da esfera temos a polarização circular para a direita, e no sul ela é circular para a esquerda.

Uma luz totalmente polarizada é um ponto $P(S_1, S_2, S_3)$ nessa esfera de tal forma que:

$$S_0^2 = S_1^2 + S_2^2 + S_3^2, \quad (4.24)$$

e se S_0 for uma intensidade normalizada, temos:

$$S_1^2 + S_2^2 + S_3^2 = 1. \quad (4.25)$$

Para a equação (4.25), a esfera de Poincaré tem raio unitário. Para os outros casos em que a polarização não é total, vale a relação [18]:

$$S_0^2 > S_1^2 + S_2^2 + S_3^2. \quad (4.26)$$

Considerando todos os pontos distribuídos ao longo da esfera de Poincaré, um ponto aleatório que esteja no interior ou na superfície da esfera representa um estado de polarização, que pode ser caracterizado por qualquer estado de polarização canônico ou por uma mistura entre os estados canônicos. Esse ponto na esfera pode ser representado por dois conjunto de parâmetros (ϵ, θ) e (ψ, Δ) conforme ilustra a figura 25 para uma polarização elíptica representada nesses dois sistemas de coordenadas. Na figura 25 (a), os eixos maiores e menores com tamanhos $2a$ e $2b$, respectivamente, são usados como

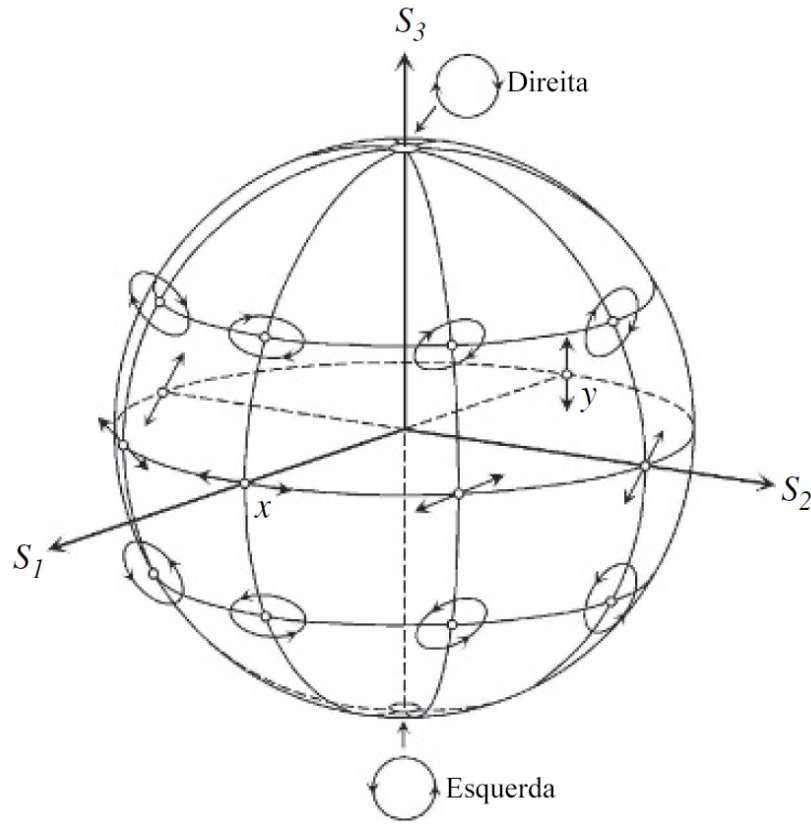


Figura 24 – Interpretação dos estados de polarização por meio esfera de Poincaré. Adaptado de [18].

coordenadas para escrever a polarização elíptica [18]. O ângulo azimutal θ é formado entre o maior eixo da elipse e o eixo E_x , enquanto o ϵ é o ângulo elíptico tal que $\epsilon = \tan \frac{b}{a}$. Nesse caso, quando $\epsilon = 0$, a polarização resultante será linear [18].

O sistema de coordenadas (θ, ϵ) usa dois ângulos como parâmetros para descrever um estado de polarização. Logo, um ponto $P(S_1, S_2, S_3)$ na esfera de Poincaré pode ser descrito como [18]:

$$S_1 = \cos 2\epsilon \cos 2\theta \quad (4.27)$$

$$S_2 = \cos 2\epsilon \sin 2\theta \quad (4.28)$$

$$S_3 = \sin 2\epsilon. \quad (4.29)$$

O uso desse sistema de coordenadas é mais compreensivo ao examinar a figura 26, que localiza um ponto P a partir dos ângulos θ e ϵ .

O sistema de coordenadas (ψ, Δ) é utilizado a partir das componentes (E_x, E_y) [18]. Embora ψ possa ter uma compreensão um tanto nebulosa em um primeiro contato, esse parâmetro representa um ângulo que relaciona as amplitudes das componentes do campo

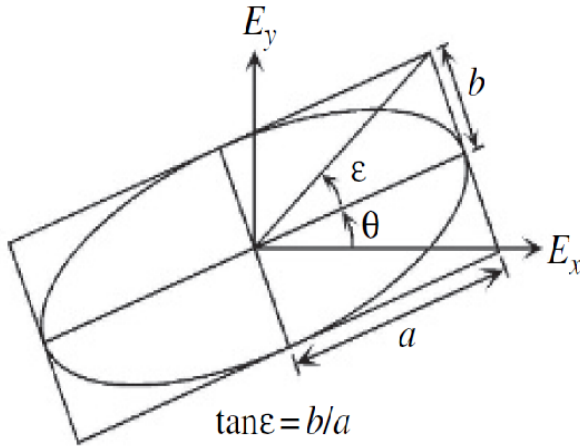
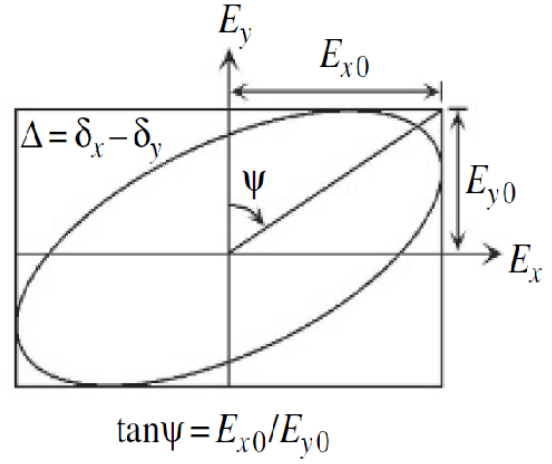
(a) Sistema de coordenadas (ϵ, θ) (b) Sistema de coordenadas (ψ, Δ) 

Figura 25 – Representação da polarização elíptica nos dois sistemas de coordenadas (ϵ, θ) e (ψ, Δ) . Adaptado de [18].

elétrico nos eixos x e y tal que:

$$\tan(\psi) = \frac{E_{x0}}{E_{y0}}. \quad (4.30)$$

O parâmetro Δ representa a diferença de fase entre essas duas componentes da onda:

$$\Delta = \delta_x - \delta_y. \quad (4.31)$$

Partindo então das equações (4.30) e (4.31), podemos usar a figura 25(b) e concluir que [18]:

$$E_{x0} = \sin \Psi, \quad E_{y0} = \cos \psi. \quad (4.32)$$

Podemos então retomar o vetor de Jones e reescrevê-lo a partir de (Ψ, Δ) como [18]:

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} E_{x0} e^{i\Delta} \\ E_{y0} \end{bmatrix} = \begin{bmatrix} \sin \Psi e^{i\Delta} \\ \cos \Psi \end{bmatrix} = \cos \Psi \begin{bmatrix} \tan \Psi e^{i\Delta} \\ 1 \end{bmatrix}. \quad (4.33)$$

É útil também representar os três últimos parâmetros de Stokes em função de Ψ e Δ . Dessa forma, os parâmetros se tornam

$$S_1 = \sin^2 \Psi - \cos^2 \Psi = -\cos 2\Psi \quad (4.34a)$$

$$S_2 = 2 \sin \Psi \cos \Psi \cos \Delta = \sin 2\Psi \cos \Delta \quad (4.34b)$$

$$S_3 = -2 \sin \Psi \cos \Psi \sin \Delta = -\sin 2\Psi \sin \Delta. \quad (4.34c)$$

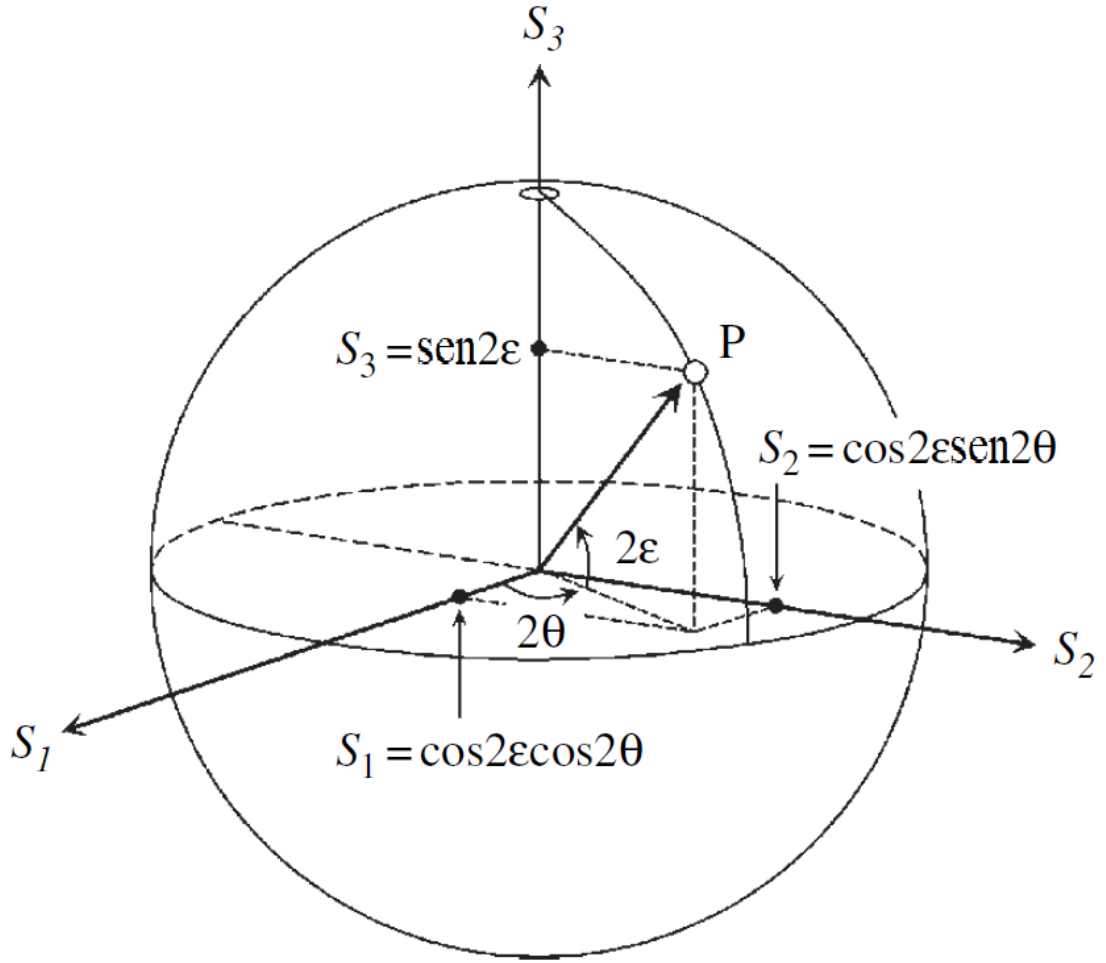


Figura 26 – Representação de um ponto P na esfera de Poincaré utilizando as coordenadas θ e ϕ . Adaptado de [18].

A esfera de Poincaré fornece uma interpretação forte para o entendimento da luz polarizada, parcialmente polarizada e não polarizada. Vamos então propor um uso direto para essas ferramentas na interpretação de medidas luminosas. Suponha um experimento que envolva um detector com interface (tela digital) e três fontes de luz, e o objetivo do experimento é determinar os estados de polarização das três fontes. A partir das informações do detector, a interface é capaz de indicar o ponto P na esfera de Poincaré que representa o estado de polarização dos feixes de luz que chegam ao detector continuamente ao longo do tempo. Então, o estado de polarização para cada fonte será dado pela média dos pontos que foram indicados na esfera de Poincaré. Por fim, vamos considerar que a intensidade S_0 está normalizada ($S_0 = 1$).

Para a primeira fonte, a interface mostrou continuamente o mesmo ponto P na superfície da esfera, que está ilustrado na figura 27(a). Isso indica que essa fonte emite uma luz totalmente polarizada. A segunda fonte provoca na interface uma distribuição

de pontos (área cinza) concentrada em uma região da esfera com um ponto P no centro dessa distribuição, que representa o ponto médio da distribuição. Esse comportamento é o da luz parcialmente polarizada e está representado na figura 27(b). No caso da terceira fonte, a interface mostra pontos espalhados por toda a esfera de Poicaré, deixando a superfície da esfera completamente sombreada pela incidência de pontos. Isso caracteriza o perfil de uma fonte não polarizada e está mostrado na figura 27(c). Como essa situação contém pontos espalhados por toda a esfera, a média desses pontos resulta em um ponto $P = (S_1, S_2, S_3) = (0, 0, 0)$. Portanto, uma luz não polarizada é representada pelo ponto que da origem ao sistemas de coordenadas, que nesse caso coincide com o centro da esfera de Poincaré.

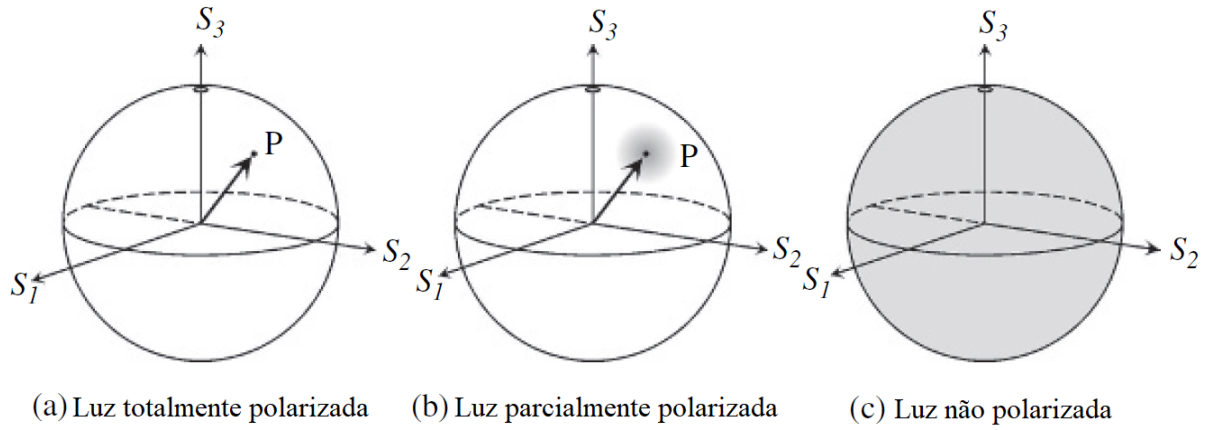


Figura 27 – Representações na esfera de Poincaré de: (a) luz totalmente polarizada, (b) luz parcialmente polarizada e (c) luz não polarizada. Adaptado de [18].

É possível avaliar o quão polarizada uma luz é a partir do grau de polarização p que ela possui. Esse grau é definido como [18]:

$$p = \frac{(S_1^2 + S_2^2 + S_3^2)^{1/2}}{S_0}. \quad (4.35)$$

Usando as equações (4.24) e (4.26) na expressão (4.35) para $S_0 = 1$, temos que o grau de polarização para a luz totalmente polarizada é $p = 1$, e para a luz parcialmente polarizada vale $p < 1$. Se a luz é não polarizada, temos $p = 0$, que é um grau nulo de polarização.

4.5 Matriz de Mueller

A expressão (4.19) representa o vetor de Stokes para a luz de uma fonte luminosa. Durante processos experimentais é comum que a luz tenha interação com superfícies e meios que possam alterar o estado de polarização dessa onda luminosa. Dessa forma, um vetor de Stokes $\vec{S}_i$ é convertido em um outro vetor $\vec{S}_o$ usando

$$\vec{S}_o = \tilde{M}\vec{S}_i, \quad (4.36)$$

em que $\tilde{M}$ é chamada de matriz de Mueller (MM), que foi desenvolvida pelo físico Hans-Mueller (1900-1965) [20]. Essa matriz descreve um sistema óptico (ou, no nosso caso deste trabalho, uma amostra de interesse) em termos de quantidades experimentalmente observáveis e tem o papel de converter uma luz inicial $\vec{S}_i$ (luz de entrada) em uma luz de saída $\vec{S}_o$ [20]. A matriz de Mueller tem uma forma 4×4 e é frequentemente representada como:

$$\tilde{M} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix}. \quad (4.37)$$

Usando as equações (4.19), (4.36) e (4.37), podemos expressar a atuação da matriz de Mueller como:

$$\begin{bmatrix} S'_0 \\ S'_1 \\ S'_2 \\ S'_3 \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix} \cdot \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}. \quad (4.38)$$

A equação (4.38) pode ser reescrita usando os parâmetros de Stokes em função das intensidades da equação (4.19):

$$\begin{bmatrix} I'_x + I'_y \\ I'_x - I'_y \\ I'_{+45^\circ} - I'_{-45^\circ} \\ I'_R - I'_S \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix} \cdot \begin{bmatrix} I_x + I_y \\ I_x - I_y \\ I_{+45^\circ} - I_{-45^\circ} \\ I_R - I_S \end{bmatrix}. \quad (4.39)$$

É comum usar os vetores de Stokes e a matriz de Mueller em suas formas normalizadas. Portanto, podemos reescrever a expressão (4.38) multiplicando $\tilde{M}$ por $\frac{M_{11}}{M_{11}}$ e $\vec{S}$ por $\frac{S_0}{S_0}$:

$$\begin{bmatrix} S'_0 \\ S'_1 \\ S'_2 \\ S'_3 \end{bmatrix} = M_{11} \begin{bmatrix} 1 & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \cdot S_0 \begin{bmatrix} 1 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix}, \quad (4.40)$$

em que $m_{ij} = \frac{M_{ij}}{M_{11}}$ e $s_i = \frac{S_i}{S_0}$. Então, os elementos da matriz de Mueller podem ser referidos como M_{ij} , MM_{ij} , m_{ij} ou mm_{ij} , em que as duas últimas maneiras representam os elementos em sua forma normalizada. Ao dividirmos por S_0 na equação (4.40), assumimos que S_0 é diferente de zero. Essa consideração não é apenas para evitar uma indeterminação matemática, mas há também um significado físico. O parâmetro S_0 está relacionado com a intensidade total da luz, então, se ele for nulo, estaríamos operando com um feixe luminoso de intensidade zero, resultando em um absurdo nesse contexto.

4.5.1 Seções da matriz de Mueller

A matriz de Mueller pode ser usada para representar uma amostra que vai interagir com uma luz incidente $\vec{S}_i$ para convertê-la em uma luz de saída $\vec{S}_o$ após a interação entre o feixe e a amostra. Como amostras diferentes tendem a interagir com a luz de maneiras distintas, é esperado que essas amostras sejam representadas por matrizes de Mueller que não sejam iguais entre si. Dessa forma, é útil entender cada elemento da matriz de Mueller, pois isso fornece meios para construir a matriz de qualquer amostra.

Começamos demarcando a matriz de Mueller em quatro seções como está ilustrado na figura 28: atenuação não polarizada (cinza), diatenuação (vermelho), polarização (azul) e retardamento (amarelo).

$$\begin{bmatrix} S_0' \\ S_1' \\ S_2' \\ S_3' \end{bmatrix}_o = \begin{matrix} & \begin{bmatrix} 1 & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \\ M_{11} & \end{matrix} \cdot \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}_i$$

Figura 28 – Matriz de Mueller dividida em seções para atenuação de luz (cinza), diatenuação (vermelho), polarização (azul) e retardamento (amarelo). Adaptado de [20].

Para uma luz não polarizada, M_{11} (seção cinza) tem relação direta com a atenuação da intensidade da luz de saída. Para um feixe completamente não polarizado, o vetor de Stokes desse feixe só não é nulo em S_0 . Dessa forma, temos:

$$\begin{bmatrix} S_0' \\ 0 \\ 0 \\ 0 \end{bmatrix} = M_{11} \begin{bmatrix} 1 & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \cdot \begin{bmatrix} S_0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (4.41)$$

Partindo da equação (4.41), obtemos:

$$S_0' = M_{11} S_0. \quad (4.42)$$

Caso a luz seja polarizada, M_{11} não será suficiente para encontrar a intensidade S_0' da luz de saída [20], pois neste caso as outras componentes do vetor de Stokes não serão mais

nulas. Partindo disso, teremos:

$$S'_0 = M_{11}S_0(1 + m_{12}S_1 + m_{13}S_2 + m_{14}S_3). \quad (4.43)$$

Por essa situação, percebemos que a primeira linha da matriz de Mueller é importante na determinação da intensidade total da luz de saída. Isso acontece porque cada elemento da primeira linha interage com o vetor de Stokes da luz de entrada de forma que as intensidades de cada tipo de polarização serão contabilizadas na intensidade final da luz de saída. Com isso, acabamos de fornecer uma interpretação para a seção vermelha da matriz de Mueller, em que os elementos m_{12} , m_{13} e m_{14} estão relacionados, respectivamente, com a atenuação de intensidade da polarização orientada em (x, y) , $(45^\circ, -45^\circ)$ e circular (direita, esquerda).

A primeira coluna da matriz de Mueller (seção azul) armazena a capacidade de uma amostra em polarizar uma luz não polarizada [20]. Para exemplificar isso, supomos que um vetor luminoso totalmente não polarizado e normalizado $\vec{S} = (1, 0, 0, 0)$ interage com uma amostra representada por uma matriz $\tilde{M}$ com a primeira coluna não nula. Podemos representar essa interação usando a expressão (4.40) como

$$\begin{bmatrix} S'_0 \\ S'_1 \\ S'_2 \\ S'_3 \end{bmatrix} = M_{11} \begin{bmatrix} 1 & * & * & * \\ m_{21} & * & * & * \\ m_{31} & * & * & * \\ m_{41} & * & * & * \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (4.44)$$

Pela equação (4.44), podemos notar que ainda que a luz de entrada tenha apenas o parâmetro S_0 não nulo (luz não polarizada), é possível que a luz de saída adquira certa polarização ao interagir com a amostra desde que a primeira coluna tenha pelo menos um elemento não nulo entre m_{21} , m_{31} e m_{41} . Esses três elementos estão relacionados, respectivamente, com a obtenção de uma polarização orientada em (x, y) , $(45^\circ, -45^\circ)$ e circular. Dessas forma, dependendo dos valores desses elementos em $\tilde{M}$, a luz pode ser totalmente polarizada, parcialmente polarizada ou não polarizada [20].

A seção amarela é chamada de retardamento e abriga os últimos 9 elementos do canto inferior direito da matriz de Mueller [20]. Os elementos da diagonal principal (m_{22} , m_{33} , m_{44}) relacionam um tipo específico de polarização na luz de entrada com esse mesmo tipo na luz de saída. Esses três elementos da diagonal principal estão relacionados, respectivamente, com os tipos de polarização: linear em (x, y) , $(45^\circ, -45^\circ)$ e circular. Se um desses elementos da diagonal principal tiver valor de 1 ou -1 , o que vai acontecer, respectivamente, é que a luz de saída preservará um determinado estado de polarização ou então terá esse estado transformado em sua contraparte ortogonal [20]. Por exemplo, se considerarmos $m_{22} = 1$, a polarização linear para os eixos x e y não será convertida em outro tipo de polarização, mas as fases e amplitudes das polarização nesses eixos podem ser alteradas [20]. De forma geral, se um elemento m_{ii} da diagonal tiver valor de ± 1 , são

nulos os elementos restantes da seção de retardamento que fazem parte da linha e coluna em que esse elemento m_{ii} está localizado. Tomando o caso em que $m_{22} = 1$, temos então uma matriz:

$$\hat{M} = \begin{bmatrix} 1 & m_{12} & m_{13} & m_{14} \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & m_{33} & m_{34} \\ m_{41} & 0 & m_{43} & m_{44} \end{bmatrix}. \quad (4.45)$$

Ao preencher todos os elementos da diagonal principal por 1 na matriz de Mueller, zerando o restante e considerando $M_{11} = 1$, teremos o caso em que a luz de entrada é igual à luz de saída:

$$\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}. \quad (4.46)$$

Esse é um caso intuitivo, pois a expressão (4.46) envolve a matriz Identidade operando um vetor coluna, resultando no mesmo vetor ao fim da operação.

4.5.2 Medições na matriz de Mueller

Considerando uma amostra alvo, montar a matriz de Mueller que representa essa amostra exige que todos os elementos dessa matriz sejam medidos. Para que isso aconteça, é preciso analisar cada estado de polarização da luz de saída com relação a sua amplitude. Podemos então entender que o processo geral para montar a MM envolve decompor a luz de saída em vários estados de polarização, medir a amplitude de cada estado e tratar essas medidas ao associá-las com os elementos da matriz de Mueller que tem relação com cada estado específico de polarização, e esse processo também é válido para o tratamento da luz não polarizada. Para entender esse processo, precisaremos usar algumas noções experimentais que envolvem o elipsômetro, que é um aparelho capaz de fazer medições de propriedades ópticas de amostras. Embora esse aparato experimental seja usado a partir de agora para explicações, os detalhes da técnica de elipsometria espectroscópica serão mais aprofundados no capítulo a seguir.

Os elementos ópticos em um elipsômetro são capazes de emitir, em uma amostra, luzes com estados de polarização específicos a partir de um instrumento chamado de gerador de estado polarizado (GEP). Após essa luz interagir com uma amostra, ela vai em direção a um conjunto de elementos ópticos conhecido como analisador de estado polarizado (AEP), que perscruta informações sobre qual é o estado de polarização que está presente na luz após a interação com a amostra [20]. Dessa forma, após a amostra ter contato com vários tipos de polarizações incidentes, é possível mapear um comportamento de como essa amostra altera ou não cada uma dessas polarizações incidentes nas polarizações de saída. Esses elementos da MM podem ser medidos por meio de um elipsômetro.

Um elipsômetro capaz de medir a matriz de Mueller modula continuamente um conjunto de elementos ópticos, e a informação detectada é um sinal modulado [20]. A análise de Fourier desse sinal permite a extração da MM de qualquer amostra desconhecida. Como são consideradas razões entre componentes dessa análise de Fourier, a intensidade absoluta da luz que incide e é detectada não é um fator importante para a precisão das medições [20]. Esse é um detalhe importante no uso do elipsômetro, pois são detectadas as mudanças relativas na polarização em vez de intensidades absolutas, como seria o caso em um refletômetro ou polarímetro. Isso possibilita medições de MM em amostras que dispersam parte da luz, amostras deformadas ou amostras menores do que o diâmetro do feixe de luz de sondagem [20]. Porém, nem todo elipsômetro é capaz de medir a matriz de Mueller completamente, pois essa capacidade depende dos elementos ópticos que constituem o aparelho. A figura 29 mostra alguns tipos de elipsômetros e quais os elementos da MM cada um tem acesso para medir [20].

Um polarizador fixo só produziria ou detectaria um tipo de polarização, mas ao girar o polarizador há uma mudança na intensidade relativa entre os elementos S_1 e S_2 , permitindo outras medições. Isso acaba determinando bem os elementos específicos da matriz de Mueller de interesse. Dessa forma, rotacionar um polarizador permite a medição dos nove elementos no canto superior esquerdo da MM, como mostrado na figura 29(A), que ilustra as configurações de um elipsômetro com polarizador ou analisador rotativo. Como o polarizador não é capaz de gerar luz circularmente polarizada, S_3 é sempre igual a 0, impedindo de acessar todas as informações da MM relacionadas à polarização circular. Portanto, um instrumento que usa apenas polarizadores no GEP ou AEP não consegue medir a última linha e coluna da MM [20].

Para medir a última linha ou coluna da MM, podemos adicionar um compensador ao GEP ou AEP, pois ele altera a fase de um campo elétrico com relação à componente ortogonal, o que pode preencher o parâmetro S_3 do vetor de Stokes [20]. Ao girar o compensador, diferentes valores de S_3 podem ser gerados na luz de interesse e analisados, tal que um elipsômetro com compensador rotativo é frequentemente utilizado para determinar propriedades adicionais da MM. Como mostrado nas figuras 29(B) e 29(C), um compensador rotativo antes da amostra permite acessar as três primeiras linhas da MM, enquanto um compensador rotativo colocado após a amostra pode medir as três primeiras colunas da MM. Se os compensadores rotativos forem colocados tanto antes quanto depois da amostra, como mostrado na figura 29(D), é possível medir toda a MM [20].

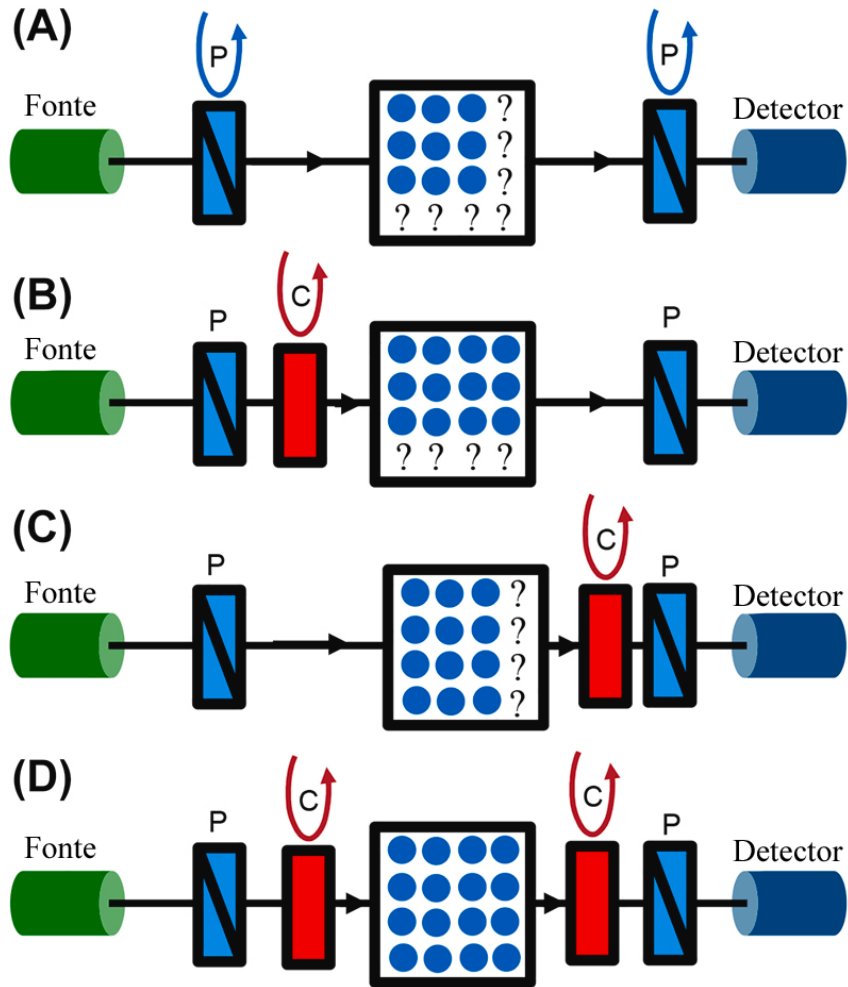


Figura 29 – Alguns tipos comuns de elipsômetros e os elementos correspondentes da matriz de Mueller que eles podem acessar. (A) Elipsômetros com analisador rotativo e polarizador rotativo não conseguem determinar a última linha e a última coluna da MM, mas podem medir os nove elementos no canto superior esquerdo. Elipsômetros com compensador rotativo podem medir (B) as três primeiras linhas ou (C) as três primeiras colunas da MM, dependendo se o compensador rotativo está antes ou depois da amostra, respectivamente. (D) Se os compensadores rotativos estiverem tanto antes quanto depois da amostra, toda a MM pode ser medida. Adaptado de [20].

5 Elipsometria espectroscópica

Anteriormente, foi abordado como a constante dielétrica é uma boa medida para captar características da matéria por meio da incidência de luz. Essa interação entre luz e matéria foi formalizada por meio do cálculo de Jones e Mueller. Então, já que construímos uma base teórica para o entendimento da função dielétrica e da sua relação com a luz e a matéria, é preciso agora propor uma técnica que seja capaz de obter medidas que sejam sensíveis à constante dielétrica. A elipsometria espectroscópica é uma técnica óptica com aplicação robusta no estudo da caracterização de materiais e com vasta aplicação na análise de filmes finos. A elipsometria mede a mudança de polarização da luz após interagir com uma amostra, e suas medidas experimentais são sensíveis à constante dielétrica.

A figura 30 ilustra o processo padrão de medição utilizando elipsometria, tal que uma luz incide sobre uma amostra desejada e essa onda pode sofrer processos de reflexão e transmissão ao interagir com o material. Essa técnica de espectroscopia é fundamentada na comparação de parâmetros ópticos de um feixe de luz antes e depois de incidir sobre a superfície do material analisado, sendo a polarização uma propriedade fundamental em todo o processo. A elipsometria pode ser usada para ângulos e comprimentos de ondas distintos de forma a montar um perfil espectral de uma amostra para cada ângulo e dentro de uma faixa do espectro eletromagnético. Dessa forma, uma grande quantidade de informações é coletada para melhor caracterizar a amostra. Na medição de elipsometria, frequentemente assumimos que a luz refletida por uma amostra é totalmente polarizada. No entanto, se a dispersão da luz ocorre na superfície rugosa da amostra, a luz incidente totalmente polarizada é frequentemente transformada em luz parcialmente polarizada [18].

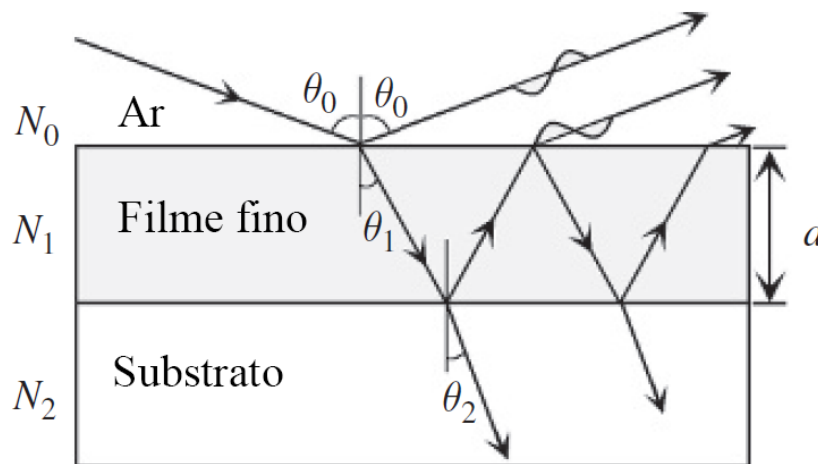


Figura 30 – Modelo óptico constituído de ar, filme fino e substrato. Adaptado de [18].

Quando a luz incidente sofre o fenômeno de reflexão pela amostra, a onda refletida tem características que podem ser consideravelmente diferentes daquelas presentes na incidente. Dessa forma, essas diferenças podem ser utilizadas em uma tentativa de coletar informações da amostra alvo.

Neste capítulo, exploraremos a técnica da elipsometria espectroscópica expondo e discutindo fundamentos e principais características. Usaremos o cálculo de Jones e Muller, que foram vistos anteriormente, para fornecer ferramentas que auxiliem no entendimento da elipsometria.

5.1 Medições de elipsometria

A coleta de valores usando a elipsometria acontece, de forma padrão, ao incidir uma luz com ângulo de Brewster em uma amostra e depois medir a porção que foi refletida. Como essa luz pode ser desmembrada em duas componentes de polarizações -p e -s, haverá polarização associada a cada uma delas. Então, quando essas componentes forem refletidas ou transmitidas pela amostra, poderão sofrer mudanças nos estados de polarização, e são essas informações que são usadas para medir constantes ópticas e a espessura da amostra. Embora o ângulo de Brewster seja frequentemente usado para coletar medidas mais sensíveis das razões de amplitudes para as polarizações -p e -s [18], outros ângulos também podem ser usados conforme as necessidades experimentais. A figura 31 ilustra um esquema com as características principais de uma medição por elipsometria. Uma luz linearmente polarizada em $+45^\circ$ de amplitude E , representada por sua decomposição em duas componentes perpendiculares $\vec{E}_{ip}$ e $\vec{E}_{is}$, incide sobre uma amostra com um ângulo θ e é refletida com as componentes $\vec{E}_{rp}$ e $\vec{E}_{rs}$. Podemos usar as componentes de incidência e de reflexão para encontrar os valores de (Ψ, Δ) , que são as principais medidas da elipsometria. Esses valores estão relacionados, respectivamente, com a razão das amplitudes e a diferença de fase entre as polarizações -p e -s [18]. Essas duas medidas são comumente expressas em conjunto na forma [18]:

$$\rho = \tan \Psi e^{i\Delta} = \frac{r_p}{r_s} = \frac{\left(\frac{E_{rp}}{E_{ip}}\right)}{\left(\frac{E_{rs}}{E_{is}}\right)}, \quad (5.1)$$

em que r_p e r_s são os coeficientes de reflexão para os planos p e s respectivamente. No caso de uma situação que envolva uma transmissão, teremos as componentes transmitidas nos planos s e p com amplitudes E_{ts} e E_{tp} . Nesse caso, a equação (5.1) é adaptada para o fenômeno de transmissão como:

$$\rho = \tan \Psi e^{i\Delta} = \frac{t_p}{t_s} = \frac{\left(\frac{E_{tp}}{E_{ip}}\right)}{\left(\frac{E_{ts}}{E_{is}}\right)}, \quad (5.2)$$

tal que t_p e t_s são, respectivamente, os coeficientes de transmissão para os planos p e s.

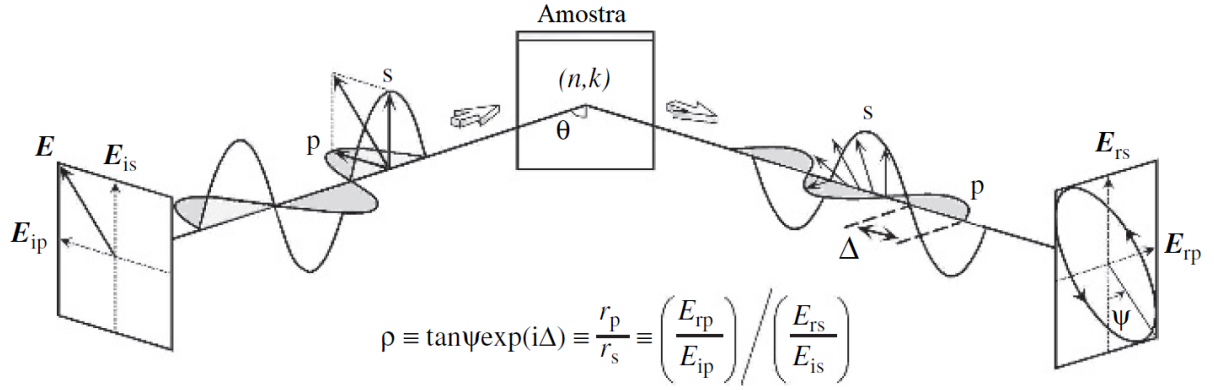


Figura 31 – Esquema e princípios de medida para elipsometria. Adaptado de [18].

Podemos escrever os parâmetros Ψ e Δ separadamente em formas que podem ser úteis para coletar esses valores. Para o caso de Ψ , podemos utilizar [18]

$$\tan \Psi = \frac{|r_p|}{|r_s|}. \quad (5.3)$$

Tomando a equação (5.1) e (5.3), podemos escrever:

$$\Psi = \tan^{-1}(|\rho|) = \tan^{-1}\left(\frac{r_p}{r_s}\right). \quad (5.4)$$

Podemos formular Δ como [18]

$$\Delta = \delta_{rp} - \delta_{rs}. \quad (5.5)$$

Nesse caso, para obtermos Δ a partir de ρ é preciso observar a parte real e imaginária de ρ na equação (5.1). Dessa forma, teremos as seguintes condições para Δ [18]:

$$\Delta = \arg(\rho) = \tan^{-1}\left(\frac{\text{Im}(\rho)}{\text{Re}(\rho)}\right) \text{ para } \text{Re}(\rho) > 0, \quad (5.6a)$$

$$\Delta = \arg(\rho) = \tan^{-1}\left(\frac{\text{Im}(\rho)}{\text{Re}(\rho)}\right) + 180^\circ \text{ para } \text{Re}(\rho) < 0 \text{ e } \text{Im}(\rho) \geq 0, \quad (5.6b)$$

$$\Delta = \arg(\rho) = \tan^{-1}\left(\frac{\text{Im}(\rho)}{\text{Re}(\rho)}\right) - 180^\circ \text{ para } \text{Re}(\rho) < 0 \text{ e } \text{Im}(\rho) < 0. \quad (5.6c)$$

As medições de Ψ e Δ certamente são as principais no contexto da elipsometria, porém há outras medidas que podem ser feitas e que também fornecem informações importantes sobre uma amostra. Um outro conjunto de medidas são os elementos da matriz de Mueller da amostra. Com um elipsômetro adequado, é possível medir os 16

elementos da matriz de Mueller normalizada, representados pelo conjunto $\{m_{ij}\}$. O valor M_{11} que multiplica a matriz de Mueller normalizada da expressão geral (4.40) é associado com a intensidade total da luz não polarizada refletida ou transmitida [20], que também pode ser chamada de uR.

Um outro conjunto de medidas que pode ser coletado pela elipsometria espectroscópica é: AnE, Aps, Asp e despolarização (dPolE). Para compreender os três primeiros parâmetros vamos considerar uma matriz de Jones $\tilde{J}$ para uma amostra anisotrópica [21]:

$$\tilde{J} = \begin{bmatrix} r_{pp} & r_{sp} \\ r_{ps} & r_{ss} \end{bmatrix}. \quad (5.7)$$

Os elementos de matriz da expressão (5.7) são definidos a partir de [21]:

$$AnE = \tan(\Psi)e^{i\Delta} = \frac{r_{pp}}{r_{ss}}, \quad (5.8a)$$

$$Aps = \tan(\Psi_{ps})e^{i\Delta_{ps}} = \frac{r_{ps}}{r_{pp}}, \quad (5.8b)$$

$$Asp = \tan(\Psi_{sp})e^{i\Delta_{sp}} = \frac{r_{sp}}{r_{ss}}. \quad (5.8c)$$

Nesse conjunto de elementos $\{r_{ij}\}$, cada um deles descreve, por meio dos índices i e j , como uma polarização- i se transforma em uma polarização- j . Como exemplo, temos que r_{ps} caracteriza como a polarização-p é convertida para uma polarização-s [22].

Por fim, a despolarização (dPolE) é uma grandeza ligada à polarização P de um sistema, tal que a despolarização é usualmente relatada como [21]:

$$dPolE = (1 - P^2).100\%, \quad (5.9)$$

em que a equação (5.9) indica a despolarização em sua forma percentual em função da polarização P . Resta agora escrever P em uma forma que seja acessível por meio das medidas de elipsometria, e então usaremos os elementos da matriz de Mueller para isso. Uma amostra isotrópica que apresenta despolarização possui uma matriz de Mueller $\tilde{M}$ da forma [21]:

$$\tilde{M} = \begin{bmatrix} 1 & -P.N & 0 & 0 \\ -P.N & P & 0 & 0 \\ 0 & 0 & P.C & P.S \\ 0 & 0 & -P.S & P.C \end{bmatrix}, \quad (5.10)$$

em que a relação de P com os elementos da expressão (5.10) é dada por [21]:

$$P = \sqrt{N^2 + C^2 + S^2}. \quad (5.11)$$

Os valores (N, C, S) da expressão (5.11) são escritos em função dos parâmetros elipsométricos (Ψ, Δ) [20]:

$$N = \cos(2\Psi) \quad (5.12a)$$

$$C = \sin(2\Psi) \cos(\Delta) \quad (5.12b)$$

$$S = \sin(2\Psi) \sin(\Delta). \quad (5.12c)$$

Dessa forma, substituindo a expressão (5.12) em na equação (5.11) e depois na expressão (5.9), temos que a despolarização pode ser escrita como:

$$dPolE = \{1 - [(\cos 2\Psi)^2 + (\sin 2\Psi \cos \Delta)^2 + (\sin 2\Psi \sin \Delta)^2]\}.100\%. \quad (5.13)$$

5.2 Considerações gerais sobre elipsometria

Como foi exposto neste capítulo, a elipsometria espectroscópica é uma técnica com alto desempenho para coletar informações sobre uma amostra e, de forma geral, é utilizada em filmes finos. As medidas de elipsometria permitem a obtenção de um perfil das amostras de modo que os dados ali coletados representam uma caracterização óptica dessa amostra. Com isso, a elipsometria se apresenta também como uma técnica com alta capacidade de aplicação em diversas áreas do conhecimento, como a criação de modelos ópticos que possam ser utilizados em contexto industrial.

No contexto da ciência pura, a elipsometria espectroscópica é uma ferramenta poderosa no processo de compreender e fomentar o estudo de propriedades ópticas da matéria. Os formalismos de Jones e Mueller, juntamente com os parâmetros de Stokes, são amplamente utilizados para entender e expandir as fronteiras do conhecimento dentro da óptica. Então, fica exposto que a elipsometria tem um grande papel dentro da física experimental e com capacidade grande de fornecer dados para a elaboração de modelos.

Ao abordarmos ciência aplicada, a elipsometria tem grande valor pela quantidade de dados que é capaz de fornecer ao combinar a técnica com interfaces, computadores e equipamentos. Particularmente, interessa a este trabalho essa grande quantidade e variedade de dados que a elipsometria consegue coletar de amostras em intervalos de tempo consideravelmente curtos. Isso é uma característica interessante, pois grandes volumes de dados podem ser empregados em técnicas de aprendizado de máquina e construção de inteligências artificiais.

6 Inteligência artificial e aprendizado de máquina

A inteligência artificial é um campo próspero com muitas aplicações práticas e vastas pesquisas na busca por modelos inteligentes para automatizar o trabalho de rotina, entender fala ou imagens, fazer diagnósticos médicos e dar suporte à pesquisa científica básica [1]. Partindo disso, uma das utilidades da inteligência artificial e aprendizado de máquina é usar os conceitos e modelos dessas áreas para resolver problemas relacionados aos seres humanos e à natureza [23].

Algumas das atuações do aprendizado de máquina envolvem criar modelos para fazer classificações ou regressões, e cada uma dessas duas modalidades possui uma aplicação específica. A classificação é uma tarefa em que um modelo deve receber informações e retornar uma resposta dentro de uma faixa discreta, em que cada valor de resposta possível é uma classe. Para uma tarefa de regressão, o modelo recebe os dados e é condicionado a fornecer respostas que podem habitar uma faixa contínua de valores. Para este trabalho, teremos um enfoque em tarefas de classificação.

A capacidade de processar dados dos computadores atuais permite fazer uso de algoritmos de aprendizado de máquina para classificar pacientes saudáveis ou reativos para uma determinada doença sem a necessidade de modelar fisicamente as amostras de modo direto. Esses algoritmos são capazes de reconhecer padrões nos dados e, a partir disso, fazer tarefas de classificação conforme a demanda do pesquisador desde que os dados usados para treino e teste do algoritmo sejam qualificados para representar bem o problema alvo.

Apesar do uso de aprendizado de máquina ser uma proposta tentadora e com grande potencial, é preciso saber como estruturar e avaliar os modelos criados. A estruturação de um modelo pressupõe a seleção de um algoritmo e a escolha de seus hiperparâmetros, que podem ser coletados por meio de técnicas de otimização com objetivo de construir um modelo de classificação robusto. Já a avaliação dos modelos é feita a fim de verificar se o desempenho deles é satisfatório o suficiente para realizar a tarefa para a qual foram criados. Para tal, é preciso estabelecer técnicas de validação dos modelos que levem em consideração as etapas de treino e teste de uma inteligência artificial.

Portanto, a base teórica a ser construída para este trabalho deve abordar questões fundamentais sobre como o aprendizado de máquina é possível de ser feito e como ele pode ser estruturado a partir dos diferentes algoritmos disponíveis. Além disso, é fundamental discutir as maneiras de avaliar a generalidade que uma inteligência artificial pode ter quando ela avalia dados que não foram apresentados a ela previamente durante o treino.

6.1 Princípios de aprendizagem de máquina

Os dados desempenham um papel crucial no aprendizado de máquina, de forma que é importante usar terminologias e representações que forneçam um bom entendimento para caracterizá-los [24], e isso também é igualmente válido para outros conceitos e entidades dentro dessa área. Considerando que os processos algorítmicos de aprendizado de máquina possuem um forte rigor matemático que fundamentam as suas funcionalidades, é preciso construir uma linguagem matemática que seja eficiente em fazer essa comunicação. A partir disso, podemos utilizar esse formalismo matemático para representar conceitos e objetos fundamentais no campo do aprendizado de máquina.

Dados e informações são entidades importantes na área de aprendizado de máquina e inteligência artificial, então é mais do que razoável que eles tenham uma representação matemática eficiente. Os algoritmos de aprendizado de máquina são alimentados por informações comumente chamadas de dados de entrada ou instâncias. Um único dado de entrada pode ser representado por um vetor $\vec{X}$ de m componentes ($m \in \mathbb{N}^*$) na forma $\vec{X} = (x_1, x_2, \dots, x_m)$. Essas componentes são chamadas de atributos (*features*), que são informações que esse vetor de entrada carrega. As *features* podem representar qualquer tipo de informação que se ache relevante para estudar um problema. Tomando um exemplo, se queremos analisar uma situação que envolve circuitos elétricos, podemos construir vetores $\vec{X} = (x_1, x_2, x_3)$ em que cada uma das três *features* representa, respectivamente, a tensão, resistência e a corrente elétrica no circuito, e os valores de x_1, x_2 e x_3 são as quantidades que o vetor armazena para cada *feature*.

Como os algoritmos são operados usando vários vetores de entrada, podemos representar todo um conjunto de vetores de entrada como sendo $\{\vec{X}_i\}$ em que o dado de entrada $\vec{X}_i$ é um vetor rotulado pelo índice $i \in \mathbb{N}^*$ dentro desse conjunto (mais a seguir o índice i será trocado por u para não causar confusões de notação). Portanto, a partir de agora, sempre que evocarmos $\vec{X}_i$ significa que estamos tomando como exemplo um vetor de instância i e com índice m para suas componentes, que pode ser escrito de maneira geral como $\vec{X}_i = (x_{1i}, x_{2i}, \dots, x_{mi})$.

Vamos agora analisar um exemplo que demonstre como esses dados de entrada podem ser utilizados. Partindo de uma idealização, propomos um problema em que o conjunto de entradas $\{\vec{X}_i\}$ está diretamente relacionado com um conjunto de saídas $\{Y_i\}$ sem qualquer tipo de erro, tal que um elemento Y_i desse conjunto pode ser um escalar ou um vetor, mas aqui usaremos Y_i como sendo um escalar. Dessa forma, acabamos de estabelecer uma relação entre uma entrada $\vec{X}_i$ com uma saída Y_i , então daremos um passo adiante ao supor que existe uma função de forma que [1, 24]:

$$F(\vec{X}) = Y. \quad (6.1)$$

No contexto de problemas de classificação, os valores em $\{Y_i\}$ podem ser chamados

de rótulos de $\{\vec{X}_i\}$. Em tarefas de classificação binária é comum que $\{Y_i\}$ seja um conjunto com n elementos, mas que só podem assumir dois valores, que geralmente são 0 e 1 ou -1 e 1. Esses dois valores que os elementos em $\{Y_i\}$ podem assumir são chamados de classes. Então, se para um dado de entrada $\vec{X}_i$ na expressão (6.1) temos $F(\vec{X}_i) = 1$, significa que a classe associada ao dado $\vec{X}_i$ é 1.

Antes de prosseguir, precisamos fazer um breve aposto sobre dois tipos diferentes de aprendizado que um algoritmo pode ter: supervisionado e não supervisionado. O aprendizado supervisionado relaciona um conjunto de entradas $\{\vec{X}_i\}$ às saídas $\{Y_i\}$, que podem ser contínuas ou discretas. O exemplo que foi fornecido anteriormente é de um aprendizado supervisionado, pois para um conjunto de dados $\{\vec{X}_i\}$ nós estamos atribuindo as classes $\{Y_i\}$ por meio de uma função $F(\vec{X})$. Para o aprendizado não supervisionado nós tentamos agrupar $\{\vec{X}_i\}$ sem atribuir diretamente a eles valores de saída. Nesse tipo de aprendizagem os algoritmos podem tratar os dados de entrada como pontos em um espaço e agrupar esses dados usando a distância entre eles e também no número de agrupamentos (classes) desejado. Boa parte do material abordado nesta escrita será direcionado ao aprendizado supervisionado, mas enunciar essas distinções é importante para prosseguir com as discussões de maneira apropriada.

Se temos um algoritmo em mãos para processar as entradas $\vec{X}_i$, é desejável que ele consiga operar essas entradas da mesma maneira que a expressão (6.1) e consiga relacionar corretamente cada elemento de $\{\vec{X}_i\}$ com cada valor em $\{Y_i\}$. O problema é que nós não temos acesso à essa função $F(\vec{X})$, denominada de função alvo [1]. O que pode ser feito é treinar um algoritmo com as entradas $\{\vec{X}_i\}$ para que ele alcance uma função $f(\vec{X})$ que seja a mais próxima possível de $F(\vec{X})$. E aqui fazemos uma distinção: enquanto $F(\vec{X})$ é a função alvo, $f(\vec{X})$ é uma função alcançada a partir do treinamento de um algoritmo. Então, podemos representar $f(\vec{X})$ como:

$$f(\vec{X}_i) = y_i, \quad (6.2)$$

em que cada vetor de $\{\vec{X}_i\}$ está ligado a uma saída $\{y_i\}$. Vale pontuar que como $F(\vec{X})$ e $f(\vec{X})$ não são necessariamente iguais, a relação $Y_i = y_i$ para os conjuntos $\{Y_i\}$ e $\{y_i\}$ não é obrigatoriamente verdadeira. Isso significa que um algoritmo que é definido por $f(\vec{X})$ pode ter um resultado $f(\vec{X}_i) = y_i$ para $y_i \neq Y_i$ mesmo que idealmente $\vec{X}_i$ esteja associado à saída Y_i no problema estudado. É claro que quanto mais parecidas forem as funções $F(\vec{X})$ e $f(\vec{X})$ melhor será o desempenho do algoritmo. Se para um dado de entrada $\vec{X}_i$ na equação (6.2) temos $f(\vec{X}_i) = 1$ e a saída (rótulo) de $\vec{X}_i$ é de fato 1, então o modelo foi capaz de acertar corretamente o dado de entrada. Porém, se $f(\vec{X}_i) = 1$ e o rótulo de $\vec{X}_i$ for 0, então o modelo não foi capaz de acertar o rótulo desse vetor de entrada.

Até o momento, utilizamos a função $f(\vec{X})$, que retorna apenas 0 ou 1 como saída, mas, em princípio, existe uma função $f'(\vec{X})$ que não trabalha com saídas binárias, mas sim com valores de probabilidade. Essa função retorna a probabilidade de um dado de

entrada ser da classe 1, então para uma entrada $\vec{X}_i$ temos $0 \leq f'(\vec{X}) \leq 1$, em que quanto mais próximo de 1 maior a probabilidade do dado de entrada pertencer a classe 1, e a maior proximidade de 0 fornece uma maior chance de ser da classe 0. Os valores entre 0 e 1 que precisam se encaixar em alguma dessas classes, e para solucionarmos isso usamos um *threshold* (limiar), que é um número entre 0 e 1 que decide a partir de qual valor um dado será classificado como 1. Como exemplo, supomos que para uma função $f'(\vec{X})$ observamos que um bom *threshold* é $th = 0.3$. Então, toda vez que $f'(\vec{X})$ atingir valores maiores ou iguais a 0.3 eles serão classificados como 1, e do contrário serão classificados como 0. É comum que o *threshold* seja 0.5 por padrão, mas essa não é uma condição necessária. Podemos então escrever a função $f(\vec{X})$ a partir de $f'(\vec{X})$ e de um *threshold* geral (th) como:

$$f(\vec{X}) = \begin{cases} 0, & \text{se } f'(\vec{X}) < th \\ 1, & \text{se } f'(\vec{X}) \leq th. \end{cases} \quad (6.3)$$

Essa função $f(\vec{X})$ habita um espaço chamado de conjunto hipóteses ($\mathcal{H}$) [1]. Esse espaço abriga todas as funções $f(\vec{X})$ que um algoritmo pode assumir para fazer as classificações dos dados de entrada $\{\vec{X}_i\}$. Esse espaço pode conter funções simples ou complexas, e isso será discutido posteriormente ao abordarmos a capacidade de um algoritmo e de um modelo.

A função $f(\vec{X})$ da expressão (6.2) é construída por meio dos dados usados para treinar um algoritmo, pois ao ser exposto a um conjunto de dados o algoritmo reconhece os padrões deles e faz um ajuste que gera essa função $f(\vec{X})$ para classificar esses dados com a maior precisão possível [1]. A partir desse ponto em que um algoritmo está completamente treinado e com uma função $f(\vec{X})$ estabelecida, temos um modelo pronto. Embora em certos contextos os termos algoritmos e modelos sejam usados como sinônimos, vamos partir do princípio de que modelo é a representação de uma função que foi obtida a partir de um treinamento usando um conjunto de dados, enquanto o algoritmo é o processo usado para gerar esse aprendizado [24].

Todo algoritmo pode ter sua estrutura alterada por meio da modificação dos seus hiperparâmetros, mas a mudança desses valores não implica que o algoritmo mudará a maneira de processar os dados. Para ilustrar isso, suponhamos uma função do primeiro grau $h(x) = ax + b$, tal que a é o coeficiente angular e b é um termo livre. Se realizarmos uma troca de a por c tal que $a \neq c$, a nova função $j(x) = cx + b$ não deixa de ser do primeiro grau, ainda que tenha tido o coeficiente alterado. Para evitar ambiguidades, vamos diferenciar neste trabalho o termo 'hiperparâmetro' de 'parâmetro'. Usaremos hiperparâmetros para referenciar valores relacionados aos algoritmos, enquanto parâmetros são associados às medidas experimentais que utilizamos neste trabalho. Os hiperparâmetros serão mais explorados na seção de algoritmos.

6.1.1 Treino e teste do modelo: o aprendizado é possível?

Embora tenha sido abordado que um algoritmo pode ser treinado a partir de um conjunto de dados, devemos dar um passo adiante e questionar: é possível que um algoritmo aprenda a partir do treino? Se quisermos ser ainda mais criteriosos, aprofundamos a pergunta para: o que é 'aprender' para uma máquina? Essas são perguntas corajosas e necessárias para fundamentar qualquer discussão sobre esse campo.

O treinamento é uma etapa essencial para a construção de um modelo de inteligência artificial, e isso foi apresentado de forma breve anteriormente. Uma outra etapa de grande importância é testar esse modelo, posto que a comparação entre métricas de desempenho para o treino e para o teste podem fornecer informações importantes sobre o modelo produzido e também a auxiliar na escolha de uma inteligência artificial com boa capacidade de generalização.

Para compreender melhor o aprendizado, supomos um conjunto de entradas $\{\vec{X}_i\}$ tal que podemos dividir esses dados em dois grupos: dados de treino $\{\vec{X}_i\}_{treino}$ e teste $\{\vec{X}_i\}_{teste}$. Se tomarmos um algoritmo, cada vez que ele processa os dados de treino ele completa uma época. Logo, devemos deixar que o algoritmo processe os dados $\{\vec{X}_i\}_{treino}$ por várias épocas para tentar compreender os padrões nesses dados. O desempenho de um algoritmo pode ser medido usando várias métricas, e elas serão abordadas com mais detalhes em outra seção, mas aqui iremos usar apenas a taxa de erro dos modelos. Como estamos usando dados de treino e teste, iremos trabalhar com dois valores de erros referentes a cada um desses grupos de dados.

O erro empírico é o percentual de erro que a IA comete ao tentar classificar os dados de treino. Após ser treinada com $\{\vec{X}_i\}_{treino}$, o modelo é posto para classificar esses dados a fim de verificar se ele comete baixos erros. O erro empírico é definido por [1]:

$$E_i = \frac{n}{N}, \quad (6.4)$$

em que n é o número de erros cometidos pelo modelo e N é a quantidade total de entradas classificadas. Para o grupo de teste, o erro esperado é o percentual de erro que o modelo comete dentro dos dados de teste $\{\vec{X}_i\}_{teste}$. Podemos definir o erro esperado como [1]:

$$E_o = \frac{m}{M}, \quad (6.5)$$

tal que, dentro do conjunto de teste, m é o número de erros cometidos pelo modelo e M é a quantidade total de dados classificados. Os erros empíricos e esperados podem fornecer interpretações sobre o modelo analisado, mas o erro esperado contém uma característica que vale a pena discutir, pois esse erro é coletado a partir de dados que o modelo não viu durante o treinamento. Então, podemos inferir que o erro esperado é uma aproximação para o desempenho real do modelo. De forma geral, o erro empírico costuma ser mais baixo do que o erro esperado, pois o modelo tende a classificar mais corretamente os dados

usados no treino do que aqueles que ele nunca processou. A depender do contexto, o erro empírico pode ser chamado de erro de treino, já o erro esperado pode ser referido como erro de teste ou erro de validação. Esses erros cometidos por um modelo são calculados a partir de uma função de perda $L(\{\vec{X}_i\}, \{Y_i\}, \{y_i\}, W)$, em que $\{\vec{X}_i\}$ são as entradas, $\{Y_i\}$ são as saídas verdadeiras, $\{y_i\}$ são as saídas previstas pelo modelo e W são variáveis do algoritmo que sejam importantes no cálculo desse erro. A função de perda pode assumir muitas formas a depender de como se queira interpretar o erro de um algoritmo. Então, uma das aspirações ao se trabalhar com aprendizado de máquina é a minimização da função de erro para selecionar modelos com desempenhos satisfatórios.

Para os propósitos que estamos abordando, um modelo adequado é aquele que alcança a seguinte condição: os erros empíricos e esperados são baixos e convergem entre si. A figura 32 apresenta a convergência dos erros empírico (E_i) e esperado (E_o) em função do número de dados visto por um algoritmo para um espaço de busca simples e outro complexo. Para da figura 32a) o erro empírico é grande, isso significa que o modelo não está sendo capaz de ter bons resultados nem mesmo para os dados de treino, então o desempenho nos dados de teste também tende a ser baixo. Nessa situação, o modelo tem uma convergência entre erro empírico e esperado, mas os dois são altos, então não é um modelo interessante para se usar. Além disso, como o espaço de hipóteses $\mathcal{H}$ alcançado pelo modelo contém funções que são simples demais para se parecer com a função alvo $F(\vec{X})$, o modelo tende a convergir para erros altos mesmo com um incremento de dados no processo de treino. No caso da figura 32b) temos os dois erros convergindo para um valor mais baixo, porém para que essa convergência seja satisfatória é necessário um número maior de dados, pois o espaço de hipóteses $\mathcal{H}$ contém funções complexas, demandando mais dados para alcançá-las a partir do treino. Essa situação é adequada à condição de um bom modelo que propomos anteriormente, então a função alcançada é próxima da função alvo.

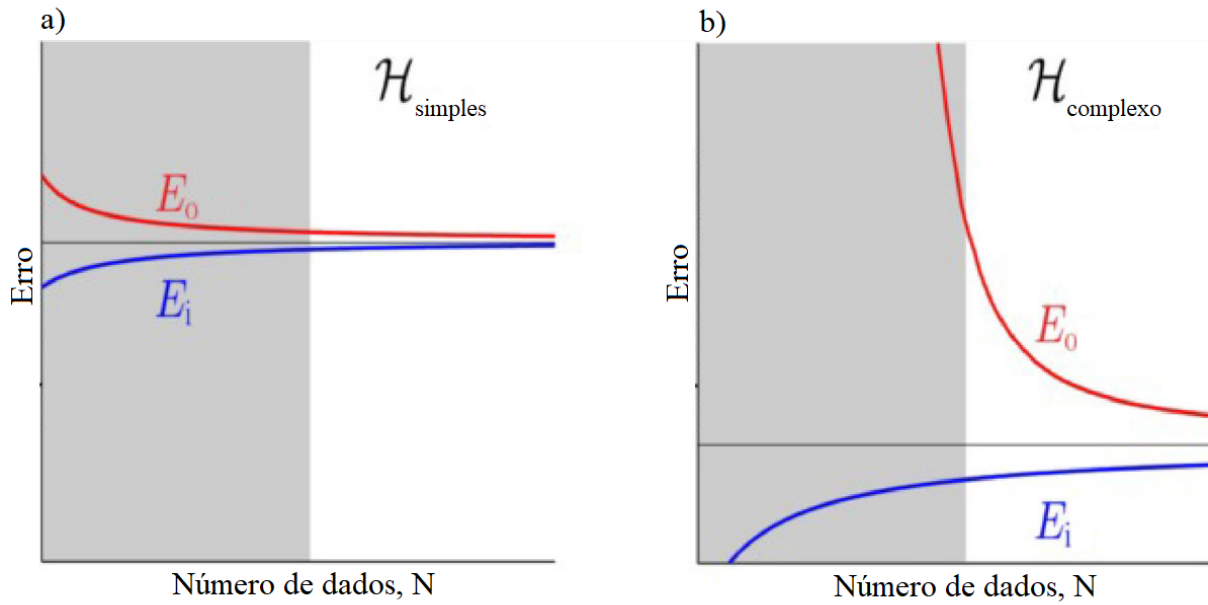


Figura 32 – Convergência dos erros empírico (E_i) e esperado (E_o) para dois espaços de busca diferentes. Para um espaço de busca a) simples a convergência dos erros ocorre para um valor alto e poucos dados, e em um espaço de busca b) complexo os erros convergem para valores menores, mas demandam mais dados para isso. Adaptado de [25].

Podemos agora retomar a discussão sobre o que seria o aprendizado para um modelo, embora o conceito seja vasto e possível de ser abordado de maneiras distintas. Para este trabalho vamos definir aprendizado como sendo a capacidade de um modelo em generalizar para os dados de teste o que foi processado por ele a partir dos dados de treino. Então, um modelo com bom aprendizado é aquele que, a partir dos dados de treino, consegue generalizar o aprendizado e obter bons resultados de classificação ou regressão para os dados fora do grupo de treino.

6.1.2 Capacidade dos modelo

Se queremos resolver um problema utilizando aprendizado de máquina, é interessante que o modelo produzido consiga mapear corretamente as entradas $\{\vec{X}_i\}$ em seus rótulos $\{Y_i\}$. Porém, para que isso aconteça é preciso ter condições para criar um modelo que seja eficiente nessa tarefa. Denominamos de capacidade o grau de competência que um modelo tem na resolução de um problema [1], e alcançar uma boa capacidade demanda alguns detalhes na construção do modelo.

Um problema de classificação pode ser caracterizado como simples ou complexo, mas em ambos os casos solucionar um problema exige uma quantidade de dados de entrada

que seja satisfatória para construir um modelo com capacidade suficiente para executar tal tarefa. Se o problema for simples, significa que existe uma função alvo $F(\vec{X})$ simples que resolve esse problema. No caso de tarefas complexas o modelo precisa de uma função que também é complexa para realizar essa tarefa de forma satisfatória. Para tarefas simples ou complexas não temos acesso direto a essa função $F(\vec{X})$ que resolve perfeitamente o problema, então o modelo pode ser representado por qualquer função no conjunto hipótese $\mathcal{H} \equiv \{f_i(\vec{X})\}$, que contém uma coleção de funções que o modelo pode representar. A existência de $\mathcal{H}$ levanta uma questão: se para um dado problema há mais de uma função que o resolve bem, qual devemos escolher? Nesse caso, valemos da navalha de Ockham: para duas soluções que pareçam igualmente boas, escolha a mais simples [1]. Aqui é importante ratificar que a solução mais simples pode ainda ser uma função de alta complexidade.

A capacidade de um modelo é uma característica que deve ser bem administrada para garantir um desempenho satisfatório, pois uma capacidade inadequada pode levar o modelo a cometer muitos erros. Então, existe um equilíbrio delicado e considerável sobre o nível de capacidade que um modelo deve ter. A figura 33 mostra as curvas de erro no teste e treino em função da capacidade do modelo. Pode ser observado que as duas curvas têm um decaimento até chegar em um certo valor de capacidade, mas após esse valor o erro de teste começa a aumentar enquanto o de treino permanece em queda. Esse ponto, que está demarcado por uma linha vertical preta tracejada, é chamado de capacidade ótima de um modelo. Um algoritmo que recebe um conjunto dados de treino está condicionado a assumir alguma função $f(\vec{X})$ permitida pelo espaço de hipótese $\mathcal{H}$. A função ótima dentro do espaço de hipótese não é necessariamente a com maior capacidade, pois o ponto de capacidade ótima é aquele em que um algoritmo consegue aproveitar ao máximo a generalização dos dados de treino e então é capaz de ter o menor erro de teste possível para um conjunto de dados. Porém, aumentar a capacidade do modelo ótimo não faz com que o erro de teste continue a cair, posto que o aumento da capacidade leva a funções $f(\vec{X})$ que podem ser tão complexas a ponto de se ajustarem demais aos dados de treino, que pode proporcionar um desempenho ruim ao classificar os dados de teste. A função ótima de um espaço de hipótese $\mathcal{H}$ não é necessariamente a função alvo $F(\vec{X})$, pois é possível que esse espaço de hipóteses não contenha a função alvo $F(\vec{X})$. Uma das razões para um espaço $\mathcal{H}$ não conter a função alvo é devido a uma quantidade inadequada de dados de treino. Porém, como foi dito anteriormente, o objetivo do aprendizado de máquina não é alcançar exatamente a função alvo, mas sim funções que sejam satisfatoriamente parecidas com ela. A figura 34 apresenta duas situações a) e b) em que cada uma tem um espaço de hipóteses $\mathcal{H}$ (área em azul escuro) e um subespaço (azul claro) contido no espaço de hipóteses que contém as funções que o modelo pode assumir a partir de diferentes dados de treino. Em 34a) temos um espaço de hipótese que contém a função alvo $F(\vec{X})$, que significa que é possível construir modelos com funções parecidas com $F(\vec{X})$. Para 34b) temos um espaço de hipótese que não contém o modelo alvo, então há uma

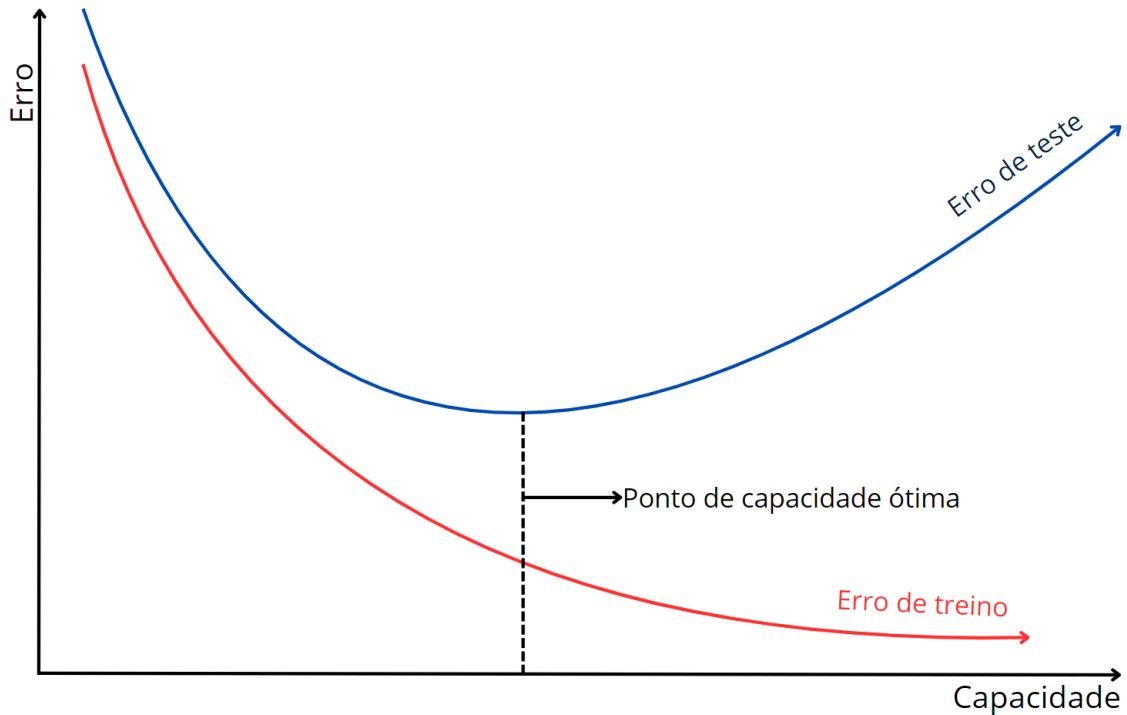


Figura 33 – Gráfico da curva de erros de treino (empírico) e teste (esperado) em função da capacidade do modelo.

possibilidade de que esse espaço não contenha funções semelhantes à função alvo, mas isso não é necessariamente verdade, pois esse espaço hipótese pode estar próximo o suficiente de $F(\vec{X})$ para ter funções semelhantes ou então estar muito afastado, ocasionando em funções muito distintas de $F(\vec{X})$. O erro de um algoritmo devido ao espaço $\mathcal{H}$ não conter a função alvo ou funções semelhantes é chamado de erro por viés [1]. Se um algoritmo é treinado com diferentes bases, ele provavelmente irá gerar modelos com funções $f(\vec{X})$ distintas. Os erros que são gerados por essas diferenças entre bases de treino são os erros por variância [1]. A figura 35 mostra as curvas dos erros por variância e viés em função da capacidade de um modelo. Ao aumentar a capacidade de um modelo, o espaço de hipóteses é capaz de englobar funções mais complexas. Então, se estamos querendo trabalhar em um problema que demanda modelos de alta complexidade, ao expandir os limites de $\mathcal{H}$ a tendência é que o algoritmo tenha acesso a funções semelhantes à função alvo, ocasionando em uma diminuição do erro por viés conforme a capacidade aumenta [1]. Por outro lado, o aumento da capacidade pode provocar o aumento no erro por variância, pois um modelo de alta capacidade consegue representar uma função que se ajusta muito bem a um conjunto de treino, mas que não se ajuste bem aos dados de teste. Esse ajuste pode ser tão preciso a ponto de gerar modelos muito distintos para bases de dados ligeiramente diferentes. Dessa forma, uma capacidade exageradamente alta para diferentes bases de treino proporciona

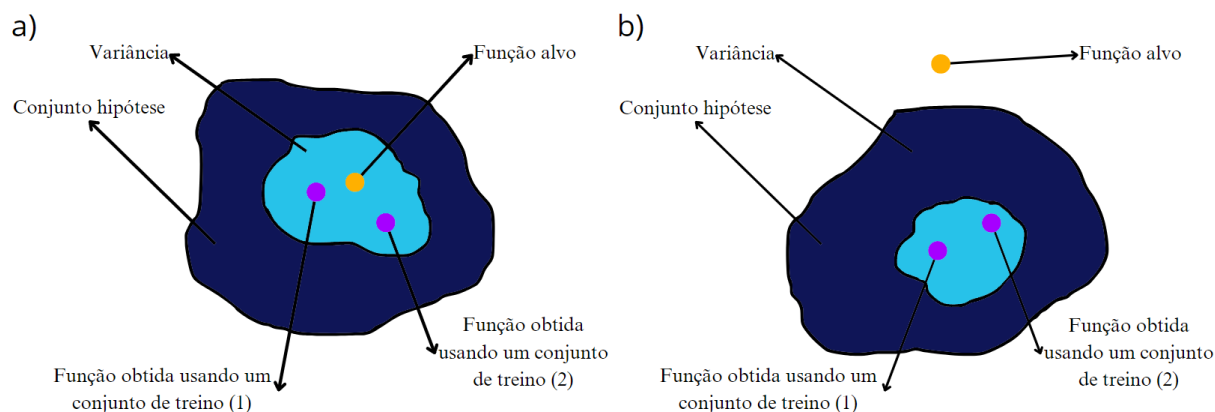


Figura 34 – Representação da variância e viés no espaço de hipóteses (azul escuro). Em a) o conjunto de funções do espaço hipótese contém a função ótima (alvo), mas em b) o espaço de hipóteses não é adequado o suficiente para conter a função alvo. A região em azul claro contém as diferentes funções que o modelo pode representar para conjuntos distintos de dados de treino.

funções que também são muito distintas entre si, ocasionando um alto erro por variância. Conforme exposto, a capacidade tem uma relação direta com o ajuste que o algoritmo

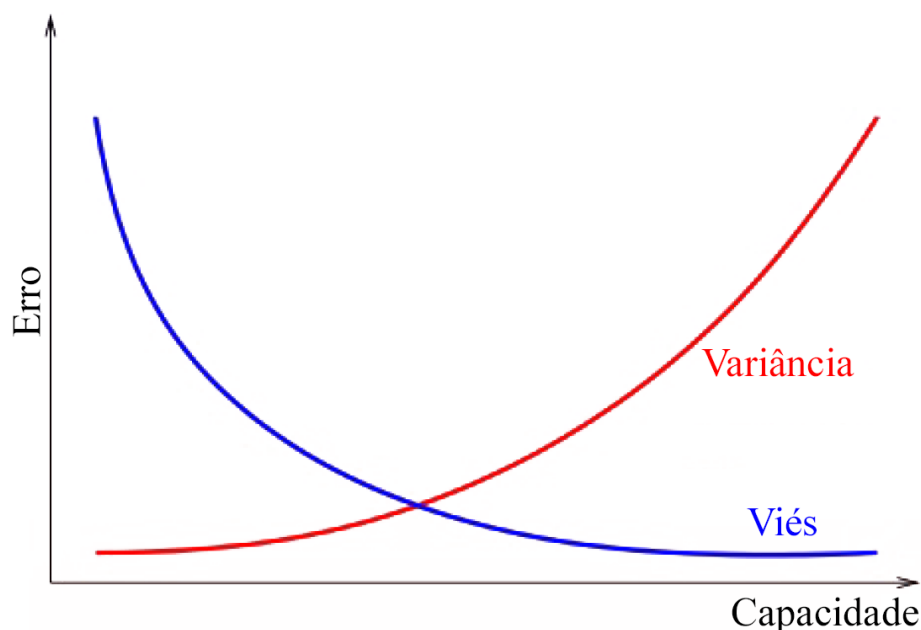


Figura 35 – Curvas dos erros por variância e viés em função da capacidade de um modelo.

fará a partir dos dados de treino, e então resta explorar como esses diferentes ajuste podem influenciar no desempenho do modelo para os dados de teste. Se tivermos um modelo com baixo erro empírico, mas alto valor de erro esperado, esse pode ser um cenário caracterizado como *overfitting*. Essa situação ocorre quando a função do modelo tem um ajuste tão restrito aos dados de treino que o modelo é capaz de ter ótimos resultados

dentro dos dados de treino, mas tem pouca eficiência para classificar os dados de teste. A figura 36 ilustra três tipos de ajustes possíveis para um mesmo conjunto de dados usando diferentes níveis de capacidade. Nos gráficos da figura 36, temos as saídas y associadas à *feature* x . Para a figura 36a) a capacidade usada é insuficiente, gerando um ajuste muito simples conhecido como *underfitting* (subajuste) [1], que provavelmente fornece altos erros empíricos e esperados. Na figura 36b) temos uma capacidade apropriada, que é capaz de fazer um bom ajuste aos dados. A consequência disso é que esse ajuste adequado tende a apresentar um valor baixo tanto no erro empírico quanto no esperado. Analisando a figura 36c), temos um ajuste que sobrepõe demais os dados (*overfitting*) [1]. Esse superajuste ocorre pois esse é um modelo com uma capacidade exagerada, capaz de alcançar funções tão complexas que podem ajustar demais aos dados de treino. Por consequência, um superajuste desse faz com que o modelo tenha bons desempenhos para os dados de treino, mas tende a ter desempenhos insatisfatórios para os dados de teste [1]. Se retomarmos a figura 33, ao aumentarmos a capacidade para além do ponto ótimo, temos ali uma situação que se assemelha bastante ao *overfitting*, pois o aumento da capacidade está causando uma diminuição do erro empírico, mas um aumento do erro esperado. Dessa forma, situações como o *underfitting* e o *overfitting* podem ser evitadas ao controlarmos a capacidade de um modelo.

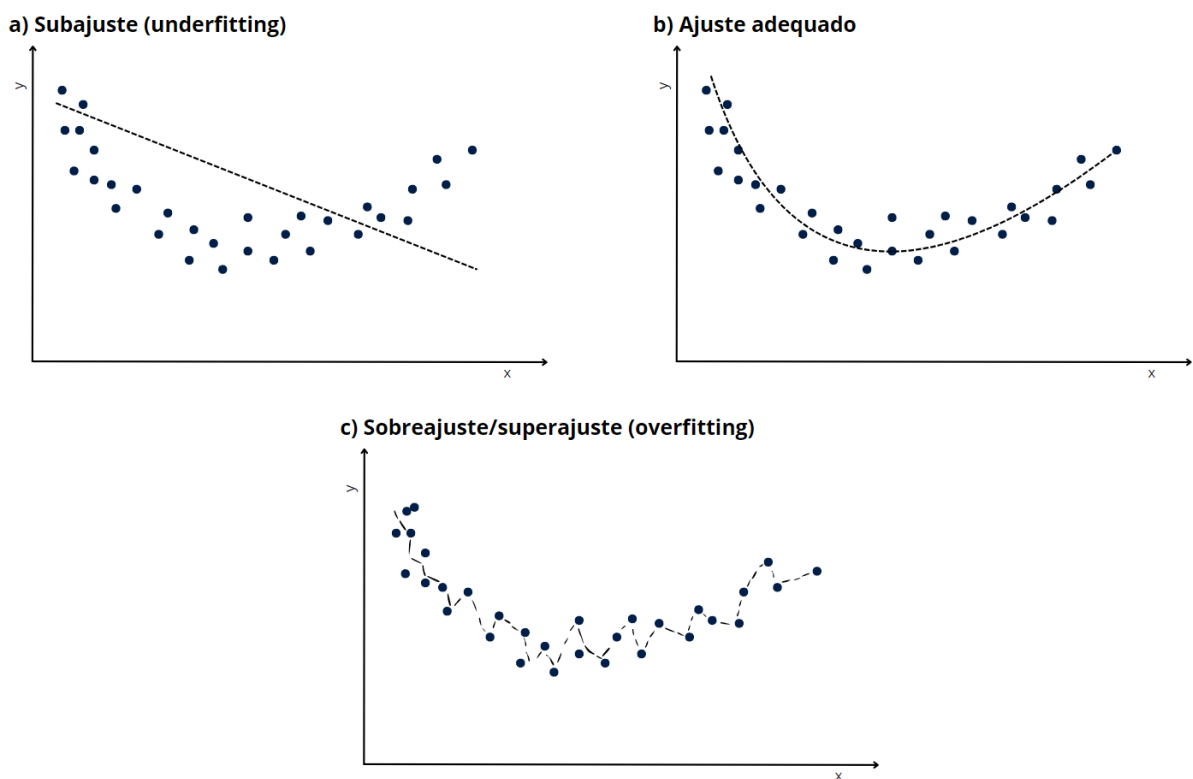


Figura 36 – Gráficos para três tipos de ajustes de uma saída y em função de uma *feature* x . Em a) temos um *underfitting*, b) é um ajuste adequado e c) é um *overfitting*.

Para administrar a capacidade de um modelo e colocá-la em níveis adequados é preciso penalizar os modelos que sejam excessivamente complexos, buscando funções que sejam mais simples, mas que tenham capacidade suficiente para aproximar os erros de treino e teste e mantê-los em baixos valores. Esse processo é chamado de regularização [1], e para aplicá-la é preciso tomar uma função de perda $L = L(\{\vec{X}_i\}, \{Y_i\}, \{y_i\}, W)$ e inserir um fator penalidade λ juntamente com uma função $C(W)$, que mede a complexidade do modelo. Então, a nova função a ser minimizada é uma combinação da função de perda L com esses fatores:

$$\mathcal{L} = L + \lambda C(W), \quad (6.6)$$

em que $\{\vec{X}_i\}$ é o conjunto de entradas do modelo, $\{Y_i\}$ são as saídas verdadeiras, $\{y_i\}$ são as saídas que o modelo previu e W são variáveis relacionadas à estrutura do algoritmo escolhido para a montagem do modelo. Na expressão (6.6) queremos que o valor de $\mathcal{L}$ seja grande para os casos em que a complexidade do modelo é exagerada, mas, por outro lado, o valor de $\mathcal{L}$ deve ser pequeno para um modelo com complexidade adequada. Um alto valor de $\mathcal{L}$ significa que o modelo analisado pode ter um nível não apropriado de complexidade. O λ deve ser escolhido de forma que ele não pode ser grande demais a ponto de sobressair em relação a L e $C(W)$, mas também não pode ser pequeno a ponto de enfraquecer a atuação do fator complexidade $C(W)$. Então, se para um modelo o erro da função de perda L é pequeno, mas a complexidade $C(W)$ é alta, então o valor final de $\mathcal{L}$ pode ser grande demais, revelando um modelo que pode não ser interessante de ser usado ainda que o erro da função de perda L seja baixo.

6.1.3 Métricas de desempenho

Até o momento, abordamos o desempenho de um modelo com base nos erros que ele comete ao classificar um conjunto de dados. Apesar dos erros serem uma medida bem direta e compreensível do desempenho de um modelo, nem sempre eles consideram algumas particularidades que podem ocorrer entre o modelo e a base de dados usada. Como exemplo, supomos que treinamos um modelo de classificação binária com uma base de treino composta de 90% de dados de uma classe 0 e os 10% restantes são da classe 1, e que temos uma base de teste que também carrega essa mesma proporção de classes. Consideremos agora que, após o treinamento, o modelo foi avaliado usando o grupo de teste e acertou todos os dados da classe zero, mas errou todos os dados da classe 1. Embora o erro seja de apenas 10%, o modelo se revelou incapaz de acertar qualquer dado de classe 1, e como isso não é uma situação muito desejada, analisar somente o erro não é suficiente para avaliar a qualidade de um modelo. Nesse exemplo, abordamos um problema relacionado a uma base desbalanceada com relação ao número de classes, demandando uma análise que deve ir além do número de erros. Para avaliar os modelos de maneira adequada, vamos explorar uma seleção de seis métricas de desempenho: acurácia (acr),

especificidade (esp), sensibilidade (sen), área abaixo da curva característica de operação do receptor (AUC), F1 e coeficiente de correlação de Matthews (MCC).

Com exceção da AUC, que independe de um valor limiar ou de corte (*threshold*) da probabilidade de classificação (outorgada pelo modelo a uma instância) para classificar as instâncias em uma determinada classe, todas as outras métricas dependem da escolha de um valor de *threshold*, e estaremos sempre considerando a atuação dele para todas as métricas em que ele é necessário. Como várias das métricas que serão apresentadas a seguir são muito presentes no contexto de diagnósticos médicos, embora não estejam restritas a ele, usaremos aqui os termos 'negativo' e 'positivo' como sinônimos das classes 0 e 1, respectivamente, como uma conveniência para facilitar o entendimento.

A acurácia é uma métrica comum e bem intuitiva de ser compreendida, mas ainda assim auxilia bastante na análise de desempenho, pois uma acurácia baixa é um indicativo de que talvez o modelo utilizado não é apropriado para fazer o tipo de tarefa que ele foi solicitado a desempenhar. A acurácia (acr) é matematicamente definida por [26]:

$$acr = \frac{VP + VN}{VP + VN + FP + FN}, \quad (6.7)$$

em que VP, VN, FP e FN são, respectivamente, os valores de: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

A acurácia por si só não é um indicador capaz de informar a proporção de acertos em cada classe analisada, então usamos a especificidade e sensibilidade para obter essas informações. A especificidade (esp) é a taxa de verdadeiros negativos previstos pelo modelo e é definida como [26]:

$$esp = \frac{VN}{VN + FP}. \quad (6.8)$$

Por outro lado, a sensibilidade representa a taxa de verdadeiros positivos de um modelo, e em um contexto de classificação binária a sensibilidade é comumente chamada de *recall* (rec ou revocação) [26, 27]. A sensibilidade (sen) é definida como [26]:

$$sen = rec = \frac{VP}{VP + FN}. \quad (6.9)$$

Essas duas métricas podem ser usadas em bases de dados desbalanceadas para evitar escolher um modelo com uma alta acurácia no geral, mas que tenha um desempenho ruim em qualquer uma das classes.

As quantidades de falsos e verdadeiros das classes positivas e negativas podem ser empregadas na métrica F1 (média harmônica entre precisão e *recall*) para avaliar o desempenho de um modelo em relação às classes, mas antes precisamos definir a quantidade precisão (pre), que também é conhecida como valor preditivo positivo [26]:

$$pre = \frac{VP}{VP + FP}. \quad (6.10)$$

Dessa forma, podemos usar as equações (6.9) e (6.10) para definir a expressão de F1 [26]:

$$F1 = 2 \frac{pre \cdot rec}{pre + rec}, \quad (6.11)$$

em que os valores de F1 ficam contidos no intervalo $[0, 1]$.

O coeficiente de correlação de Matthews (MCC) mede uma correlação entre os valores verdadeiros e previstos para as instâncias e é definido por [26, 28]:

$$MCC = \frac{VN \cdot VP - FN \cdot FP}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}, \quad (6.12)$$

e o valor de MCC fica restrito ao intervalo $[-1, 1]$. Tanto o MCC quanto F1 são métricas que proporcionam um entendimento mais geral e minucioso do modelo a partir das taxas de verdadeiros e falsos de cada classe. Essas duas métricas são mais sensíveis a desbalanços de acertos e erros entre classes quando comparadas à acurácia, especificidade e sensibilidade.

Até o momento abordamos métricas que dependem de um *threshold* fixo para coletar as probabilidades das amostras e classificá-las. Porém, esse limiar pode ser uma escolha com um grau de arbitrariedade, e isso é capaz de introduzir um viés durante o processo de classificação que dificulta a comparação de desempenho entre os modelos. A métrica AUC tem a vantagem de ser independente do valor do *threshold* escolhido pelo supervisor do modelo, e dessa forma ela não carrega vieses e permite a comparação direta do desempenho de diversos modelos. Para calcular a AUC, considere um modelo já treinado que será avaliado com um conjunto de dados de teste $\{\vec{X}_i\}$. Ao enviar esses dados de teste ao modelo, ele retornará um conjunto de probabilidades $\{P_i\}$, que estão relacionadas a cada instância. O que faremos agora é escolher uma coleção de todos os *thresholds* possíveis $\{th_j\}$ no intervalo $[0, 1]$ e vamos usar cada um desses valores para classificar esses dados de teste a partir de suas probabilidades no conjunto $\{P_i\}$. Então, cada *threshold* utilizado na classificação está associado a um quarteto de valores de falsos e verdadeiros positivos e negativos (VP_j, FP_j, VN_j, FN_j) . Ao fim desse processo, teremos então um conjunto de taxas $\{VP_j, FP_j, VN_j, FN_j\}$ que foram coletadas para cada *threshold* no conjunto $\{th_j\}$. Com isso, podemos formar uma coleção de especificidades e sensibilidades $\{especificidade_j, sensibilidade_j\}$ usando as equações (6.8) e (6.9). Se coletarmos esse conjunto de pares e elaborarmos um gráfico em que o eixo vertical contém os valores de *sensibilidade* e o eixo horizontal recebe $(1 - especificidade)$, que é a taxa de falsos positivos, teremos a formação da curva característica de operação do receptor (ROC) [26]. Como a especificidade e sensibilidade são quantidades que ficam no intervalo $[0, 1]$, os eixos vertical e horizontal da curva ROC também respeitam esses intervalos.

A figura 37 ilustra uma curva ROC (linha preta), em que a linha vermelha tracejada é a curva referente a um modelo com classificação aleatória. Curvas ROC que estejam acima e para a esquerda da linha vermelha estão associadas a modelos que tendem a ter um bom desempenho, enquanto curvas abaixo e para a direita da linha costumam

representar modelos com desempenho não satisfatório. Após traçar a curva ROC, a AUC é coletada ao calcular a área que está abaixo dessa curva. No caso da figura 37 temos uma AUC de 0.83.

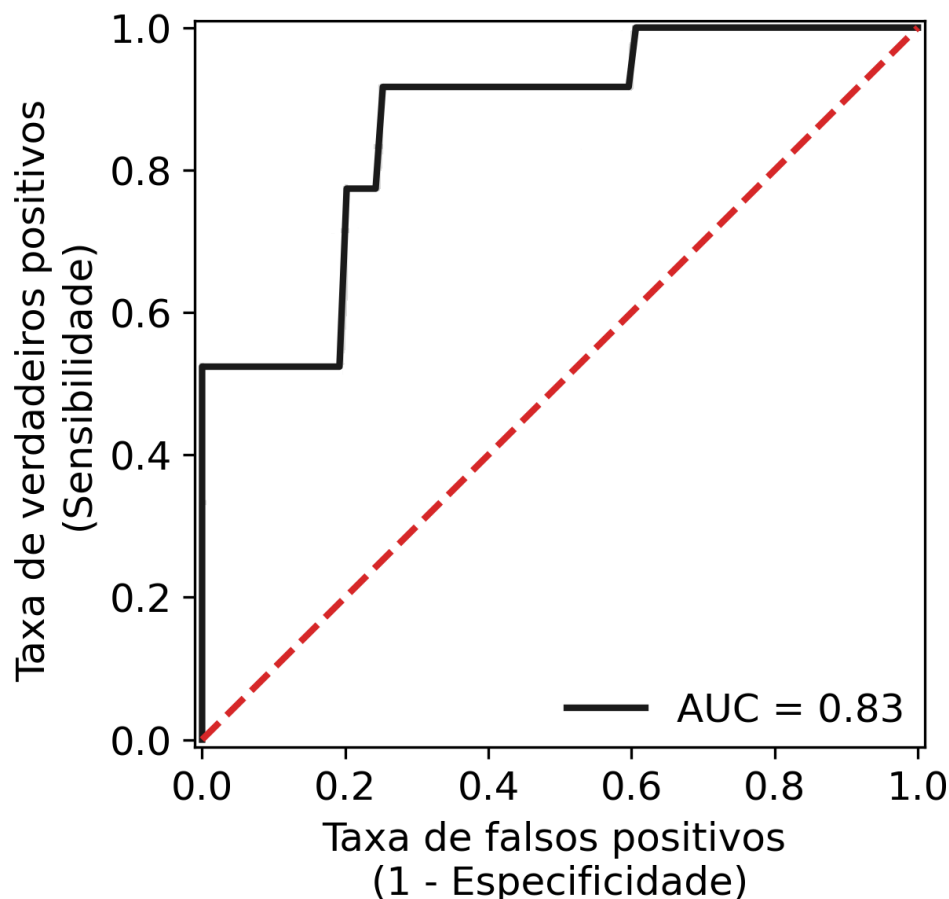


Figura 37 – Demonstração de uma curva ROC para uma $AUC = 0.83$.

Os valores mínimo e máximo da AUC são, respectivamente 0 e 1. Se um modelo tem $AUC = 0$ o desempenho dele é o pior possível, mas se $AUC = 1$ então o modelo representa um classificador perfeito. Por fim, temos o caso em que $AUC = 0.5$, e ele ocorre quando a curva ROC coincide com a reta tracejada vermelha da figura 37, que é a linha que representa um modelo com classificação aleatória como foi mencionado anteriormente. A figura 38 ilustra uma curva ROC com $AUC = 1$, e, em uma situação como essa, uma AUC máxima significa que o modelo usado contém um valor de *threshold* que é capaz de classificar todas as amostras positivas e negativas sem nenhum erro. O ponto na curva que representa esse classificador perfeito é o (0,1), pois nele temos que a taxa de falsos positivos é 0 e taxa de verdadeiros positivos é 1. Em geral, o ponto da curva ROC mais próximo da coordenada (0,1) define o *threshold* que maximiza a acurácia. Esse será o valor de *threshold* utilizado neste trabalho para o cálculo de todas as métricas com exceção de AUC.

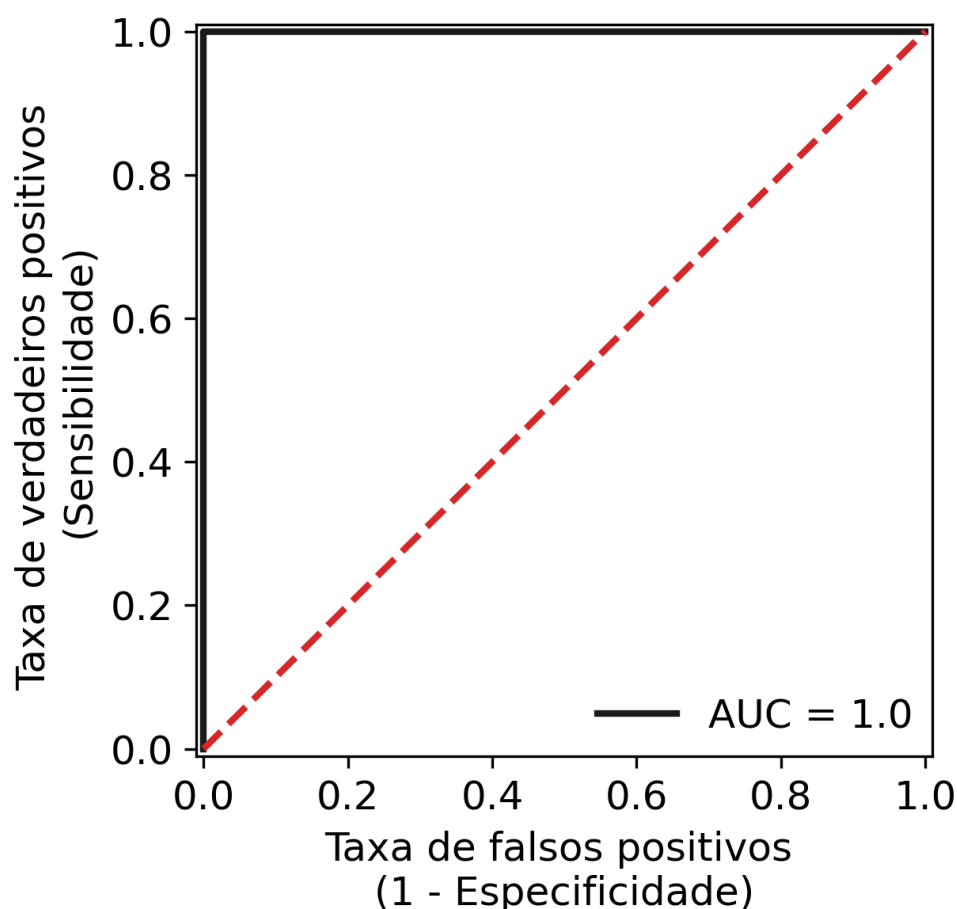


Figura 38 – Curva ROC de um classificador perfeito (AUC máxima).

No contexto de diagnósticos clínicos, podemos considerar que modelos com AUC acima de 0.9 são de alto desempenho, enquanto 0.7-0.9 indicam um desempenho moderado e 0.5-0.7 possuem baixo desempenho [29].

6.1.4 Otimização de um modelo

Conforme apresentado anteriormente, a arquitetura e estrutura base de um modelo dependem dos valores de hiperparâmetros associados a cada tipo de algoritmo. Considerando que os hiperparâmetros podem ser de natureza contínua, discreta ou categórica e podem então assumir uma variedade de formas, fica o questionamento: dada uma tarefa que se queira resolver, como escolher os hiperparâmetros para solucionar esse problema da melhor maneira atingível? Para resolver o dilema desse questionamento, usamos a otimização dos modelos para coletar hiperparâmetros ótimos.

A otimização é uma técnica usada no aprendizado de máquina com objetivo de encontrar um conjunto de hiperparâmetros ótimos para um modelo. Esse processo é geralmente feito a partir da função objetivo, que é uma função que treina um algoritmo a partir de um conjunto de hiperparâmetros candidatos e depois testa o modelo resultante

com base em alguma métrica escolhida previamente. Esse processo é repetido para vários hiperparâmetros candidatos quantas vezes se queira, e o conjunto de hiperparâmetros escolhidos ao fim do processo é aquele que proporcionou o maior valor da métrica escolhida. Cada vez que um conjunto de hiperparâmetros é formado ele pode ter sido gerado aleatoriamente ou então com base em algum cálculo que leve em consideração os resultados da métrica do conjunto anterior. Esse é o processo geral da otimização, e ele pode envolver nuances e particularidades. A otimização será abordada novamente na seção seguinte sobre métodos de validação.

A figura 39 mostra um exemplo de otimização simples. Os dados são divididos entre treino e teste, em que o primeiro grupo é usado para gerar um modelo a partir de um grupo de candidatos a hiperparâmetros, que são gerados de forma aleatória. Após o treinamento, o modelo é testado pelo outro grupo e a sua acurácia dentro do grupo de teste é armazenada. Esse processo é repetido vinte vezes (linha cinza tracejada), em que cada tentativa (trial) é feita para um conjunto de hiperparâmetros diferentes da tentativa anterior. Então, ao total são coletadas vinte acurácias, sendo que cada uma é associada ao conjunto de hiperparâmetros que foi usado para gerá-la. Ao fim do processo, é escolhido o conjunto de hiperparâmetros que resultou na melhor acurácia.

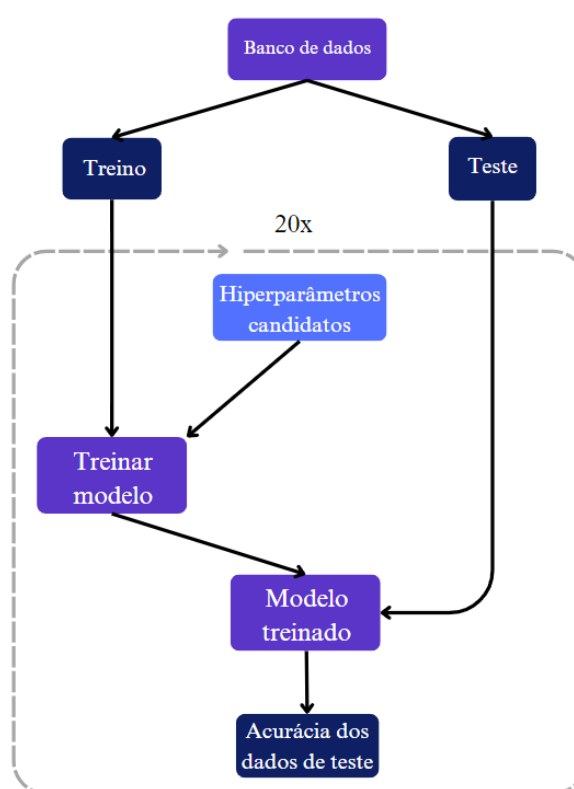


Figura 39 – Esquema de otimização de um modelo com vinte tentativas com base na acurácia dos dados de teste.

6.2 Métodos de validação

No contexto de aprendizado de máquina, sempre há um dilema com relação à divisão da base de dados para treino e teste. Usar muitos dados para treino pode garantir um modelo mais bem treinado e com melhor capacidade de generalização, mas o efeito colateral disso é uma quantidade pequena de dados de teste. Isso é um problema pois uma baixa quantidade de dados de teste pode fornecer resultados muito otimistas ou muito pessimistas sobre o desempenho do modelo, posto que é possível que uma base de dados de teste pequena não represente uma quantidade adequada para estimar o desempenho esperado do modelo. A outra via é usar poucos dados no treino e muitos no teste, porém uma quantidade de dados baixa no treino pode ser insuficiente para que o modelo resultante seja capaz de construir padrões de identificação entre cada classe de dados, gerando um modelo mal treinado, ainda que tenha muitos dados para teste de desempenho.

Há uma outra via para estimar o desempenho do modelo, em que podemos pegar a base de dados completa, dividi-la em partes e permutar os dados entre treino e teste. Há várias formas de fazer isso, mas aqui iremos abordar duas: *stratified K-fold cross-validation* (validação cruzada estratificada em k-partições) e *nested cross-validation* (validação cruzada aninhada).

6.2.1 Validação cruzada estratificada em k-partições

Para a validação cruzada estratificada em k-partições (SKFCV) a base de dados é dividida em k partes (*folds*) de forma que temos uma permutação sobre qual *fold* será usado para testar o modelo, enquanto o restante é utilizado para treiná-lo [30]. Além disso, a estratificação dessa validação cruzada garante que cada *fold* gerado contenha a mesma proporção de amostras de cada classe [30]. A figura 40 mostra uma divisão de modo estratificado para uma base de dados com duas classes distintas. A base de dados possui quinze dados positivos (azuis) e dez dados negativos (vermelhos), resultando em uma proporção de 3:2. Ao dividir essa base em cinco partições de modo estratificado, cada repartição deve preservar essa proporção de 3:2, então cada *fold* terá, nesse caso, três dados positivos e dois dados negativos. A figura 41 mostra o exemplo de uma validação cruzada para cinco partições. O banco de dados é dividido em cinco *folds* distintos, e cada uma das cinco iterações terá *folds* de teste diferentes, tal que todos os *folds* terão sido usados para treino ou teste do modelo em alguma iteração. Cada iteração será treinada e testada a partir dos *folds* designados para tais tarefas. Então, após o treino de cada iteração, o modelo gerado por cada uma é usado para coletar a acurácia nos dados de teste, e a acurácia final é dada pela média das cinco acurácias coletadas de cada iteração. Embora esse exemplo use média das acurácias, é possível utilizar outros tratamentos estatísticos e também outras métricas. É importante destacar que cada conjunto de *folds* de treino é usado para fazer um processo individual de ajuste para a produção do modelo.

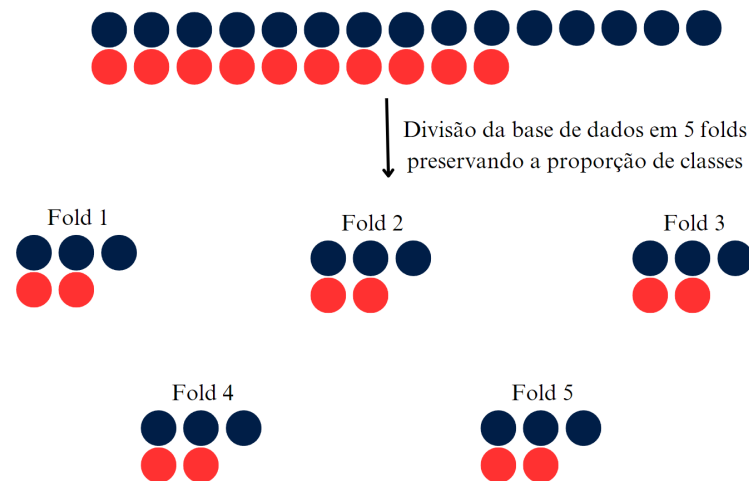


Figura 40 – Divisão de uma base de dados em cinco *folds* de modo estratificado para duas classes.

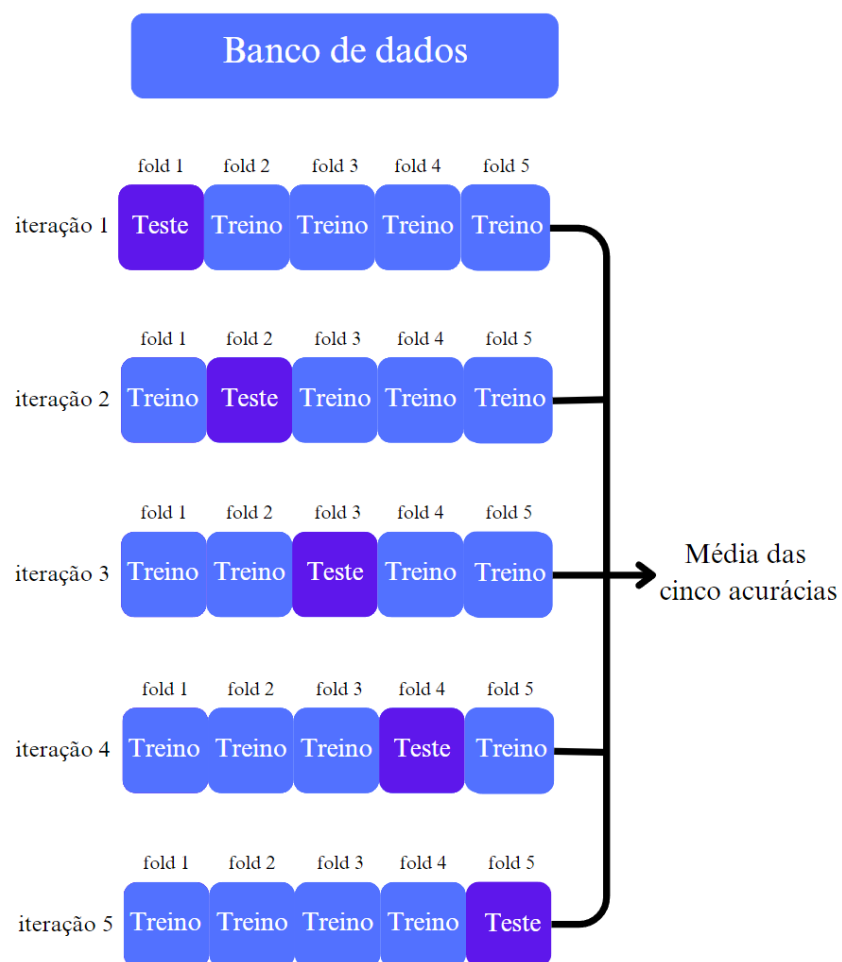


Figura 41 – Exemplo de uma validação cruzada de cinco partições avaliada por acurácia.

Portanto, ao ajustar um algoritmo para um conjunto de *folds* de treino, esse ajuste deve ser completamente descartado quando um novo conjunto de *folds* de treino for selecionado

nas etapas seguintes da validação cruzada. Esse processo é necessário posto que, se os ajustes feitos não forem redefinidos ao trocar os grupos de treino, haverá um vazamento de informações que fará o modelo ter um desempenho com um viés otimista. Quando um algoritmo é ajustado aos dados de treino, o modelo resultante aprendeu os padrões desses dados e posteriormente esse aprendizado é verificado nos dados de teste. Se os ajustes não forem eliminados para os novos grupos de treino e teste durante as etapas da validação cruzada, o modelo conterà informações sobre o grupo de teste (vazamento de informações), pois anteriormente ele foi usado para treinar o modelo. A figura 42 mostra gráficos comparativos para duas validações cruzadas de oito partições feitas de maneiras distintas para uma rede neural. As curvas dos gráficos são as acurácias (eixo y) para os dados de treino e teste associados a cada iteração (eixo x). Para a situação a) temos o método mais apropriado, que é quando o ajuste do treino é reiniciado toda vez em que um novo grupo de treino e teste é formado. Nesse caso, temos então resultados de acurácias que representam medidas mais realistas ao longo das validações cruzadas. Esse cenário demonstra uma situação mais confiável para o real desempenho dessa rede neural, e essa deve ser a metodologia de ajuste usada para buscar as melhores estimativas dos modelos. Em b) a validação cruzada ocorre, mas sem que os ajustes sejam reiniciados entre cada iteração. É notável que o desempenho tende a aumentar conforme mais iterações são treinadas e testadas, pois o modelo é alimentado cada vez mais com os ajustes dos dados que ele já viu, proporcionando um desempenho enviesado.

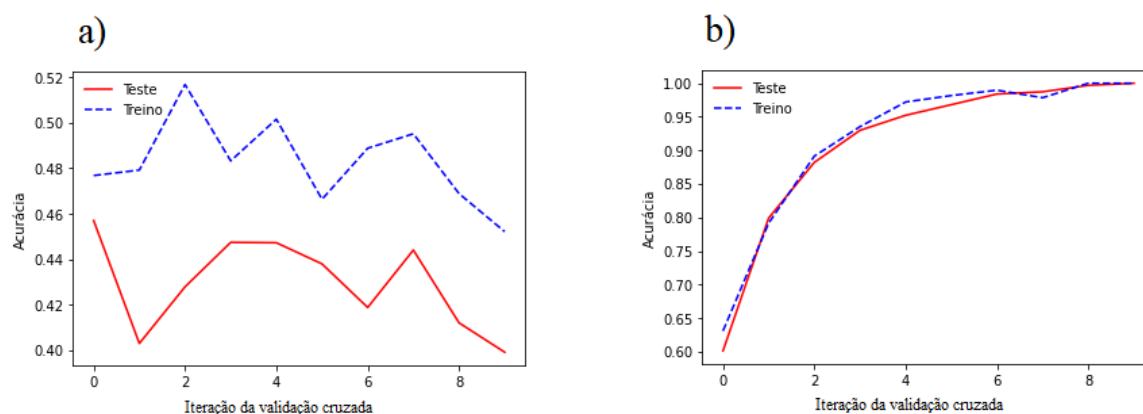


Figura 42 – Comparativo de validações cruzadas para uma rede neural. Em a) os ajustes são reiniciados a cada mudança de iteração, proporcionando resultados mais confiáveis. Na situação b) temos uma validação cruzada sem que os ajustes sejam reiniciados a cada nova iteração, provocando um vazamento de dados que proporciona um falso melhor desempenho.

6.2.2 Validação cruzada *Nested*

A validação cruzada *Nested* (*Nested CV*) é uma maneira alternativa de avaliar os resultados de um aprendizado de máquina. Esse método é particularmente útil, pois separa os dados que vão otimizar um modelo dos dados que irão testar esse modelo otimizado. Então, é uma maneira eficaz de aplicar seleção de *features* e ajuste de hiperparâmetros de aprendizado de máquina para treinar um modelo e fazer uma boa previsão de resultados a partir da otimização [31]. Esse método faz uso do mesmo princípio da validação cruzada estratificada, porém ele contém uma divisão adicional nos dados de forma que agora a base de dados é dividida em três grupos: treino, validação e teste. Na prática, a *nested CV* apresenta uma validação cruzada externa e uma interna de forma que os dados da divisão externa não contribuem para a otimização do modelo, sendo eles utilizados exclusivamente para teste.

A figura 43 apresenta um exemplo de funcionamento da *nested CV*. O método se inicia ao fazer uma divisão da base de dados de modo estratificado em cinco partições, tal que uma dessas partes é selecionada e se torna o grupo de teste enquanto as quatro repartições restantes formam o grupo A. Então, esse grupo A é dividido em 10 partes de forma que uma partição é o grupo de validação e o restante é usado para montar o grupo B. Esses dois grupos definem uma validação cruzada de 10 *folds*, e a otimização é feita nessa validação cruzada ao buscar os parâmetros e hiperparâmetros ótimos. O algoritmo é treinado usando os nove *folds* do grupo B e depois o modelo gerado passa por uma validação coletando, por exemplo, a acurácia do modelo para os dados do grupo de validação. Os *folds* das dez divisões permutam entre si para treinar e validar o modelo, de forma que teremos uma acurácia para cada *fold* de validação, e a otimização entregará o modelo com maior média dessas dez acurácias. Essa busca pelos parâmetros e hiperparâmetros ótimos acontece vinte vezes, e ao final temos um conjunto otimizado de parâmetros e hiperparâmetros, que são usados para treinar o modelo usando o grupo A. Então, o modelo treinado é posto para classificar o grupo de teste. Com a coleta da acurácia do grupo de teste, o grupo A é modificado, e o mesmo acontece com o de teste. Esse processo ocorre até que todos os 5 *folds* da primeira divisão de dados tenham permutado entre si no grupo de teste. Como teremos um total de cinco grupos de teste distintos, ao fim do processo teremos cinco resultados de acurácias, e o desempenho será a média entre elas. Então, quando a *nested CV* se encerrar, terão tido cinco modelos otimizados, treinados e testados e cada um deles tem um conjunto de métricas de desempenho e de hiperparâmetros. Embora o exemplo esteja usando a média das acurácias para avaliar o modelo por meio da *nested CV*, essa não é a única maneira de proceder. A partir desse ponto, é possível escolher a maneira que pareça mais adequada para reportar os resultados desejados. Uma abordagem possível é a aplicação da média de uma ou mais métricas dos cinco modelos construídos, como por exemplo informar a média entre os modelo para a acurácia, especificidade e sensibilidade.

Uma outra opção viável é analisar os cinco modelos e emitir o melhor resultado entre eles.

Se feita de maneira apropriada, a *nested CV* tem uma aplicação muito adequada, pois como a otimização feita no processo não envolve os dados que são usados para avaliar o modelo ótimo, então não ocorre o risco de vazarem informações do teste para os parâmetros e hiperparâmetros ótimos. Isso proporciona um resultado que permite estimar bem como seria o desempenho do modelo para dados que não foram usados para otimizar e treinar o modelo [31].

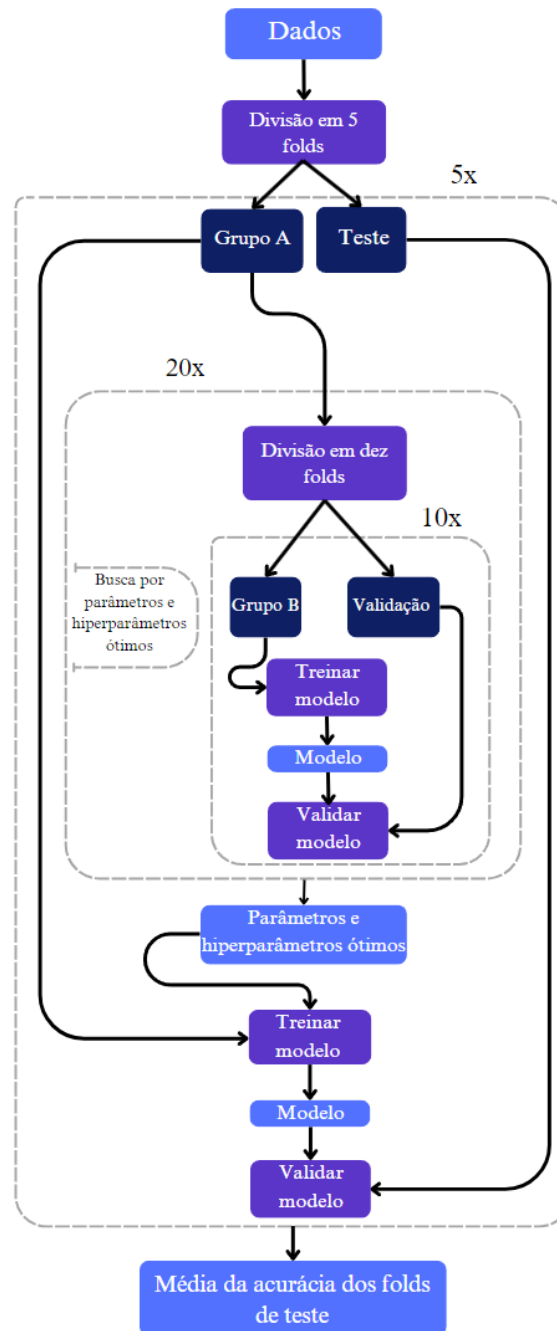


Figura 43 – Esquema de funcionamento da *nested cross validation*. Os dados são divididos de forma que uma porção será de dados para otimizar um modelo e a outra contém dados que não foram utilizados na otimização. Essa outra porção é usada para testar o modelo otimizado. Permutando os dados de teste e treino, teremos um número modelos otimizados equivalente à quantidade de permutações.

6.3 Algoritmos

Dentro da área de aprendizado de máquina há uma quantidade vasta de algoritmos que se propõem a buscar padrões em dados com o objetivo de executar uma tarefa de

classificação. Porém, uma análise desses vários algoritmos revela que muitos deles possuem estruturas e bases em comum, formando o que será chamada de família de algoritmos. Partindo de agora, será percorrido sobre cinco algoritmos, tal que cada um deles é baseado em cinco grandes famílias de algoritmos: hiperplanos, regressões, perceptrons, distância entre dados e árvore de decisão. Tomando essa lista, os algoritmos que usaremos para representar cada item dela são, respectivamente: Support Vector Machine (SVM), Logistic Regression (LR), Multilayer perceptron (MLP), K-Nearest Neighbor (KNN) e Extreme Gradient Boosting (XGBoost). Esses algoritmos serão referidos frequentemente por suas abreviações por serem usualmente difundidas. Cada um desses algoritmos por si só já é capaz de render muitas considerações, de modo que focaremos em suas principais características e funcionamento geral. Nesta seção a quantidade índices para identificar dados aumenta bastante, então para não causar confusões, os dados de entrada anteriormente escritos como $\vec{X}_i$ serão escritos como $\vec{X}_u$, pois o índice i será reservado para matrizes, tensores de terceira ordem e somatórios.

6.3.1 Support Vector Machine (SVM)

O SVM é um algoritmo capaz de separar dados de diferentes classes por meio de hiperplanos em um espaço. Por exemplo, se temos um conjunto de instâncias $\{\vec{X}_u\}$ com duas componentes (*features*) e classes 0 ou 1, esses dados de entradas habitam um espaço de duas dimensões como pontos de coordenadas $\{(x_{1u}, x_{2u})\}$ e podem ser separados por uma curva que estabelece os limites de onde começa e termina cada classe [24]. Para entendermos melhor o SVM, focaremos em alguns hiperparâmetros e estruturas do algoritmo.

A figura 44 mostra um conjunto de dados representados por duas *features* e com duas classes distintas (azul e vermelha). Esses dados são usados para treinar três modelos diferentes, que conseguem separar por classe completamente todos esses dados usando qualquer uma das três retas da figura 44 a), b) e c). Todos os dados vermelhos estão em um lado do separador e os azuis estão do outro, indicando que não há erros de classificação para esses dados de treino (erro empírico nulo). Se temos três separadores que conseguem discriminar os dados de maneira perfeita, qual deles seria o mais adequado para usar em um modelo? É evidente que a escolha do modelo pode depender do que o pesquisador julga mais apropriado, mas vamos aqui apresentar algumas ferramentas que proporcionem uma intuição de como essa escolha pode ser feita.

Começaremos então comparando esses três separadores para verificar o que há de diferente entre eles. Na figura 44, temos uma linha tracejada que indica distância entre o separador para os pontos de cada classe que estão mais próximos do separador. Chamaremos de margem a soma da distância entre esses dois pontos até o separador, e ela pode ser visualizada ao traçar para cada um desses dois pontos uma reta paralela ao separador e depois ligá-las pela menor distância entre elas. Se o algoritmo iniciar na

situação da figura 44a), temos uma margem que é curta e o separador chega muito próximo dos pontos com menores distâncias até ele. Se rotacionarmos esse separador no sentido horário, teremos a configuração da figura 44b), em que a margem aumenta o seu tamanho e, por consequência, a distância entre os pontos mais próximos e o separador também. Se continuarmos rotacionando esse separador no sentido horário, temos a estrutura da figura 44c), com uma margem tão pequena quanto da figura 44a), mas a inclinação dos dois separadores é diferente. Das três situações, a que apresenta maior tamanho de margem é a da figura 44b), e geralmente essa é uma condição adequada para um modelo de SVM assumir [24]. A busca por um modelo com uma margem adequada é uma boa prática

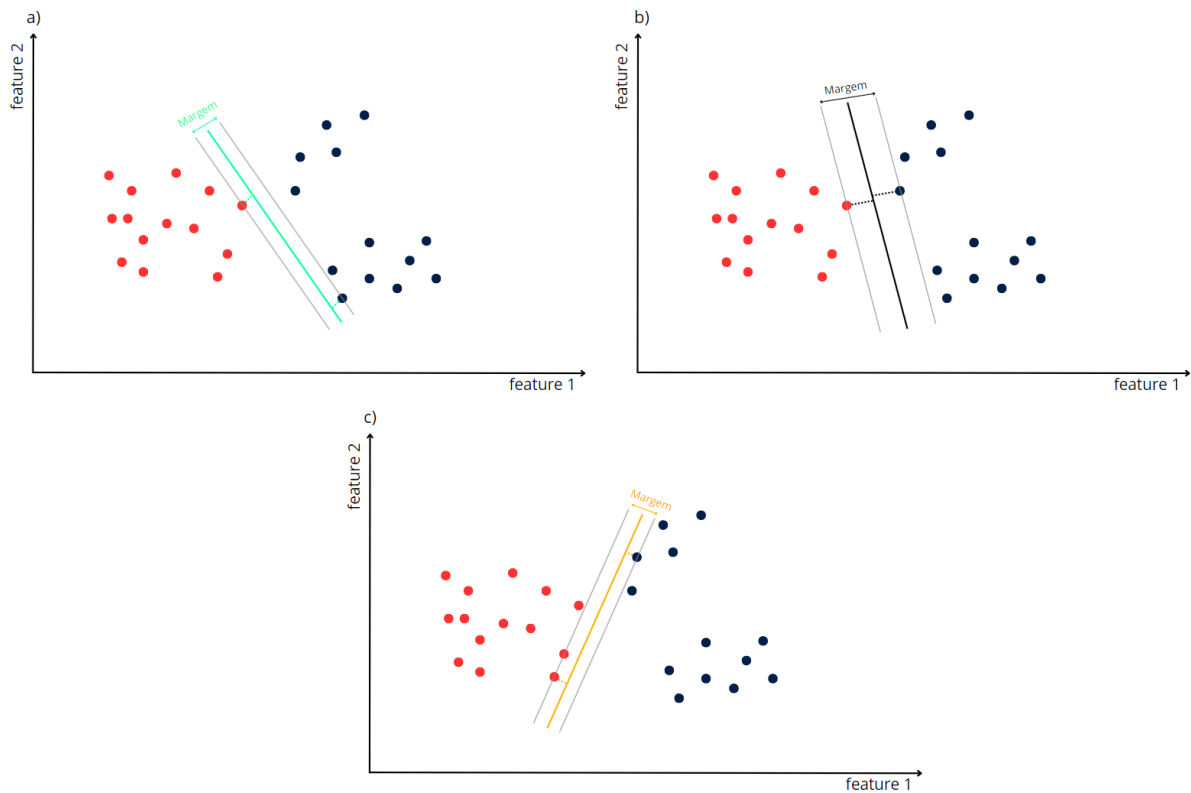


Figura 44 – Três diferentes situações para separar um conjunto de dados em duas dimensões. As linhas tracejadas são as menores distâncias do separador até os pontos mais próximos a ele para cada classe. Em a) o separador é uma reta de inclinação negativa com uma margem pequena. Para o caso b) a inclinação do separador é negativa e o tamanho da margem é o maior dos três casos. Em c) a margem também é pequena e tem um separador de inclinação positiva.

pois ela tende a diminuir o erro esperado. Esse tamanho da margem é influenciado pelo hiperparâmetro C , em que valores grandes de C tendem a selecionar margens pequenas, e valores menores de C proporcionam margens maiores. Se tomarmos um ponto da figura 44 e criarmos ao redor dele outros pontos, podemos encarar que esses novos dados criados são pontos para verificar o desempenho de cada modelo em um grupo de teste formando por esses novos dados. A figura 45 pega um ponto da classe vermelha da figura 44, marca-o

com a cor cinza e cria outros seis pontos da classe vermelha ao redor dele. Na figura 45a) vemos que o modelo cometeu dois erros, pois há dois pontos vermelhos que estão no lado azul do classificador. Para 45b) o modelo teve um erro esperado nulo, e no caso da 45c) o modelo classificaria um ponto vermelho como sendo da classe azul. Apesar dos três casos terem um erro empírico zero baseado na figura 44, foi possível observar agora que o classificador com maior tamanho de margem foi aquele em que o erro esperado foi o mais baixo. Porém, é importante destacar que nem sempre o modelo com maior margem será aquele com melhores resultados e o mais adequado, mas a busca por um tamanho apropriado de margem é uma boa prática a ser utilizada ao se trabalhar com o SVM, pois tende a evitar grandes erros por variância [24, 32].

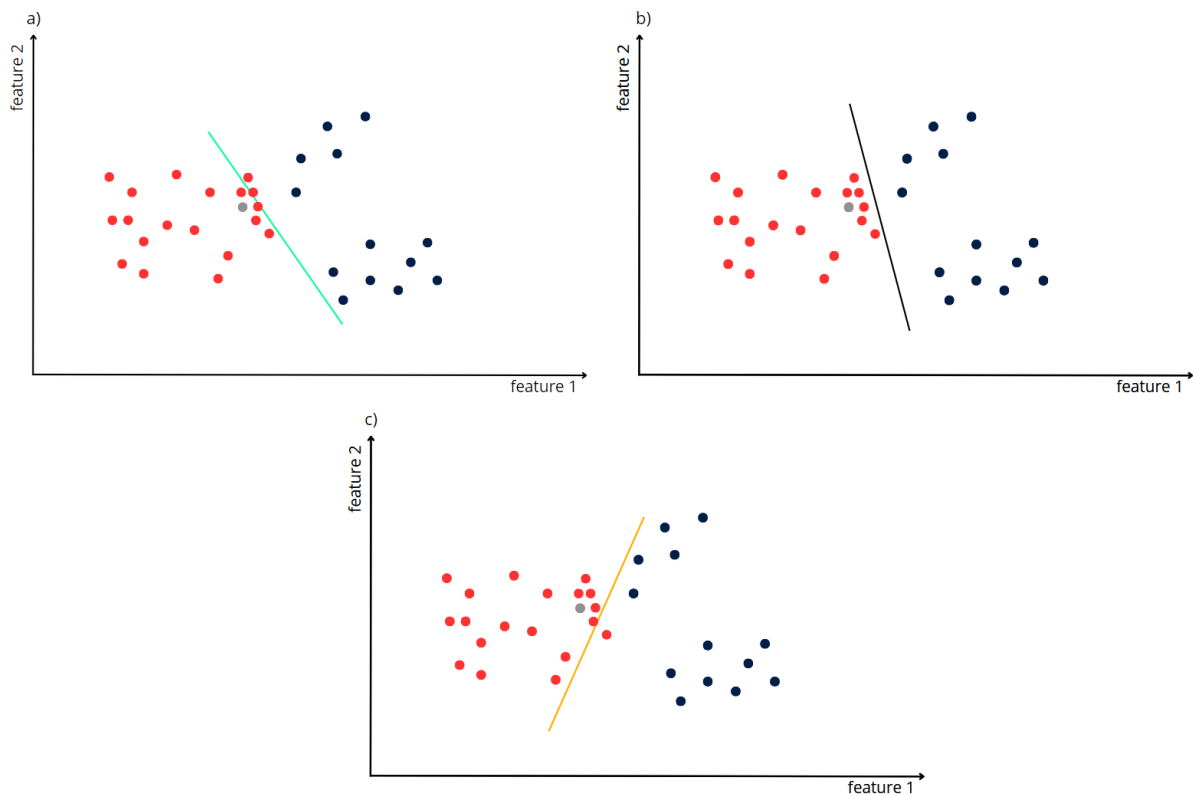


Figura 45 – A inserção de novos dados ao redor do ponto cinza fornece uma estimativa de como cada modelo poderia se comportar em um grupo de dados de teste. Em a) temos dois dados vermelhos que seriam erroneamente classificados como azuis, em b) não houve erro de classificação e a situação c) apresenta um dado vermelho que seria equivocadamente classificado como azul. Isso ilustra como uma margem maior pode evitar os erros por variância.

Como o SVM faz o processo de classificação ao separar os dados que estão em um espaço, é importante compreender qual a relação entre os espaços e suas dimensões com a eficiência que o algoritmo pode ter para classificar os vetores de entrada. Na classificação binária para um espaço 2D, por exemplo, se temos três pontos quaisquer é sempre possível

separá-los por classes desde que os três não sejam colineares, pois esse é o caso em que os três pontos estariam confinados a apenas uma dimensão. O número máximo de pontos que permite a um modelo de hiperplanos separá-los por classes é chamado de dimensão Vapnik–Chervonenkis (dimensão-VC) e, para um espaço de dimensão N , a dimensão-VC é dada por

$$\text{dimensão-VC} = N + 1. \quad (6.13)$$

Se o número de pontos para um espaço de dimensão N ultrapassar o valor de $N + 1$, então esses pontos não podem ser perfeitamente separáveis por um hiperplano para um espaço de dimensão N . Nesse caso, se existe um número máximo de pontos $N + 1$ que podem ser separados em classes por um hiperplano para uma dada dimensão N , o que fazer quando o número de pontos for superior a esse limite de $N + 1$? Uma abordagem possível é aumentar a dimensionalidade dos dados a partir de funções chamadas de kernels, que processam esses dados para projetá-los em maiores dimensões. Dessa forma, ao projetar um conjunto de pontos em maiores dimensões, pode ser possível separá-los em classes usando hiperplanos, pois dimensões extras afetam como os dados serão distribuídos no novo espaço e, se essa nova organização dos dados for apropriada, as classes podem ser separáveis por um hiperplano. Porém, nem sempre é possível usar uma nova *feature* nos dados para analisá-los em uma dimensão a mais para buscar uma separação de classes, pois em algumas situações a separação não está sendo adequada mesmo usando todas as *features* disponíveis. Nesse caso, podemos aumentar a dimensionalidade do problema ao mapear os dados de entrada para dimensões mais altas a partir de similaridades entre as entradas do conjunto $\{\vec{X}_u\}$ [32], criando novas *features* a partir dessa base de dados. A maneira de fazer isso é comumente conhecida como kernels, que são funções capazes de exercer esse tipo de tarefa. A medida de similaridade feita por um kernel linear é dada por:

$$K(\vec{X}_u, \vec{X}_a) = \vec{X}_u \cdot \vec{X}_a, \quad (6.14)$$

em que $\vec{X}_u$ e $\vec{X}_a$ são instâncias de índices u e a .

A figura 46 ilustra um exemplo de atuação de kernel. Para a situação da figura 46a) temos quatro pontos divididos em duas classes (azul e vermelha) e projetados em um plano com duas *features*. O espaço em duas dimensões que esses pontos habitam não proporciona que as classes sejam separadas usando uma reta. Então, na figura 46b) um kernel $K(\vec{X}_u, \vec{X}_a)$ é utilizado para projetar esses dados em uma terceira dimensão (kernel *feature*) a fim de tentar uma classificação satisfatória. Para a figura 46c) podemos notar que o hiperplano verde consegue separar completamente os dados vermelhos (acima do hiperplano) dos azuis (abaixo do hiperplano).

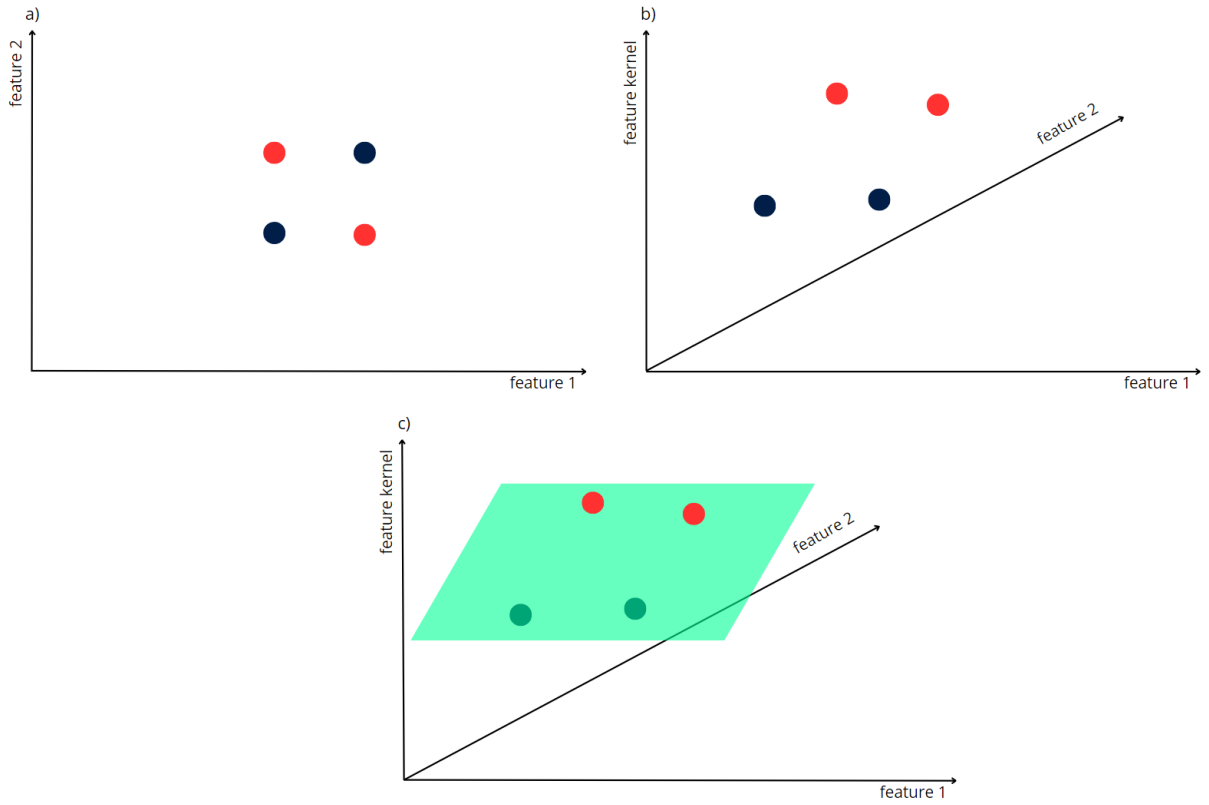


Figura 46 – Exemplo da aplicação de um kernel. Em a) temos quatro pontos de duas classe (azul e vermelha) que não podem ser separados em duas dimensões por uma reta. A situação b) projeta todos esses dados em uma terceira dimensão a partir do uso de um kernel, e a consequência disso é que em c) as classes dos pontos são completamente separáveis por um hiperplano.

6.3.2 Regressão Logística (LR)

Embora o algoritmo tenha o termo 'regressão' em seu nome, a regressão logística é utilizada em análise de dados e aprendizado de máquina para modelar a probabilidade de eventos binários [33]. Se temos uma coleção de instâncias $\{\vec{X}_u\}$ associadas a um conjunto de rótulos de previsão $\{y_u\}$ tal que y_u vale 0 ou 1, podemos construir uma matriz $\tilde{X}$ em que cada linha dela é um elemento de $\{\vec{X}_u\}$, e também é possível definir um vetor $\vec{y}$ para o qual cada componente dele é um elemento de $\{y_u\}$. Dessa forma, podemos fazer duas abordagens com esses dados: linear ou logística. Um modelo de regressão linear pode ser montado a partir desses dados como [33]

$$\vec{y} = \tilde{X}\vec{\beta} + \vec{\epsilon}, \quad (6.15)$$

tal que $\vec{\beta}$ é um vetor de parâmetros para fazer ajustes diretamente em $\tilde{X}$ e $\vec{\epsilon}$ é o vetor de erro com um conjunto de componentes $\{\epsilon_u\}$ relacionadas a cada instância u . No contexto de classificação, a regressão linear pode não ser muito adequada, pois ela pode gerar valores

de previsões menores do que 0 e maiores do que 1, e isso dificulta em uma interpretação probabilística [33].

Uma alternativa à regressão linear é usar regressão logística, pois, ao contrário da regressão linear, a LR aplica a função logística para garantir que as previsões fiquem entre os limites de 0 a 1, permitindo que os resultados possam ser interpretados como probabilidades [33]. A fórmula básica da função logística é

$$P(y_u = 1|\vec{X}_u) = \frac{1}{1 + e^{-\vec{X}_u\beta}}, \quad (6.16)$$

tal que $P(y_u = 1|\vec{X}_u)$ é a probabilidade da instância $\vec{X}_u$ ser classificada como sendo da classe 1.

A figura 47 mostra a intuição da regressão logística, em que um conjunto de dados negativos e positivos (vermelhos e azuis) são avaliados a partir de uma *feature* no eixo x , e no eixo y temos a probabilidade positiva associada a cada ponto. Na situação a), temos uma reta feita a partir de regressão linear. Podemos reparar que a reta não satisfaz as condições de probabilidade, pois ao prolongá-la os dados positivos que estão à direita dela terão probabilidade maior do que 1, e os dados negativos que estão à esquerda terão probabilidade menor do que 0. Em b) e c) temos a torção dessa reta formando curvas que já começam a se aproximar dos pontos de cada classe, fornecendo um melhor ajuste. Então, na situação d) temos a aplicação de uma curva formada a partir da regressão logística, em que a curva já toca vários pontos e ainda respeita os limites para o uso da interpretação de probabilidade.

De forma geral, um modelo de LR tenta gerar a curva que passa pela maior quantidade possível de pontos positivos e negativos. Porém, isso deve ser feito de uma maneira cautelosa, pois uma curva que passe por muitos pontos de maneira inadequada pode gerar um modelo com *overfitting*, e uma curva que passe por poucos pontos será propensa ao *underfitting*.

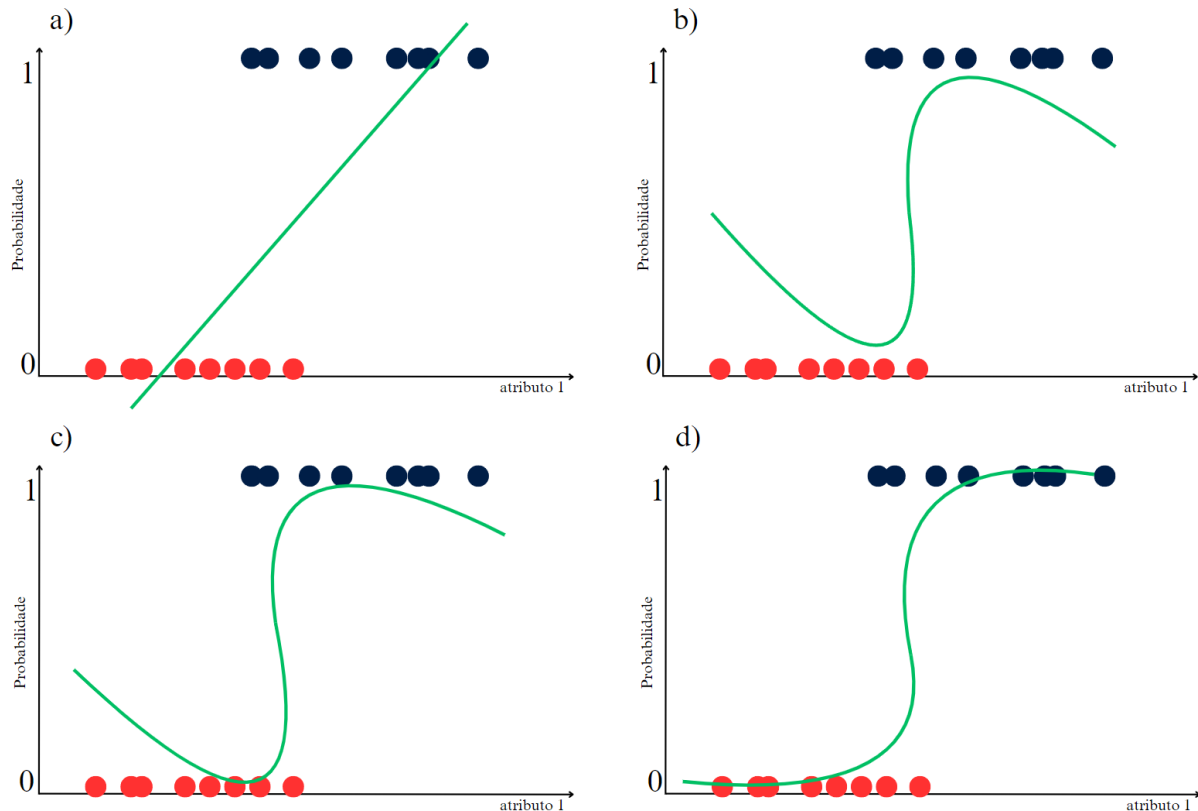


Figura 47 – Exemplo de ajuste a dados pelo algoritmo de regressão logística. Em a) temos uma reta feita a partir de regressão linear. Na situação b) a curva é gerada por meio de uma torção da reta de a) e começa a se aproximar cada vez mais dos pontos de cada classe. Na passagem de c) para d) temos a curva que melhor se ajusta a esses dados.

6.3.3 Perceptron e Rede Neural (*Neural Network*)

A rede neural é um dos mais populares algoritmos de aprendizado de máquina supervisionado e com frequência é citada e usada devido à sua capacidade de processar grandes volumes de dados. As redes neurais modernas possuem origem em um algoritmo chamado perceptron [1], e um conjunto de perceptrons alinhados forma uma grande rede, caracterizando o nome de multilayer perceptron, que é uma outra nomenclatura para se referir à rede neural. Portanto, para explorar o funcionamento de uma rede neural é recomendável compreender a atuação do perceptron. A figura 48 mostra duas estruturas básicas para o entendimento das redes neurais, tal que 48a) é um perceptron e 48b) é uma rede neural com uma camada oculta. Na figura 48a) temos os neurônios da camada de entrada, os pesos e um neurônio na camada de saída. Cada neurônio é representado por um círculo e é responsável por receber os dados de entrada. As setas caracterizam os pesos da rede, que são aplicados nos dados de entrada e posteriormente a interação entre pesos e dados de entrada fornece um resultado que é processado pelo neurônio de saída. Então, os perceptrons são considerados redes neurais com apenas uma camada de entrada

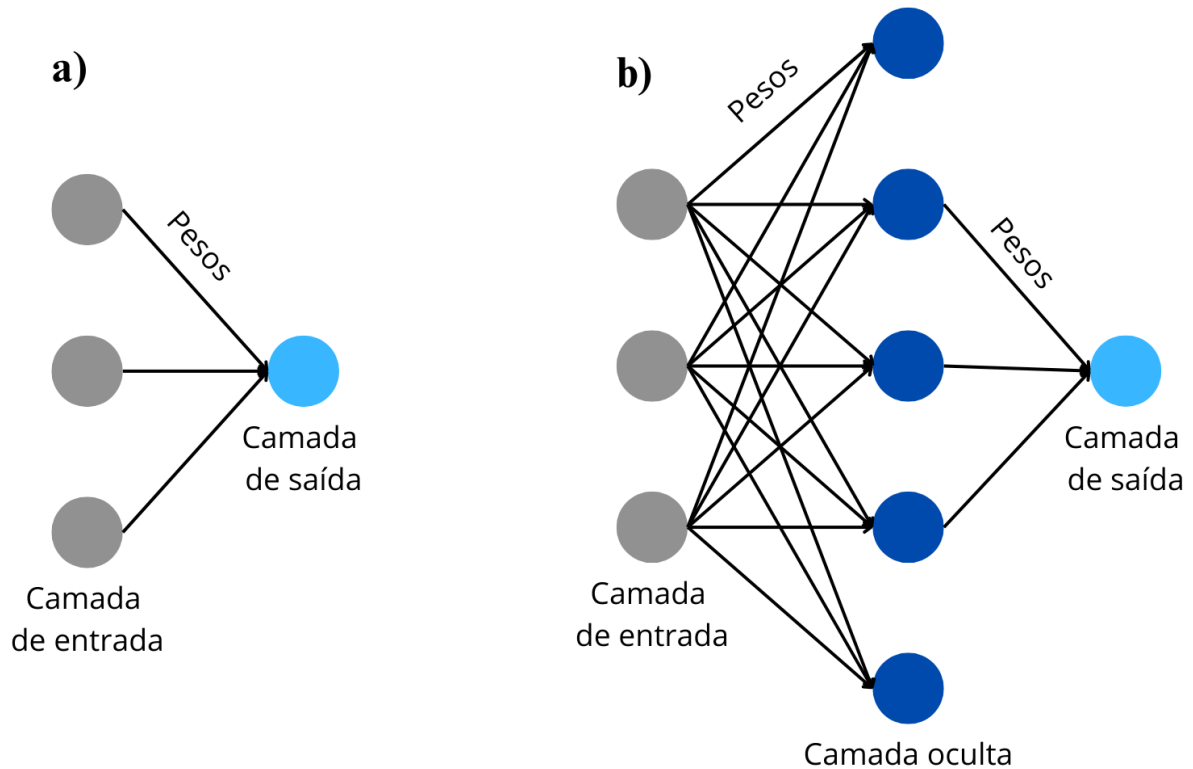


Figura 48 – Diferentes tipos de redes neurais. a) O Perceptron é uma rede neural com apenas uma camada de entrada e uma de saída. b) Uma rede neural com uma camada de neurônios adicional é outra estrutura possível para o algoritmo.

e saída. No caso da figura 48b) temos uma camada a mais em relação ao perceptron, que é chamada de camada oculta. Uma outra diferença é que entre a camada de entrada e a oculta há uma quantidade maior de pesos que ligam os neurônios de cada uma dessas camadas, e isso será explorado mais adiante.

As redes neurais funcionam com ajustes de pesos entre as camadas com base nos dados que recebe e então são capazes de realizar tarefas de classificação a partir disso, embora haja outras aplicações. As principais características que exploraremos da rede neural são: camadas, pesos, método de solução (solver), função de ativação, taxa de aprendizado (r) e momento (p). Vamos discutir primeiramente sobre a rede neural com apenas uma camada de entrada e saída (perceptron) e depois generalizar esses resultados para redes com camadas adicionais. É importante pontuar que todas as redes neurais que serão abordadas são aquelas com um único neurônio na camada de saída.

O funcionamento geral do perceptron é baseado em ajustes de pesos para que o modelo seja capaz de aprender a generalizar as informações dos dados de entrada que recebe, que podem ser representados por um vetor $\vec{X}_u = (x_{1u}, x_{2u}, \dots, x_{mu})$ com u sendo o índice da instância e m informando o índice das componentes desse vetor de entrada.

Diferentemente do apresentado anteriormente, agora os índices das instâncias são u e não i , pois a letra i será reservada para operar com tensores, e isso é feito para não causar ambiguidade durante a leitura.

Os pesos de uma rede neural podem ser representados por um vetor $\vec{W}$. A atuação dos pesos $\vec{W}$ em uma instância $\vec{X}_u$ é dada por

$$\vec{W}^T \vec{X}_u = v_u, \quad (6.17)$$

em que v_u é um escalar resultante da operação da equação (6.17). Esse resultado passa por uma função de ativação $\phi(v)$, que é responsável por processar esse valor e fornecer a informação de saída da rede na forma

$$\phi(v_u) = y_u. \quad (6.18)$$

Uma das funções de ativação mais usadas é a sigmoide (σ), que é dada por

$$\sigma(v_u) = \frac{1}{1 + e^{-v_u}}. \quad (6.19)$$

Para um dado escalar v_u , essa função transforma esse valor em 0 se $v_u < 0.5$ ou 1 se $v_u > 0.5$. Então, se aplicamos a função sigmoide a um neurônio representado pela expressão (6.17), temos [1]

$$\sigma(\vec{W}^T \vec{X}_u) = \begin{cases} 0, & \text{se } (\vec{W}^T \vec{X}_u) < 0.5 \\ 1, & \text{se } (\vec{W}^T \vec{X}_u) > 0.5. \end{cases} \quad (6.20)$$

Além da sigmoide, há outras opções para função de ativação como: tangente hiperbólica ($\tanh$), função unitária linear retificada (relu) e a identidade ($f(v_u) = v_u$).

Os pesos em $\vec{W}$ podem ser atualizados para construir um modelo de rede neural que seja mais eficiente no processo de classificação. Essa atualização de pesos é feita a partir da minimização de uma função de perda. Supondo que $L(\vec{W})$ seja uma função de perda de erro quadrático, temos

$$L(\vec{W}) = \frac{1}{2} \sum_i^n (\sigma(\vec{W}^T \vec{X}_u) - y_u)^2. \quad (6.21)$$

Para minimizar ao máximo os erros da rede neural, devemos tomar o gradiente $\nabla \cdot L(\vec{W})$, que minimiza a função de perda $L(\vec{W})$ com relação aos pesos $\vec{W}$. Aplicando o gradiente em 6.21, podemos escrever [1]

$$\nabla \cdot L(\vec{W}) = \sum_u^n [\sigma(\vec{W}^T \vec{X}_u) - y_u] \sigma(\vec{W}^T \vec{X}_u) (1 - \sigma(\vec{W}^T \vec{X}_u)) \vec{X}_u. \quad (6.22)$$

Se usarmos agora a expressão (6.22), é possível atualizar os pesos $\vec{W}$ da seguinte maneira [1]:

$$\vec{W} \rightarrow \vec{W} - r \sum_u^n [\sigma(\vec{W}^T \vec{X}_u) - y_u] \sigma(\vec{W}^T \vec{X}_u) (1 - \sigma(\vec{W}^T \vec{X}_u)) \vec{X}_u, \quad (6.23)$$

em que r é a taxa de aprendizado. Então, expressão (6.23) pode ser escrita de modo mais simplificado como

$$\vec{W} \rightarrow \vec{W} - r \left(\sum_u^n \text{Erro}_u \cdot \dot{\sigma}(\vec{W}^T \vec{X}_u) \vec{X}_u \right), \quad (6.24)$$

tal que $\text{Erro}_u = [\sigma(\vec{W}^T \vec{X}_u) - y_u]$.

Esse processo de apresentar um conjunto de dados de entrada $\{\vec{X}_u\}$ e atualizar os pesos $\vec{W}$ pode ser feito até que uma condição específica seja atingida como, por exemplo, o erro atingir um certo valor que seja considerável aceitável por aquele que construiu a rede neural. Porém, atingir esse valor não é uma tarefa trivial, e a taxa de aprendizado auxilia nessa tarefa. O valor de r deve ser tal que proporcione uma convergência para essa condição desejada. Como estamos usando a função de erro $L(\vec{W})$ como exemplo, se durante o treino da rede neural o erro converge para um valor desejado, significa que a taxa de aprendizado está adequada, mas se a convergência não ocorre é uma indicação de que o valor de r está grande demais [1]. Um terceiro caso ocorre quando a taxa de aprendizado é muito pequena, e a consequência disso é que a convergência é possível matematicamente, mas o tempo de espera para isso pode ser longo. Uma maneira de mitigar esse problema com o tamanho da taxa de aprendizado é fazendo um ajuste dela ao longo do treinamento do modelo.

Tão importante quanto a modificação dos pesos é a escolha de como eles serão modificados, e vamos aqui explorar três modos de fazer isso: descida de gradiente (GD), descida de gradiente estocástica (SGD) e mini-batch SGD. Essa seleção fornece uma intuição sobre o processo, embora esses modos não sejam as únicas maneiras de fazer isso. O GD atualiza os pesos $\vec{W}$ após processar todo o conjunto de entrada de dados $\{\vec{X}_u\}$. Para o mini-batch SGD a atualização dos pesos acontece a cada n entradas vistas no conjunto $\{\vec{X}_u\}$, e podemos generalizar a atualização dos pesos como [1]

$$\vec{W} \rightarrow \vec{W} - \frac{r}{b} \sum_i^b \text{Erro}_u \cdot \sigma'(\vec{W}^T \vec{X}_u) \vec{X}_u. \quad (6.25)$$

Então, se $n = 10$, os pesos da rede serão modificados a cada 10 entradas processadas pelo algoritmo. Por fim, o SGD promove um ajuste dos pesos feito a cada vez que o algoritmo processa uma entrada $\vec{X}_u$, e isso faz com que o treinamento seja mais demorado quando comparado aos outros métodos. O SGD de certa forma pode ser entendido como um mini-batch SGD para $n = 1$.

Tudo isso que foi desenvolvido e apresentado pode ser generalizado, de maneira cautelosa, para redes neurais com mais de uma camada oculta. Porém, é importante discutir o porquê de se incrementar mais camadas em uma rede neural. O número de camadas de uma rede neural é um hiperparâmetro que tem relação com a complexidade que um modelo pode ter e revela se ele tem baixa ou alta capacidade [1]. Redes com apenas uma camada são capazes de construir apenas hiperplanos lineares, enquanto redes

com duas camadas podem construir qualquer função contínua desde que haja o número de neurônios ocultos adequados para isso. No caso de redes com três ou mais camadas, elas também podem aproximar qualquer função contínua e tendem a ser mais rápidas e com menor tendência ao *overfitting*. Dependendo da tarefa de classificação que um algoritmo precisa exercer é desejado que ele possa alcançar funções $f(\vec{X})$ robustas para que o modelo seja capaz de fazer um ajuste adequado nos dados, pois modelos com baixa capacidade são incapazes de resolver tarefas complexas [1]. Porém, aumentar o número de camadas não é o suficiente para criar um bom modelo, pois é preciso também considerar o número de neurônios que cada camada deve ter, que é um outro hiperparâmetro.

De forma geral, o número de neurônios também tem relação com a complexidade da função $f(\vec{X})$ que o modelo atinge, então o balanço entre esses hiperparâmetros é um equilíbrio a ser respeitado e testado, e o número alto ou baixo desses hiperparâmetros depende do quão complexa deve ser a função $f(\vec{X})$ desejada. Se essa função é muito complexa, é provável que ela demande um número grande de neurônios e camadas. Nesse caso, é preciso também uma grande quantidade de dados, pois um modelo com muitos neurônios usa um grande número de *features*, e nesse caso uma baixa quantidade de dados pode não permitir que o algoritmo capte os padrões necessários para fazer uma boa tarefa de classificação, resultando em um *underfitting* [1]. Por outro lado, se a tarefa a ser feita é de baixa complexidade, usar uma estrutura com muitos neurônios e camadas pode gerar um modelo muito ajustado a apenas os dados que ele viu durante o treinamento, gerando condições favoráveis para um *overfitting* [1].

Apesar de já ter sido apresentado um conjunto de maneiras de como representar instâncias, valores e pesos, as redes neurais demandam alguns cuidados adicionais na notação devido à complexidade estrutural que elas têm quando comparadas ao perceptron. A notação e formalismo que serão apresentados a seguir foram elaborados e adaptados para o contexto deste trabalho, e não representam necessariamente a única forma de construir a álgebra das redes neurais. Dessa forma, abordaremos o funcionamento da rede neural juntamente com a álgebra necessária para compreender o algoritmo.

As redes neurais com múltiplas camadas também possuem um conjunto de pesos. Para o perceptron, temos uma rede neural com apenas uma camada de entrada e saída, e todos os pesos são armazenados em um vetor $\vec{W}$. Porém, para uma rede de múltiplas camadas temos mais de um vetor com pesos, e eles podem ser armazenados em um tensor $\overleftrightarrow{W}$. Para esse tensor, $\overleftrightarrow{W}_{ijk} = w_{ijk}$ representa um peso do tensor, em que k é o índice referente à camada, e i e j indicam, respectivamente, a linha e coluna para uma camada k . Em contexto de redes neurais, i e j possuem significados adicionais além de indicarem linhas e colunas. O índice i representa o neurônio de uma camada $k + 1$ que recebe os pesos que saem dos neurônios da camada k , em que cada neurônio dessa camada k é identificado por um índice j distinto. O conjunto de neurônios das redes neurais pode ser representado

como $\{h_{jk}\}$, em que j nesse contexto também é o identificador de um neurônio da camada k . Essa notação indica que os pesos que saem de uma camada k até uma camada $k + 1$ são designados por $\overleftrightarrow{W}_{::k}$. Por definição, uma rede com n camadas não possui pesos saindo da última camada, pois não existe uma camada $n + 1$. De modo semelhante ao perceptron, uma rede neural recebe uma instância $\vec{X}_u$, que é então processada pelos pesos e funções de ativação da rede neural até que a última camada devolva uma saída y_u . Porém, como a rede neural possui camadas adicionais em relação ao perceptron, é preciso saber indicar todos os valores que entram e saem de um determinado neurônio. O valor que chega a um neurônio $\{h_{jk}\}$ é representado por x_{ujk} , em que u é o índice da instância do vetor de entrada $\vec{X}_u$, o indicador i é o número do neurônio que recebe o valor e k é a camada em que ele se encontra. De modo semelhante, o valor de saída do neurônio h_{jk} é dado por y_{ujk} . A relação entre o valor de entrada v_{ujk} e o valor de saída y_{ujk} do neurônio $\{h_{jk}\}$ é mediada pela função de ativação ϕ :

$$\phi(v_{ujk}) = y_{ujk}. \quad (6.26)$$

Antes de prosseguir, precisamos apresentar uma notação alternativa para as instâncias, pois redes neurais demandam um grande número de índices, e a notação que usamos até o momento para as instâncias pode se tornar ambígua neste contexto. Então, a partir de agora uma instância de índice u não será mais representada como $\vec{X}_u = (x_{1u}, x_{2u}, \dots, x_{mu})$, mas sim como $\vec{X}^{(u)} = (x_1^{(u)}, x_2^{(u)}, \dots, x_m^{(u)})$.

Como uma instância $\vec{X}^{(u)}$ é processada pela rede de forma que surgem vários valores de entradas e saídas em neurônios a partir dessa instância, o conjunto de todas as entradas e saídas de neurônios gerados a partir de $\vec{X}^{(u)}$ são representados por, respectivamente, $\{v_{jk}^{(u)}\}$ e $\{y_{jk}^{(u)}\}$. Essa notação apresentada funciona bem desde que sejam estabelecidas algumas condições para as camadas de entrada e saída da rede. Por definição, para a camada de entrada da rede temos $\{v_{j1}^{(u)}\} = \{y_{j1}^{(u)}\}$. Para uma rede de n camadas, podemos escrever de modo conveniente que $y_{jn}^{(u)} = y^{(u)}$, em que essas duas formas representam a última saída da rede para uma instância $\vec{X}^{(u)}$. Se queremos descobrir qual é o valor de entrada $x_{a(b+1)}^{(u)}$ em um neurônio $h_{a(b+1)}$ a partir dos neurônios da camada (b) e de uma entrada $\vec{V}_b^{(u)}$, podemos usar a expressão geral

$$x_{a(b+1)}^{(u)} = \overleftrightarrow{W}_{a:b} \vec{V}_b^{(u)}, \quad (6.27)$$

em que $\overleftrightarrow{W}_{a:b}$ são os pesos de todos os neurônios da camada b conectados ao neurônio a da camada $(b + 1)$.

Podemos atualizar os pesos da rede durante o treinamento para construir um modelo que seja mais eficiente no processo de classificação. Para fazer isso, é preciso calcular os gradientes locais, que são os valores das derivadas das funções de ativação de cada neurônio para as entradas nesses neurônios. A reformulação dos pesos pelos gradientes

locais é chamada de retropropagação (*backpropagation*) [1]. Esse método de atualização de pesos por si só já é extenso e não cabe demonstrá-lo em detalhes neste trabalho. A figura 49 mostra um esquema geral de uma rede neural com *backpropagation*. Para cada instância $\vec{X}^{(u)}$ devemos calcular a previsão $f(\vec{X}^{(u)})$ e, se $f(\vec{X}^{(u)}) \neq Y^{(u)}$, o erro deve ser propagado para ajustar os pesos [1]. Esse processo deve ocorrer por um número suficiente de épocas até que se atinja alguma condição de parada, que é estabelecida antes de iniciar o treinamento da rede.

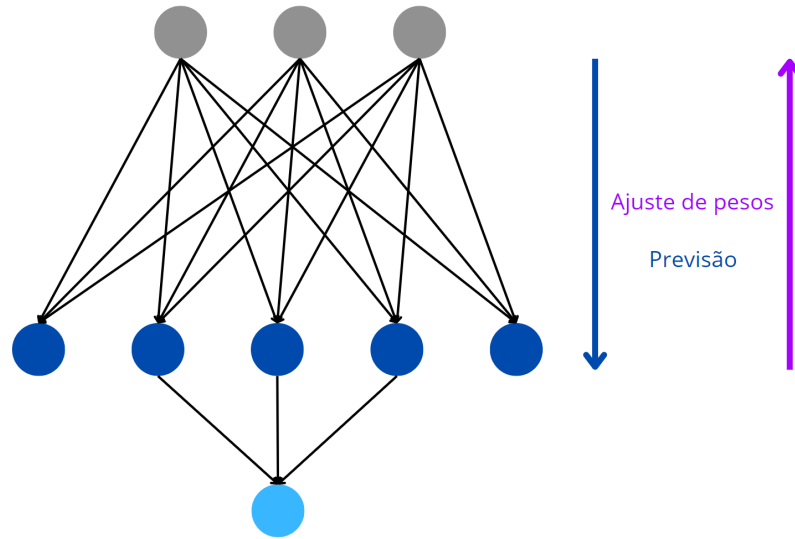


Figura 49 – Esquematização de uma rede neural com *backpropagation* a partir de dois processos: previsão de $f(\vec{X}^{(u)})$ a partir de uma instância $\vec{X}^{(u)}$ e ajuste dos pesos caso $f(\vec{X}^{(u)}) \neq Y^{(u)}$.

6.3.4 Árvore de decisão e Extreme Gradient Boosting

No contexto de aprendizado de máquina é desejável que o modelo produzido possa ser interpretável com relação aos dados que ele trabalha. Isso significa que é interessante entender como um modelo utiliza as *features* dos dados de entrada e como cada uma dessas *features* pode nos fornecer uma interpretação maior para compreender o problema que está sendo estudado. Porém, nem todos os algoritmos de aprendizado são capazes de fornecer uma interpretação intuitiva daquilo que está sendo processado a partir dos dados. É nessa situação que o algoritmo de árvore de decisão tem um destaque considerável, pois a maneira em que ele processa os dados permite uma interpretação de como as *features* são usadas para fazer a discriminação dos dados a partir de suas classes.

O processo de classificação feito pelas árvores de decisão é baseado em projetar os dados em um espaço usando suas *features* como eixos e depois separando cada classe por meio de cortes (intervalos) no espaço. A figura 50a) mostra um conjunto de pontos de duas classes projetados em um espaço com duas *features* (eixos horizontal e vertical).

Podemos separar esses pontos em classes ao colocar uma linha vertical tracejada no valor de p da *feature 1* (eixo horizontal). Dessa forma, todos os pontos azuis ficam confinados à esquerda e todos os pontos vermelhos ficam isolados à direita. Apesar de termos duas *features*, os dados são completamente separáveis utilizando apenas a *feature 1*, e isso significa que a projeção desses dados no eixo da *feature 1*, conforme mostrado na figura 50b), já resolve o problema com o uso de apenas uma dimensão. Para a figura 50c) temos a árvore de decisão que representa esse problema, em que os pontos são enviados para verificar se eles possuem um valor de *feature 1* menor ou maior do que p . Se o valor da *feature 1* de um dado é menor do que p , então o ponto pertence à classe azul, mas se a *feature 1* é maior do que p o ponto é classificado como vermelho. Na figura 50c) o bloco marrom é chamado de raiz ou nó raiz, as setas são os galhos e os dados ao fim dos galhos são chamados de folhas ou folhas nó, que são os blocos verdes. Na figura 51 temos a

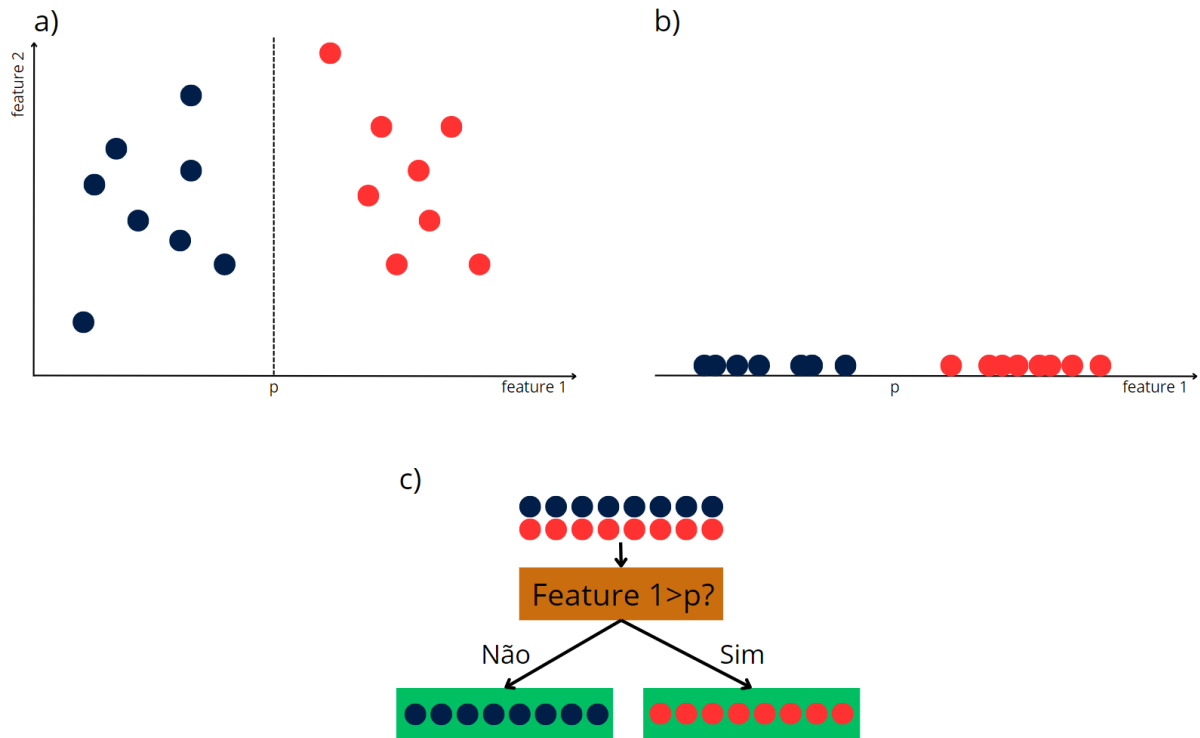


Figura 50 – a) Os pontos de duas classes diferentes (azul e vermelha) podem ser separados completamente a partir de uma linha vertical tracejada feita no valor p da *feature 1*. b) O problema pode ser projetado em apenas uma dimensão e mesmo assim separa perfeitamente os pontos de cada classe a partir do valor p da *feature 1*. Em c) temos esse problema representado em forma de uma árvore de decisão.

estrutura básica de uma árvore de decisão. Os dados entram em cada nó e, com base em alguma condição contida em cada nó, eles são direcionados pelos galhos para outros nós ou folhas. Se um conjunto de dados foi direcionado a outro nó, significa que os dados

desse conjunto não tiveram nenhuma separação completa de classes. Caso um grupo de dados seja direcionado para uma folha, temos duas opções: ou parte desses dados foi completamente separada e temos uma folha que contém dados de apenas uma classe ou então a árvore de decisão foi podada naquele ponto por escolha de quem criou a árvore. Nesse caso, a poda da árvore pode ser uma estratégia para atender a alguma condição que o criador da árvore julga adequada, mesmo que a folha gerada ainda contenha dados com mais de uma classe [32]. Cada nó vai conter alguma condição relacionada a alguma *feature* dos dados, e os galhos são responsáveis por direcionar os dados de acordo com a condição proposta. A construção de uma árvore depende totalmente da ordem que os atributos são

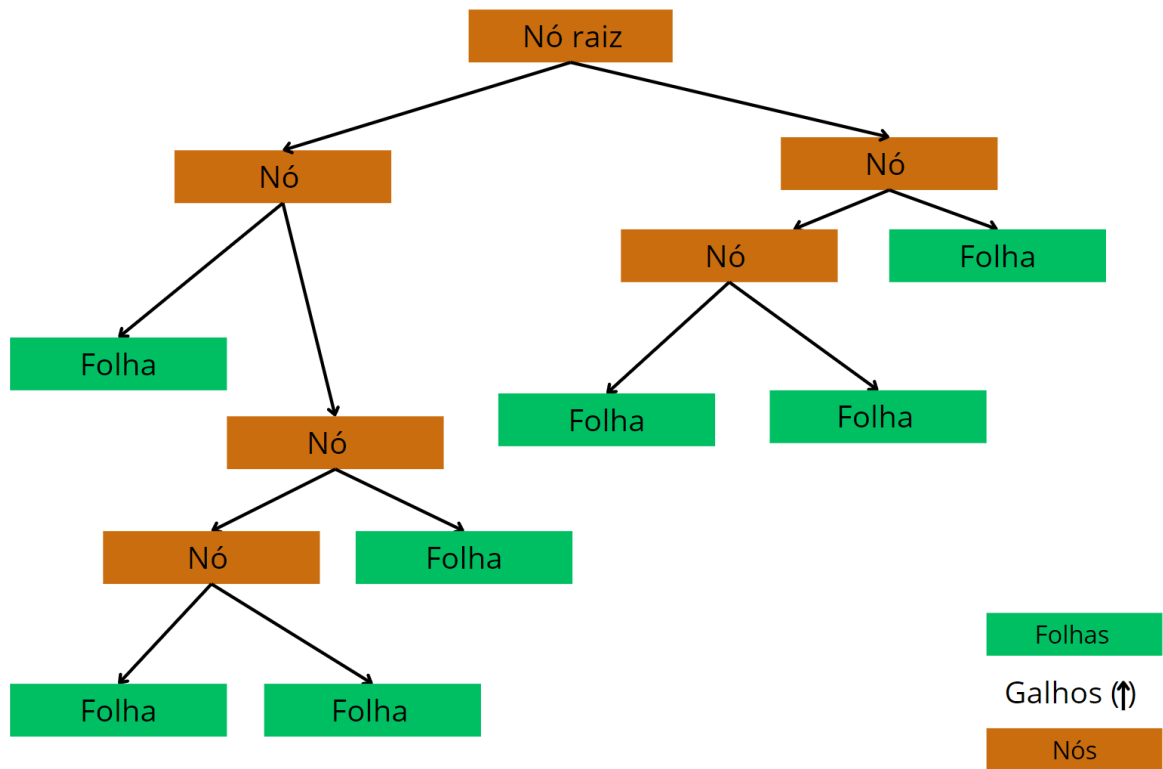


Figura 51 – Estrutura de uma árvore de decisão com folhas, galhos e nós.

colocados nas árvores, mas como escolher a melhor ordenação dos atributos? Para essa tarefa, temos que selecionar primeiro sempre aquele atributo que melhor divide os dados após cada galho. Quanto menor a porcentagem de mistura de classes após cada galho, melhor será o atributo relacionado a essa divisão de dados. Chamaremos de entropia H o grau de desordem e mistura que um atributo causa ao tentar dividir um grupo de dados. Quanto maior a entropia H ocasionada por um atributo, pior ele é para separar as classes. A entropia H de um nó pode ser calculada como [32]:

$$H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log \frac{p}{p+n} - \frac{n}{p+n} \log \frac{n}{p+n}, \quad (6.28)$$

tal que p é o número de exemplos da classe positiva e n é a quantidade de membros da classe negativa. Se a partir de um atributo a nós conseguimos formar m nós, a entropia esperada (EH) dada pela soma de m entropias geradas a partir desse atributo é [32]:

$$EH(a) = \sum_{i=1}^m \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right), \quad (6.29)$$

em que p_i e n_i são, respectivamente, o número de exemplos positivos e negativos para o nó i , e temos p e n como sendo a quantidade de positivos e negativos para o nó raiz respectivamente.

Podemos usar agora a equação (6.29) para calcular o ganho de informação (I) que obtemos ao escolher um atributo para tentar separar os dados. O ganho de informação pode ser entendido como uma medida do quanto a entropia de um nó diminui quando ele é dividido em m nós por um atributo a . Então, o ganho de informação para um atributo a pode ser escrito como

$$I(a) = H\left(\frac{p}{p + n}, \frac{n}{p + n}\right) - EH(a). \quad (6.30)$$

Como exemplo de cálculo de entropia, usaremos a figura 52 com dezesseis dados igualmente divididos entre positivos (azuis) e negativos (vermelhos). Temos a situação a) em que os dados são separados usando o atributo 1 e em b) eles são separados usando o atributo 2. A partir da expressão (6.28), a entropia no nó raiz é dada por:

$$H_{raiz} = -0.5 \log(0.5) - 0.5 \log(0.5). \quad (6.31)$$

Usando as equações 6.28 e 6.29, temos que a entropia esperada em cada caso é dada por

$$EA(\text{atributo 1}) = \frac{4}{16} H(1, 0) + \frac{6}{16} H(0, 1) + \frac{6}{16} H\left(\frac{4}{6}, \frac{2}{6}\right) = 0.11 \quad (6.32a)$$

$$EA(\text{atributo 2}) = \frac{4}{16} H(0.5, 0.5) + \frac{6}{16} H(0.5, 0.5) + \frac{6}{16} H(0.5, 0.5) = 0.30. \quad (6.32b)$$

Usando os valores de entropia esperada na expressão (6.30) juntamente com H_{raiz} , o ganho de informação em cada cenário é

$$I(\text{atributo 1}) = H_{raiz} - EH(\text{atributo 1}) = 0.30 - 0.11 = 0.19 \quad (6.33a)$$

$$I(\text{atributo 2}) = H_{raiz} - EH(\text{atributo 2}) = 0.30 - 0.30 = 0. \quad (6.33b)$$

Portanto, podemos ver que o atributo 1 retorna um maior ganho de informação, sendo melhor para separar os dados em classes quando comparado com o atributo 2, que tem um ganho de informação nulo. Na figura 52b), é possível ver que para o atributo 2 os dados foram enviados para outros nós, mas a proporção de dados positivos e negativos em cada um desses nós é a mesma do que a proporção no nó de origem, significando que de fato não houve nenhum ganho de informação para realizar uma separação adequada dos dados. Se adicionarmos novos atributos na figura 52 depois do atributo 1 ou 2, nós estaremos

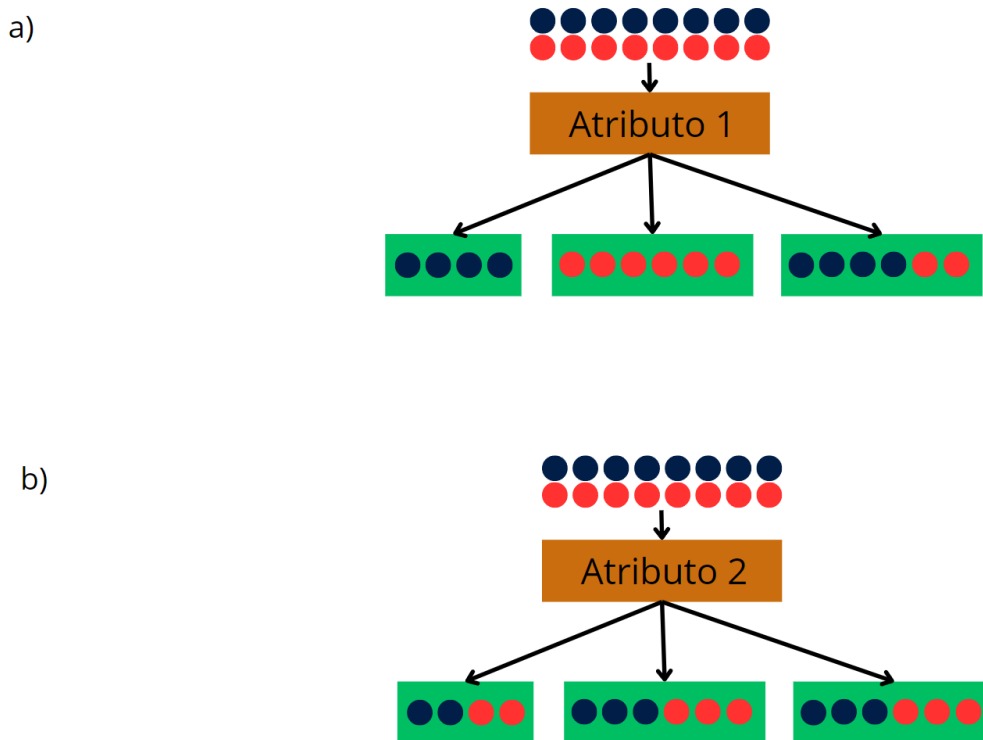


Figura 52 – Duas árvores de decisão construídas a partir de dois atributos diferentes para o nó raiz. Em a) temos uma entropia total menor, portanto um maior ganho de informação. Já em b) o ganho de informação é zero.

aumentando o tamanho ou profundidade (*depth*) dessas árvores. O tamanho de uma árvore é diretamente proporcional ao número de nós [32] e, portanto, há uma relação com a quantidade de *features*. Ao aumentarmos o tamanho de uma árvore de decisão estamos também aumentando sua capacidade e, portanto, o modelo que tem uma árvore grande é capaz de alcançar funções $f(\vec{X})$ mais complexas, e isso é desejável pois pode ser uma boa alternativa para melhorar o desempenho de um modelo. Na prática, isso significa que uma árvore tão complexa quanto se queira pode ter em seus nós finais os dados totalmente separados em suas respectivas classes. Porém, como discutido anteriormente, o aumento da complexidade de um modelo pode levar ao *overfitting*, então é preciso usar técnicas para regularizar a árvore. Uma das maneiras de fazer isso é podando os nós mais afastados do nó raiz, pois assim o tamanho da árvore diminui, e também sua complexidade.

Uma alternativa para alcançar melhores desempenhos é utilizar mais de um modelo de árvore ao mesmo tempo para processos de classificação. Essa união de árvores é comumente chamada de random forest (floresta aleatória). O algoritmo random forest é uma técnica de aprendizado de máquina que usa a construção de múltiplas árvores de decisão [34]. Cada árvore é treinada em subconjuntos diferentes do banco de dados total, contribuindo para a generalidade do modelo tal que, para tomar uma decisão, a

random forest combina os resultados de todas as árvores utilizadas em seus respectivos processos de classificação [34]. Então, essa abordagem reduz o risco de *overfitting*, pois a decisão final de classificação não depende de uma única árvore [34]. Dessa forma, seu poder de generalização a torna uma ferramenta amplamente utilizada, pois pode oferecer um bom desempenho em dados de teste [34]. Esse algoritmo de agregação de árvores é particularmente útil em problemas que envolvem muitas variáveis, pois o caráter aleatório na seleção das amostras para a criação de cada árvore ajuda a aumentar a diversidade e a precisão do modelo [34].

A figura 53 demonstra o funcionamento simplificado de uma random forest. O banco de dados é dividido em quatro partes: três conjuntos distintos para treinar cada árvore e um para testá-las. Como os conjuntos dos dados de treinamento são diferentes entre si, os modelos gerados também são. Então, os modelos prontos são usados para fazer previsões usando o conjunto de teste, e as previsões de cada modelo são coletadas e submetidas a alguma métrica como, por exemplo, a média. Então, a previsão final é obtida ao aplicar uma métrica nas probabilidades geradas por cada uma dessas árvores. Apesar da

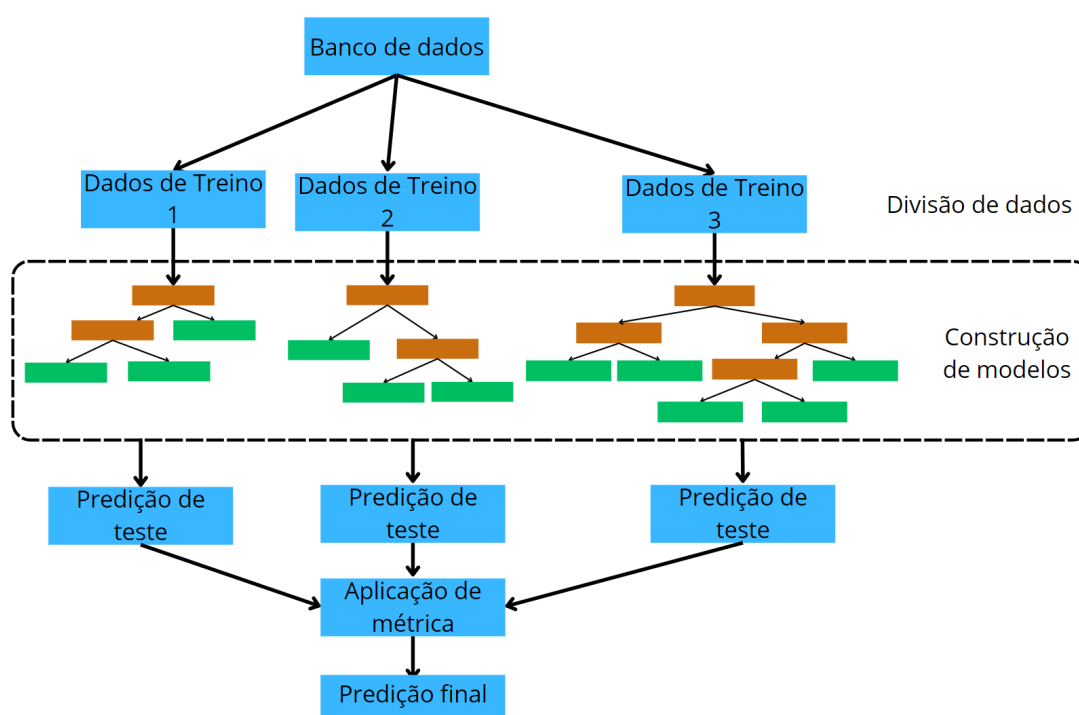


Figura 53 – Esquema simplificado do funcionamento de uma random forest para três árvores.

random forest ser um modelo com uma estrutura que ajuda a evitar o *overfitting*, podemos explorar essa ideia de várias árvores e construir um modelo com ainda menos tendência ao *overfitting*. O que faremos agora é usar a ideia de múltiplas árvores da *random forest*, mas cada árvore será a mais simples possível, que é quando elas apresentam apenas o nó raiz e os nós folhas, assim como a figura 50c). Esse tipo de árvore com apenas um nó é chamada

de *decision stump*, e a ideia dessa proposta é construir um modelo forte a partir da união de classificadores fracos (simples).

Esse algoritmo proposto é chamado de *Extreme Gradient Boosting* (XGBoost), que é uma técnica utilizada em aprendizado de máquina supervisionado, desenvolvida para melhorar a precisão de modelos baseados em árvores de decisão [35]. Esse algoritmo segue o princípio de combinação de múltiplos modelos fracos (boosting) de forma sequencial para formar um modelo mais robusto [35]. De forma geral, o XGBoost constrói árvores de decisão em sequências, onde cada árvore nova tenta corrigir os erros cometidos pelas árvores anteriores [35]. O algoritmo é baseado na minimização de uma função de perda, que quantifica o erro entre as previsões feitas pelo modelo e os valores reais. Ele utiliza o gradiente descendente para ajustar iterativamente as árvores, de modo que, a cada iteração, as previsões se aproximem cada vez mais dos valores reais [35]. Um dos principais diferenciais do XGBoost é a regularização, que é aplicada diretamente aos pesos das árvores, ajudando a controlar a complexidade do modelo e a evitar o *overfitting*. Isso significa que, ao contrário de outros algoritmos de *boosting*, o XGBoost tem mecanismos internos para garantir que o modelo não se ajuste excessivamente aos dados de treinamento, mantendo sua capacidade de generalização para dados fora do conjunto de treino. Outro ponto importante é que o XGBoost utiliza computação paralela para acelerar o processo de construção das árvores de decisão, tornando-o adequado para conjuntos de dados muito grandes [35].

A figura 54 é uma demonstração simplificada do funcionamento do XGBoost com três árvores do tipo *decision stump*. O banco de dados é usado para treinar a árvore 1, que classifica de forma errada os três dados que estão circulados. Então, o banco de dados é modificado (DM1) de forma que os dados classificados de forma errada na etapa anterior agora ganham um peso a mais, significando que a próxima árvore é recompensada ao focar sua tarefa em acertar a classificação desses dados. Na figura 54, os dados que recebem um peso a mais são representados com uma tamanho maior em relação aos outros. Então, os dados de DM1 são usados para treinar a árvore dois, que classifica corretamente os dados errados pela árvore anterior, mas comete três erros. Esses dados classificados de maneira errada recebem pesos para serem classificados de forma correta e formam o conjunto DM2, que é então usado para treinar a árvore 3. Ao fim do processo, temos um classificador que é uma combinação das três árvores que, em princípio, é um modelo melhor do que as três isoladas.

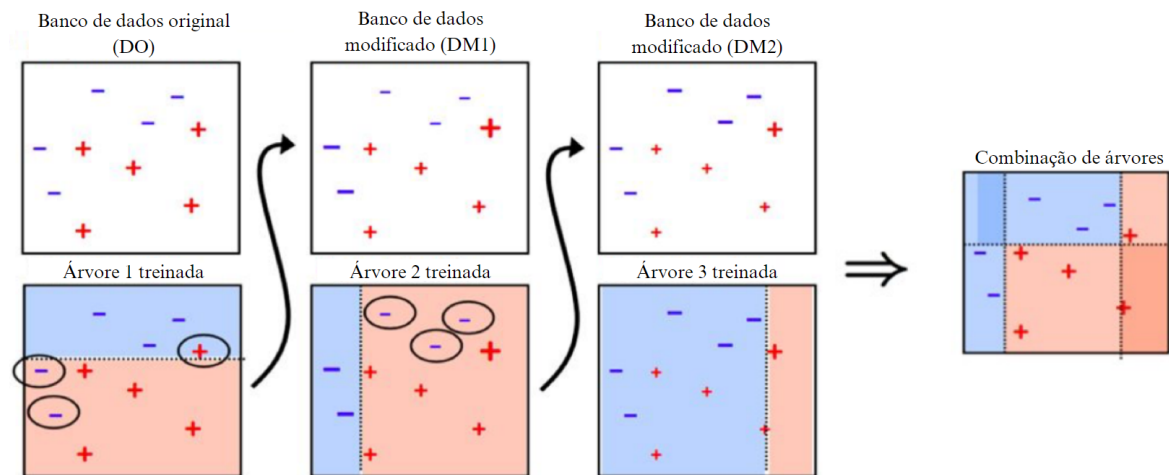


Figura 54 – Estrutura básica de funcionamento de um XGBoost. A partir dos dados originais, uma árvore é gerada, e a árvore seguinte é criada condicionada a não errar os dados que a anterior errou, pois os dados errados recebem um peso maior para serem classificados corretamente. Esse processo ocorre até a última árvore, e, ao fim de tudo, um classificador final é construído a partir da combinação dessas árvores. Adaptado de [36].

6.3.5 *K-Nearest Neighbors* (KNN)

O KNN é um algoritmo com uma interpretação e entendimento com boa acessibilidade, pois ele utiliza conceitos clássicos e de uso cotidiano, como o de distância. Então, esse algoritmo é uma técnica simples e eficaz de aprendizado supervisionado, amplamente utilizada para tarefas de classificação e regressão [24], embora nossa abordagem aqui seja focada na tarefa de classificação. Este método é baseado na ideia de que dados (pontos) próximos em um espaço de atributos (*features*) possuem rótulos semelhantes, o que permite a classificação de novos dados com base na análise deles em relação a seus vizinhos mais próximos.

O KNN não constrói um modelo de aprendizado durante a fase de treinamento da mesma maneira que os modelos anteriores. Em vez disso, o algoritmo armazena o conjunto de dados de treinamento e realiza as previsões dos dados de teste com base nos pontos vizinhos mais próximos. Quando um novo ponto de dados é apresentado, o KNN calcula a distância entre este ponto e todos os pontos no conjunto de treinamento. Há uma variedade de distâncias que podem ser utilizadas, mas a comumente empregada é a Euclidiana. Dependendo das características dos dados, também podemos utilizar as distâncias Manhattan, hamming e Minkowski, sendo a última uma generalização das distâncias Euclidiana e Manhattan [24]. A distância Manhattan, também chamada de City Block Distance, mede a distância entre dois vetores ao calcular a soma das diferenças absolutas entre suas coordenadas [24]. Ela é apropriada para dados que variam em tipos e

unidades, como idade, altura ou gênero, pois ela não depende da magnitude das variáveis. A distância Hamming é uma métrica que calcula a distância entre dois vetores binários [24], ou seja, o número de posições em que os bits são diferentes. Ela é adequada em problemas onde os dados são categóricos ou binários.

Após calcular a distância, o algoritmo seleciona os K vizinhos mais próximos do ponto, ou seja, os K pontos com menor distância em relação ao ponto que se quer classificar. A previsão é feita por votação majoritária entre os vizinhos, tal que a classe mais frequente dos K pontos mais perto é atribuída ao novo ponto [24].

A escolha do valor de K é fundamental para o desempenho do algoritmo, pois um valor muito baixo pode fazer com que o modelo seja sensível a ruídos, já que ele considera a classificação com base apenas nos K vizinhos mais próximos, sem considerar a distribuição dos dados ao redor. Já valores muito altos podem considerar pontos não muito adequados, dificultando a captura de padrões locais eficientes para classificar um ponto de teste. Ajustar o valor de K envolve encontrar um equilíbrio entre o *overfitting*, que ocorre com valores baixos de K , e o *underfitting*, que ocorre com valores altos. A boa prática é testar diferentes valores de K e avaliar o desempenho do modelo para determinar a melhor escolha em um problema específico.

Uma das principais desvantagens do KNN é o seu custo computacional, pois o algoritmo precisa calcular a distância entre o ponto a ser classificado e todos os pontos do conjunto de dados de treinamento. Esse processo é repetido para cada instância de dados, o que pode se tornar muito custoso em bases de dados grandes.

A figura 55 mostra um exemplo de funcionamento do algoritmo KNN para quatro valores diferentes de K . Em um espaço com dois atributos (bidimensional), é projetado um conjunto de dados positivos (azuis), negativos (vermelhos) e um ponto de teste (cinza) para ser classificado a partir de sua distância euclidiana (linha tracejada) em relação aos outros pontos. Em a) temos $K=1$, então devemos considerar apenas o primeiro ponto mais perto do ponto cinza, e o dado mais próximo dele é da classe positiva, então ele é classificado como positivo. Na situação b) temos $K=2$, e para os dois pontos mais próximos do ponto de teste temos duas classes diferentes, então com base apenas nesse valor de K não é possível classificar o ponto. Nesse caso, seria preciso estabelecer um outro critério adicional, e os valores pares para K podem ser um problema por possibilitarem um empate entre as classes. Para c) temos que $K=3$, e o ponto então é classificado como sendo negativo, pois o terceiro ponto a ser considerado é da classe negativa, portanto fica um ponto positivo contra dois pontos negativos. De modo semelhante, isso também é válido para o caso d) com $K=4$, pois agora temos um ponto positivo contra três negativos.

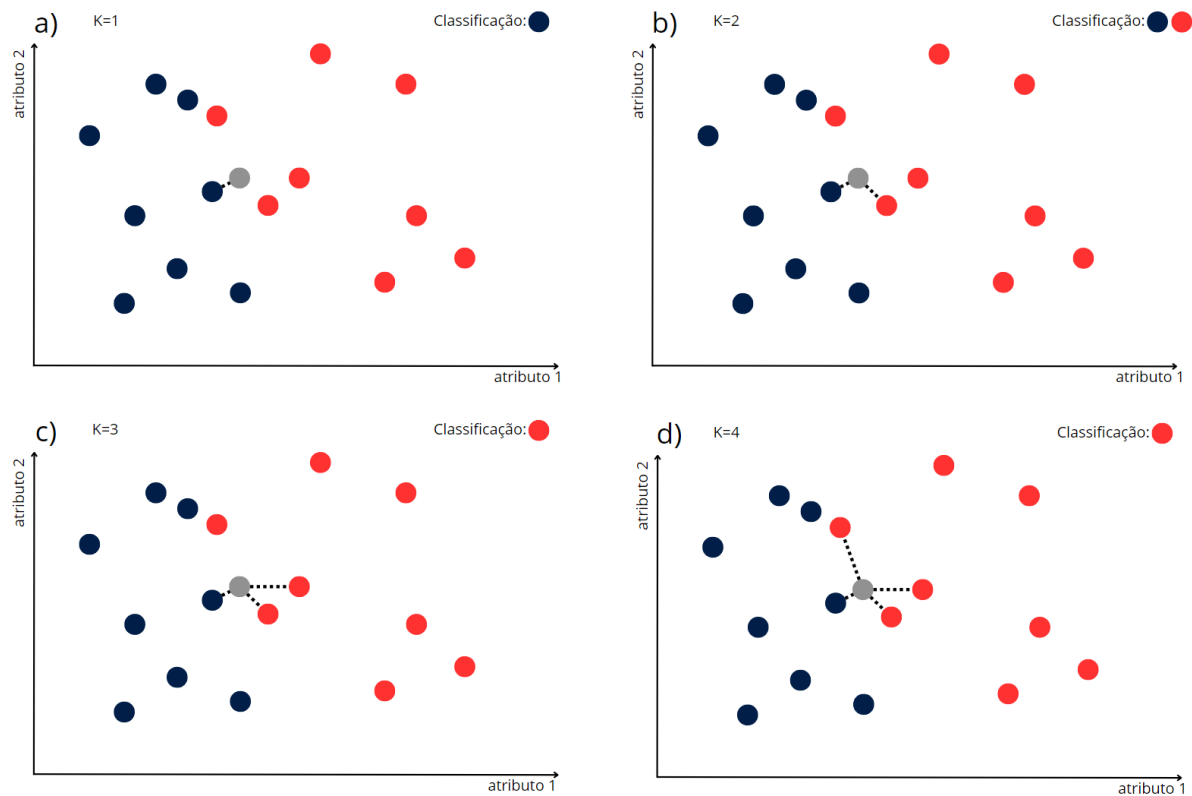


Figura 55 – Esquema do funcionamento do KNN em duas dimensões para um conjunto de dados positivos (azuis), negativos (vermelhos) e um ponto de teste (cinza) para ser classificado com base em distância euclidiana (linha tracejada). Em a) temos $K=1$ e o dado mais próximo do ponto cinza é da classe positiva, então ele é classificado como positivo. A situação b) mostra $K=2$, e os dois pontos mais próximos do ponto de teste são de classes diferentes, então com base apenas nisso não é possível classificar o ponto. Para c) a situação é de $K=3$, e o ponto é classificado como sendo negativo, e o mesmo vale para o caso d) com $K=4$, pois nos dois casos a maioria dos pontos mais próximos do dado de teste é da classe negativa.

7 Diagnósticos clínicos

Ao propor um novo método de diagnóstico clínico por aprendizado de máquina para uma enfermidade, é preciso estabelecer passos metodológicos que sejam objetivos e que possam ser validados desde a medição das amostras até a classificação delas pelo modelo proposto. Para isso, é importante escolher procedimentos e métricas para a avaliação de desempenho do modelo de diagnóstico tais que elas possam refletir resultados precisos e coerentes com a natureza do problema que se quer propor um diagnóstico. Além disso, um novo modelo clínico deve considerar outros fatores além das métricas de precisão, pois é preciso analisar a praticidade e viabilidade de aplicação desse novo método.

7.1 Análise e validação de testes clínicos

No contexto de método de classificação para doenças infecciosas, uma grande parte dos testes clínicos utilizados para chegar a um diagnóstico é feita em amostras *in vitro*, que são derivadas do corpo humano (como sangue, expectoração, tecido). O objetivo desse processo é de fornecer informações sobre o estado atual ou futuro da pessoa que foi submetida ao teste [37]. Para tal, precisamos de testes de diagnósticos que sejam adequados, e para garantir isso podemos propor um conjunto de quatro fases, que estão esquematizadas na figura 56, para verificar a eficácia e segurança dos testes, que englobam desde sua concepção até a sua aplicação em casos clínicos [37].

Na primeira fase temos a discriminação entre grupos, que envolve testar se há uma correlação entre o teste proposto e a condição clínica que se deseja verificar [37]. Essa fase tem objetivo de demonstrar a validade científica do teste, tal que queremos verificar se ele consegue diferenciar de modo adequado os indivíduos doentes dos saudáveis. Nessa etapa, a sensibilidade e especificidade podem ser medidas em grupos de indivíduos doentes e saudáveis. Essa fase inicial busca garantir que a proposta do teste seja viável, fornecendo uma justificativa para seu desenvolvimento posterior [37].

A segunda etapa trabalha com a validade técnica do processo. Após estabelecer a correlação entre o teste proposto e as condições clínicas, o próximo passo é verificar essa validade técnica do teste [37]. Essa etapa envolve garantir que o teste seja reproduzível e confiável, para que ele produza resultados consistentes em diferentes situações. Nesta fase, são analisadas medidas como sensibilidade analítica, precisão e os limites de detecção [37]. Os testes contínuos e categóricos demandam diferentes abordagens para garantir que suas medições sejam adequadas e reproduzíveis [37].

Para a terceira fase é crucial determinar como o teste se comporta em situações

clínicas reais. Nessa etapa são examinadas a sensibilidade diagnóstica, especificidade, valores preditivos positivos e negativos e a viabilidade da aplicação clínica do teste [37]. Essa avaliação considera o contexto clínico de forma geral, como quem será testado e em quais condições o teste será aplicado. Isso significa considerar o público-alvo do teste em situações de triagem e confirmação de diagnóstico, que significa avaliar, respectivamente, indivíduos assintomáticos e sintomáticos. Então, o objetivo dessa fase é garantir que o teste ofereça um valor clínico para precisão diagnóstica [37].

A quarta e última fase de avaliação é concentrada em verificar o impacto do teste sobre a saúde dos pacientes e sobre a sociedade em geral [37]. Um teste pode até se mostrar preciso, mas a questão mais importante e central é se ele realmente é capaz de auxiliar na melhora da condição do paciente ao influenciar decisões de diagnósticos e de tratamento. Estudos clínicos preparados aleatoriamente podem ser necessários para avaliar se o uso do teste revela melhores resultados clínicos, como uma redução em mortalidade ou uma melhoria na qualidade de vida dos pacientes [37]. Portanto, a precisão de um teste

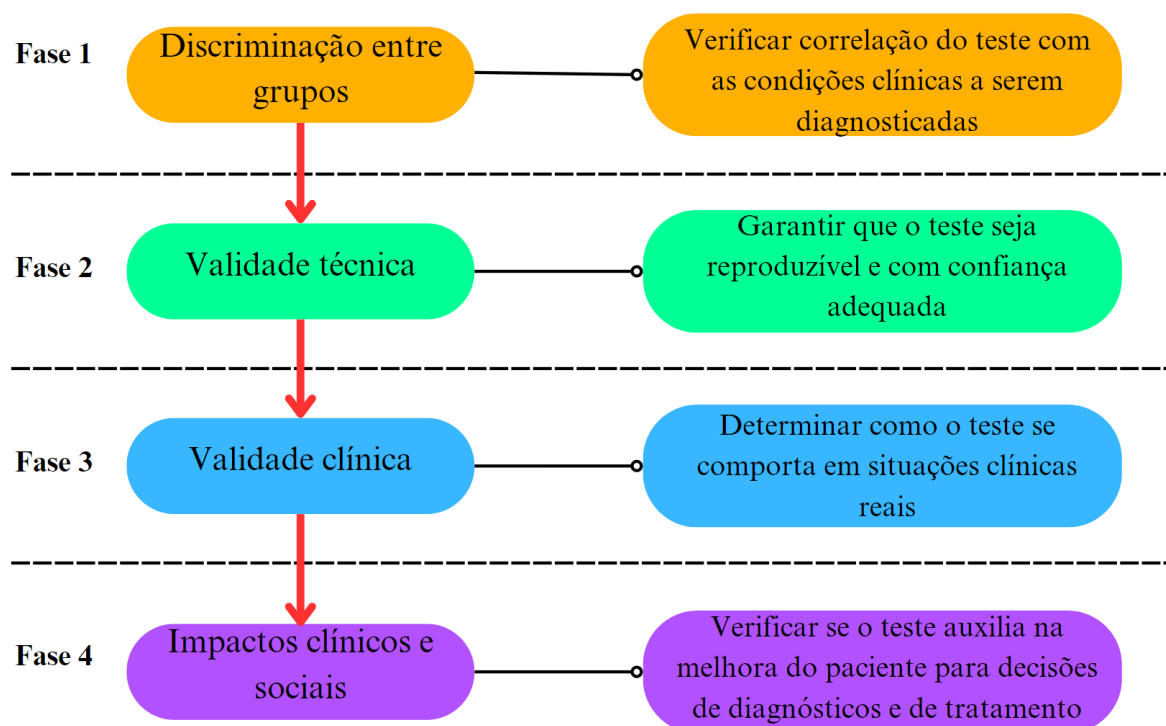


Figura 56 – Esquema das quatro etapas para avaliação de um novo diagnóstico clínico.

diagnóstico é muito importante, mas não é suficiente para indicar se o método é adequado ou não, pois é apenas uma parte do quadro maior. O verdadeiro valor de um teste está em sua capacidade de impactar positivamente os desfechos clínicos e sociais dos indivíduos que estão envolvidos no processo [37]. Logo, é fundamental que esses testes passem por avaliações rigorosas, sistemáticas e bem orquestradas antes de serem amplamente

adotados como boas práticas clínicas. Dessa forma, fica demonstrada a importância de uma abordagem baseada em evidências eficientes na avaliação de novos testes diagnósticos, tal que a precisão é apenas um dos aspectos a serem considerados durante todo o processo.

7.2 Testes clínicos por aprendizado de máquina

É particularmente interessante compreender como essas quatro etapas podem ser usadas em um contexto que envolve diagnóstico usando aprendizado de máquina e inteligência artificial. Embora a análise desse contexto possa ser vista de maneira ampla, aprofundada e por mais de uma abordagem, vamos usar de base os conhecimentos sobre aprendizado de máquina que foram construídos até o momento.

Para a primeira fase, necessitamos de um banco de dados contendo amostras de todos os grupos (classes) que queremos discriminar, tal que as boas práticas indicam que é apropriado que não haja uma diferença muito grande na quantidade de dados entre cada classe. Nesse caso, é preciso manter uma base de dados balanceada entre as classes. Na segunda fase, precisamos garantir a integridade do código ou software que foi criado para conter o modelo de diagnóstico e garantir que o aparelho que utilizará esse modelo seja apropriado para executar suas funções de maneira adequada. Considerando a fase três, o modelo precisa apresentar uma boa garantia de que é funcional e eficiente para avaliar dados que estão fora de sua base de treino. Para tal, são coletadas as informações que foram discutidas no capítulo sobre métricas de desempenho, que servem para avaliar o modelo de diferentes maneiras. Além disso, as técnicas de validação cruzada regular e *nested* são capazes de fornecer boas aproximações para o desempenho do modelo em situações clínicas reais desde que o banco de dados e as *features* utilizadas sejam adequadas para representar bem o problema a ser solucionado. Por fim, a fase quatro exige que o modelo produza um impacto real na recuperação de pacientes a depender de vários fatores. É desejável que modelo feito por aprendizado de máquina seja eficiente, rápido e com baixos custos para fazer diagnósticos. Do contrário, se o modelo for apenas eficiente, os impactos reais que ele pode gerar talvez não sejam tão relevantes, pois um modelo robusto com difícil acesso não será viável para aplicar em muitos casos e, portanto, talvez tenha um impacto pouco significativo para a evolução clínica de um indivíduo. Por outro lado, um modelo feito por aprendizado de máquina que seja de baixo custo, eficiente e rápido tem potencial de auxiliar de forma ágil os diagnósticos de indivíduos para que as decisões clínicas mais apropriadas sejam tomadas, auxiliando para buscar uma melhora do quadro do indivíduo.

8 Metodologia

Este trabalho apresenta uma proposta de diagnóstico clínico usando inteligência artificial com baixo custo e sem biomarcadores para detectar indivíduos expostos ao *P. vivax*. Um modelo de aprendizado de máquina otimizado foi desenvolvido para classificar medições de espectroscopia da matriz de Mueller para amostras de soro sanguíneo. Um total de 58 indivíduos de regiões endêmicas e não endêmicas de malária no Brasil compõem o banco de amostras utilizados nesta pesquisa. Para garantir resultados confiáveis, foram utilizadas técnicas de validação cruzada e regularização na criação de uma metodologia de classificação robusta dos indivíduos. A seguir, todas as principais etapas e considerações sobre o projeto serão apresentadas.

8.1 Amostras de soro sanguíneo

Para este trabalho, usamos amostras de soro sanguíneo de 58 indivíduos. O soro sanguíneo é obtido por centrifugação de uma amostra de sangue que foi deixada coagular. Dessa forma, o soro é a parte líquida do sangue sem as hemácias e sem os fatores de coagulação como plaquetas e fibrinogênio. Ele é composto principalmente por água ($\sim 92\%$), componentes orgânicos, inorgânicos, lipídeos e proteínas. Os processos biológicos que acontecem durante e após a infecção de uma pessoa pelo parasita da Malária levam a modificações físico-químicas no soro sanguíneo [9].

O banco de dados utilizado neste trabalho usa amostras de soro sanguíneo de indivíduos que são reativos ou negativos para malária, tal que cada um desses indivíduos pode ser de área endêmica ou não endêmica de malária. As amostras de soro reativas são de indivíduos que resistiram à doença e agora são sadios (não têm o parasita, mas apresentam níveis de anticorpos IgG e IgM específicos no sangue, denominados de reativos para IgG e IgM), e as amostras negativas são de um grupo de controle que nunca teve a doença e consequentemente não apresenta anticorpos específicos no sangue. Essas amostras foram fornecidas em uma colaboração com duas especialistas no estudo da malária: Dra. Luzia Helena Carvalho e Dra. Flora Satiko Kano do Instituto René Rachou (Fiocruz Minas).

Como as amostras podem ser rotuladas de acordo com a região geográfica (áreas endêmicas ou não endêmicas) e a exposição ao parasita (grupos negativos ou reativos), isso permitiu a criação de subgrupos amostrais distintos no banco de dados. O conjunto de 58 amostras foi, então, organizado em cinco grupos. No primeiro grupo, todas as amostras de áreas não endêmicas foram selecionadas, e então foram comparadas as amostras de indivíduos primo-infectados com as amostras negativas da mesma região. No segundo grupo, todas as amostras reativas foram incluídas para comparar as amostras reativas de

áreas não endêmicas com as reativas provenientes de regiões endêmicas. O terceiro grupo concentra-se exclusivamente em amostras de áreas endêmicas, comparando-se amostras com anticorpos específicos (reativas) com amostras sem anticorpos específicos (negativas). O quarto grupo avalia todas as amostras negativas, contrapondo as negativas das áreas não endêmicas às das áreas endêmicas. Por fim, o quinto grupo reúne todas as amostras reativas para compará-las a todas as amostras negativas, independentemente da localização geográfica. A Tabela 1 apresenta um resumo desses cinco grupos mencionados, detalhando as classes binárias (0 e 1) atribuídas a cada um. É importante apontar nessa tabela que apesar do banco de dados ter um total de 58 amostras, a soma das quantidades dos subgrupos endêmicos com não endêmicos totaliza 57, pois uma das amostras nós não temos informações sobre sua endemicidade.

Tabela 1 – Grupos de amostras de malária

Grupo amostral	Total	Classe 0	Classe 1
Não endêmicas	29	9 (negativas)	20 (reativas)
Reativas	35	20 (não endêmicas)	15 (endêmicas)
Endêmicas	28	13 (negativas)	15 (reativas)
Negativas	22	9 (não endêmicas)	13 (endêmicas)
Todas as amostras	58	23 (negativas)	35 (reativas)

8.2 Medidas espectrais de soro sanguíneo

A preparação das amostras para as medidas experimentais e o processo de aquisição de dados estão ilustrados na Figura 57. Visando simplicidade, baixo custo e viabilidade para seu uso no setor da saúde em todo o território nacional, as amostras são preparadas depositando 20 μL de soro sobre uma lamínula de vidro #1 $\frac{1}{2}$ de 22 mm x 22 mm, e depois o biofilme líquido formado é coberto com uma segunda lamínula de vidro. As amostras de soro foram caracterizadas opticamente utilizando um elipsômetro espectroscópico de ângulo variável M-2000 integrado à interface CompleteEASE [38] da J. A. Woollam Company. De modo geral, o aparelho ilumina a amostra com luz polarizada na faixa espectral entre 245 nm e 1700 nm em determinado ângulo de incidência θ e coleta a luz refletida especularmente. A razão entre as intensidades da luz incidente e refletida em diferentes estados de polarização é medida em função do comprimento de onda da luz incidente de forma a determinar os 12 primeiros elementos da matriz de Mueller e as propriedades Ψ , Δ , depolarização, Aps, Asp e AnE, pois o elipsômetro utilizado é do tipo da figura 29 b).

Esse processo de medida é repetido em nove regiões diferentes de cada amostra para mitigar as complicações que as heterogeneidades espaciais do biofilme líquido podem trazer durante o processo de medição. Dessa forma, para cada amostra utilizada medimos 18 propriedades em 640 comprimentos de onda diferentes, para 6 ângulos de incidência em

9 regiões distintas da amostra. Então, coletamos um total de 622080 dados experimentais por paciente.

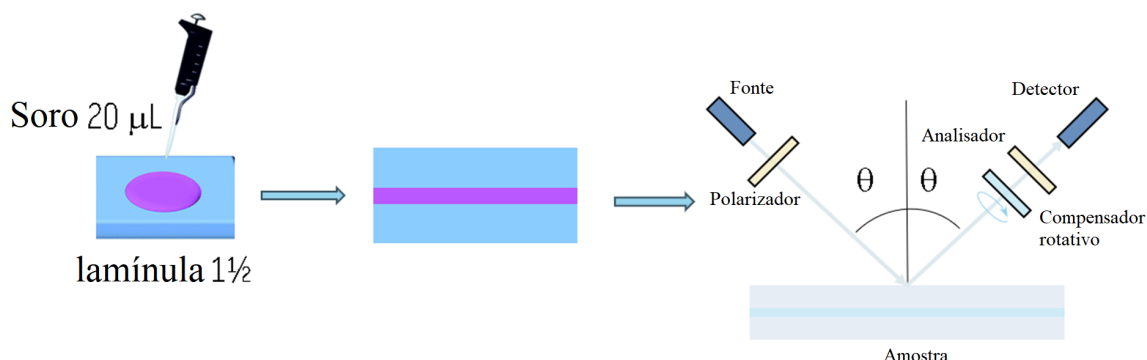


Figura 57 – Preparação de amostras de soro e medições de elipsometria espectral de ângulo variável da matriz Mueller.

Antes de prosseguirmos, devemos retomar uma pergunta central: por que medidas ópticas estão sendo cogitadas para diferenciar amostras negativas das reativas para malária? A realização e uso dessas medidas para um processo de diagnóstico clínico de malária por modelos de inteligência artificial se dá pelo fato de que os processos biológicos que acontecem durante e após a infecção de uma pessoa pelo parasita da Malária levam a modificações físico-químicas em fluidos corporais como o sangue e, em particular, o soro sanguíneo [17]. Logo, a proposta é que as propriedades ópticas desses fluidos corporais são sensíveis às modificações físico-químicas que ocorrem em sua composição. Dessa forma, sugerimos que a determinação direta ou indireta com alta precisão dessas propriedades ópticas pode permitir diferenciar pacientes reativos ou não em relação à malária para constituir um método de diagnóstico clínico usando esses dados aliados ao aprendizado de máquina. Essa sensibilidade às diferenças entre amostras negativas e reativas para malária está relacionada à função dielétrica das amostras, conforme foi discutido anteriormente.

Como visto no Cálculo de Mueller, a interação da luz com um material pode ser completamente descrita pela matriz de Mueller. Cada elemento dessa matriz está associado a uma ou mais propriedades ópticas do material analisado. A matriz de Mueller pode ser determinada com alta precisão por técnicas de elipsometria. Usando a elipsometria espectral de ângulo variável (VASE) é possível extrair para vários ângulos os elementos da matriz de Mueller (mm_{ij}) em diversas regiões espectrais onde as modificações físico-químicas das amostras sejam particularmente relevantes e acentuadas.

8.3 Construção do Modelo de aprendizado de máquina

Os dados obtidos por elipsometria tradicionalmente se processam utilizando modelos ópticos das amostras. Porém, dada a elevada complexidade das amostras biológicas utilizadas neste trabalho, é muito difícil criar modelos físicos realistas que permitam, a partir de medidas de propriedades ópticas, discriminar entre indivíduos infectados ou não por alguma doença. Então, uma alternativa a isso é usar algoritmos que sejam capazes de processar grandes quantidades de dados brutos para encontrar padrões entre essas informações com finalidade de usá-las em uma tarefa de classificação de amostras. Nesse contexto, o uso de aprendizado de máquina e seus algoritmos é uma opção para o processamento desses dados com o objetivo de construir um modelo final de classificação binária.

Todos os modelos criados neste trabalho foram desenvolvidos em linguagem Python, e as condições gerais da criação do modelo serão expostas em seguida. Usaremos 17 propriedades ópticas das 18 que foram medidas, pois para este trabalho não utilizamos as medidas de Δ para criação de modelos, mas todas as outras 17 propriedades são usadas para essa finalidade, totalizando então 587520 dados por paciente. Para processar os dados, utilizamos os 5 algoritmos de aprendizagem de máquina que foram apresentados na seção de algoritmos: SVM, rede neural, regressão logística, XGBoost e KNN. Apesar de existirem diversos outros algoritmos, esses cinco selecionados são bastante representativos das famílias de algoritmos presentes na literatura e estão entre os que apresentam os melhores resultados para o volume de dados com o qual trabalhamos. Como o banco de dados total pode ser dividido em 5 grupos distintos, conforme a tabela 1, cada um desses grupos é utilizado para criar modelos a partir das amostras e classes que eles contêm. Logo, usando as 17 propriedades, 5 algoritmos e 5 grupos de amostras, podemos produzir um total de 425 modelos.

A figura 58 apresenta a metodologia geral para criação e avaliação dos modelos deste trabalho. Nessa abordagem, utiliza-se inteligência artificial para processar dados espectrais provenientes do banco de dados, focando em uma propriedade óptica por vez para produção de um modelo a partir de um algoritmo escolhido. As entradas $\{\vec{X}_i\}$ apresentadas ao algoritmo são estruturadas de forma que as componentes de cada entrada são os valores da propriedade óptica escolhida para cada comprimento de onda utilizado nas medições. Podemos então interpretar isso de tal forma que os comprimentos de onda são *features* e os valores associados a eles são as medidas experimentais da propriedade óptica escolhida. Os valores espectrais dessa propriedade são extraídos para cada amostra, em que cada entrada $\vec{X}$ é associada à medida de um ponto na amostra de um indivíduo. Então, após a seleção do algoritmo e propriedade óptica, o código para criação dos modelos passa por uma etapa de busca dos melhores hiperparâmetros ao maximizar uma função objetivo baseada na métrica AUC (área sob a curva ROC), seguida pelas etapas de treinamento e

validação a partir desse conjunto de hiperparâmetros ótimos. A base de dados passa por um processo de normalização, remoção de dados ruidosos (*outliers*) e aumento de dados (data augmentation) com o objetivo de melhorar a qualidade e robustez do modelo. Após aplicar a validação cruzada *k-fold* (SKFCV), as combinações de algoritmo e propriedade física com maiores AUC de cada grupo amostral da tabela 1 são selecionadas para serem aplicadas no processo de validação cruzada *Nested* (*Nested-CV*). Essa aplicação da *Nested-CV* é feita porque esse método é mais robusto e serve como uma maneira de verificar os resultados obtidos pela SKFCV. Todo o processo geral descrito anteriormente é aplicado para a SKFCV e a *Nested-CV*, cujos detalhes serão discutidos para cada caso em seguida.

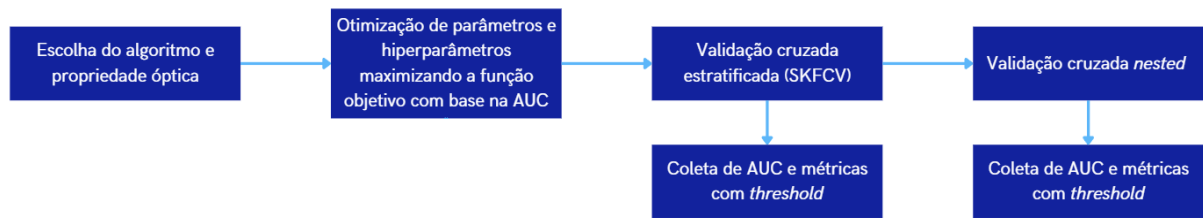


Figura 58 – Metodologia geral para a criação e avaliação dos modelos a partir da validação cruzada estratificada e validação cruzada *nested*.

A figura 59 descreve a nossa implementação da SKFCV para este trabalho. O conjunto de dados foi dividido em dez partes, sendo 9 partes usadas como grupo de treino e a parte restante para o grupo de teste. Essa divisão é feita para buscar os parâmetros e hiperparâmetros ótimos utilizando uma função objetivo de otimização com o auxílio do Optuna [39]. Os parâmetros e hiperparâmetros otimizados com seus espaços de busca estão apresentados nas tabelas 6 e 7 do Apêndice A.

Para o processamento adequado de todos os valores dos atributos, foi realizado um processo de escalonamento usando os dados de treino como referência para o ajuste, e posteriormente esse ajuste de escala foi aplicado aos dados de treino e validação. O escalonador usado foi o Robust Scaler da biblioteca Scikit-Learn, que apresenta bons resultados descritos pela literatura [40]. Esse procedimento é importante para garantir que a base de dados fique na mesma escala. Após escalonamento, fazemos uma busca por outliers com o algoritmo KNN, removendo amostras que poderiam comprometer o desempenho do modelo [41]. Além disso, o SMOTE [42] foi empregado para balancear as classes positivas e negativas, gerando dados artificiais para o grupo de treino. Então, em cada iteração da validação cruzada, as nove partes do conjunto de dados foram utilizadas para treinar o modelo, enquanto a parte restante foi usada para validá-lo por meio da previsão de probabilidades. A probabilidade final de uma amostra ser reativa para malária foi definida a partir da média das probabilidades obtidas da análise de cada uma das nove posições medidas na amostra e, caso uma ou mais dessas posições tenha sido detectada

anteriormente como *outlier*, a média é feita a partir das amostras que restaram. Todo esse processo é repetido, permutando a parte de validação até que todas as dez partes fossem utilizadas para validação, resultando em dez conjuntos de probabilidades preditas. A partir disso, foi calculada uma pontuação da área sob a curva ROC (AUC) nos dados de validação. Esse procedimento anterior foi repetido dez vezes com diferentes sementes aleatórias (distribuições de dados), gerando um valor de AUC por semente, resultando em um total de dez valores de AUC. Assim, a função objetivo visava otimizar a média desses dez valores de AUC calculados a partir dos dados de validação. No total, foram realizadas 150 tentativas para garantir que a busca pelos parâmetros e hiperparâmetros ótimos convergisse para uma solução apropriada.

Após a etapa de otimização, os hiperparâmetros que retornaram maior valor de AUC durante a otimização são coletados e aplicados para o treinamento, validação e avaliação dos resultados em um novo processo de validação cruzada de dez *folds*. Utilizando esses hiperparâmetros ótimos, o conjunto de dados foi dividido novamente em dez partes, e as métricas significativas foram calculadas a partir das probabilidades preditas geradas pelo grupo de validação. Esse processo de validação cruzada foi repetido 30 vezes usando uma semente aleatória diferente a cada repetição para obter resultados generalizados em várias distribuições de dados. Após concluir essas 30 iterações, o valor de cada métrica foi calculado pela média dos valores obtidos em cada semente aleatória. O *threshold* para métricas que exigiam um valor de corte foi determinado a partir das probabilidades previstas do grupo de treinamento.

Após a montagem dos modelos usando SKFCV, foram selecionadas as combinações de algoritmos e propriedades com os melhores resultados de AUC para serem avaliados pela *Nested-CV*, e a figura 60 ilustra como essa validação foi utilizada neste trabalho. A *Nested-CV* segue um procedimento semelhante à SKFCV, mas agora inclui uma validação cruzada dentro de outra, permitindo uma divisão para treinamento, validação e teste, conforme foi exposto na seção sobre métodos de validação. Inicialmente, o conjunto de dados foi dividido em cinco partes: quatro partes formam o grupo A, que é utilizado para otimização e validação, enquanto a parte restante foi reservada para teste do modelo otimizado. Neste contexto, "teste" e "validação" representam etapas distintas: os dados de validação foram usados exclusivamente para otimizar o modelo, enquanto o conjunto de teste foi composto de dados distintos, sem uso em nenhuma outra etapa do processo.

O grupo A foi subdividido em 4 partes para dar início à otimização. Nessa divisão, 3 partes foram utilizadas para treinar o modelo, enquanto a parte restante serviu para validação. O escalonamento é construído a partir dos dados de treino, e depois é usado para ajustar as partes de treino e validação. O processo de revezar os grupos de treino e validação foi repetido 4 vezes, e depois que todas as 4 partes foram utilizadas para validação, as probabilidades previstas são usadas para gerar um valor de AUC para cada

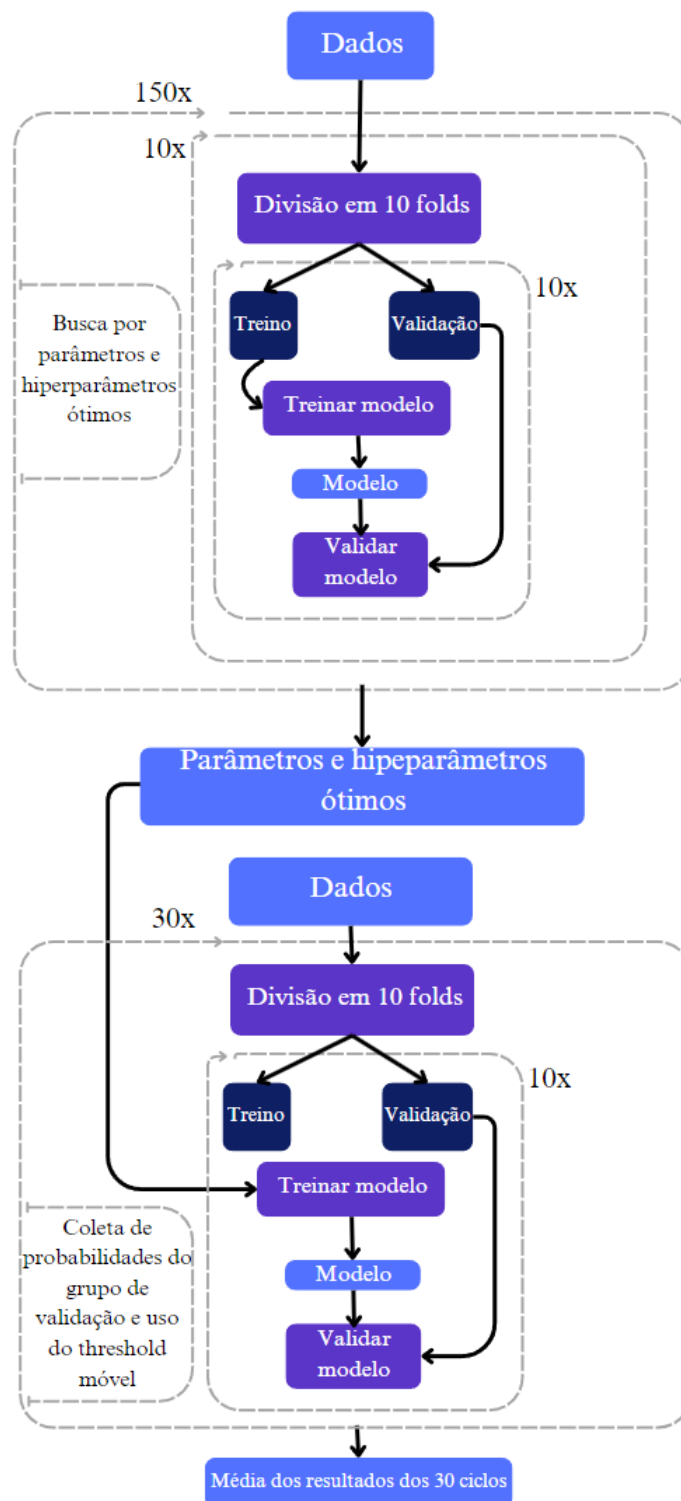


Figura 59 – Metodologia da validação cruzada estratificada para 10-*folds*. Para 10 distribuições de dados, é feita uma validação cruzada de 10 *folds* para otimizar um modelo com base na AUC por 150 tentativas. O modelo ótimo é validado usando grupos de treino e teste para 30 distribuições de dados distintas.

parte. A função objetivo, então, otimizou a média desses valores de AUC e retorna o conjunto de parâmetros e hiperparâmetros que resultaram no maior valor de AUC dentre

as 20 tentativas de busca.

Em seguida, o grupo A foi usado para treinar o modelo com o conjunto de parâmetros e hiperparâmetros ótimos obtidos anteriormente. Assim como nas descrições anteriores, o escalonamento é montado usando os dados de treino (grupo A) e depois é aplicado aos dados de treino e teste. Após treinar o modelo com o grupo A, as probabilidades de classificação são coletadas a partir do grupo de teste, e a partir delas calculamos as métricas: AUC, acurácia, F1, MCC, especificidade e sensibilidade. Após isso, essas métricas são guardadas e o processo se reinicia, mas agora para um grupo A e de teste diferentes dos anteriores. Portanto, para cada vez que os grupos A e de teste forem alterados teremos um novo conjunto de hiperparâmetros ótimos e de métricas, e como dividimos a base de dados em 5 partes, esse processo acontece por 5 vezes. Finalizando essa etapa, a base de dados total é embaralhada para alterar a distribuição dos dados (sementes aleatórias para embaralhar os dados) em relação à distribuição anterior. Ao total, a base de dados sofre uma redistribuição por 5 vezes. Então, ao fim da *Nested-CV*, teremos um total de 25 conjuntos de hiperparâmetros ótimos e de métricas de validação, em que para cada distribuição da base de dados nós teremos os processos de otimização, treino e teste.

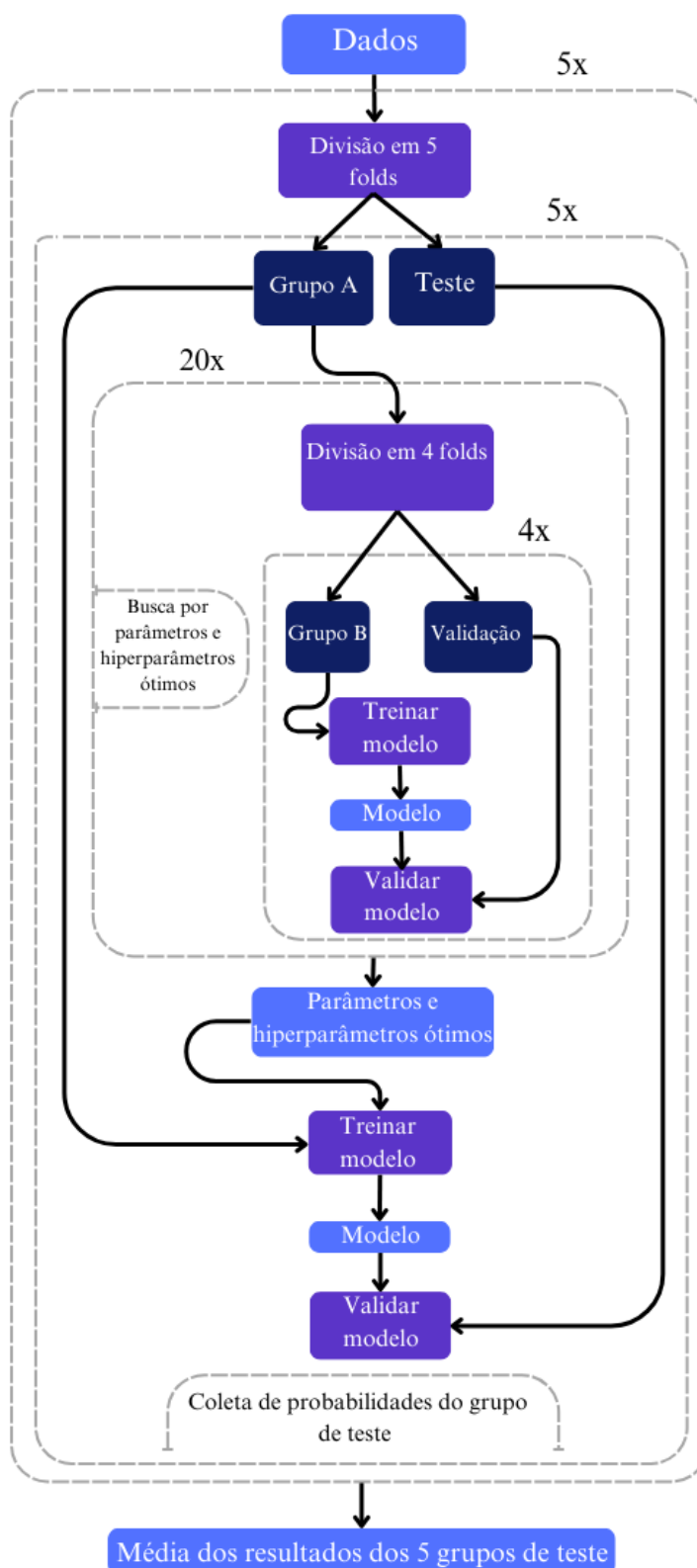


Figura 60 – Metodologia da *Nested-CV* utilizada para avaliar os modelos. Os dados são embaralhados 5 vezes para gerar 5 distribuições distintas de dados. Cada uma dessas distribuições contém um processo interno de otimização, treino e teste.

9 Resultados e discussão

As amostras de soro de malária foram medidas e posteriormente utilizadas para criar e avaliar modelos de aprendizado de máquina na tarefa de classificação de indivíduos reativos ou negativos para a doença. As características espectrais das amostras podem ser analisadas para obter interpretações físicas relevantes, enquanto os modelos criados foram avaliados com base em suas métricas e, posteriormente, alguns deles foram usados para testar suas especificidades contra outras doenças para verificar se os modelos não confundem malária com outras infecções.

9.1 Análise espectral do soro sanguíneo

A análise espectroscópica da matriz de Mueller do soro de pacientes revelou características isotrópicas e despolarizantes entre 1000 nm e 1700 nm para algumas amostras. A figura 61 apresenta a matriz de Mueller teórica para uma amostra isotrópica e despolarizante [20], conforme foi indicado pela expressão (5.10). Na figura 62 temos o espectro da matriz de Mueller para uma amostra de soro de malária do banco de dados para 6 ângulos de incidência entre 45° e 70° em um passo de 5°. O espectro da matriz de Mueller experimental da figura 62 mostra que para a região espectral do infravermelho próximo as amostras com esse tipo de espectro podem ser consideradas transparentes, isotrópicas e com despolarização, significando que é possível relacionar a matriz da figura 61 com os resultados experimentais da figura 62.

Conforme foi descrito pela metodologia, o elipsômetro usado para caracterizar as amostras mede apenas os 12 primeiros elementos da matriz de Mueller. Porém, a partir das figuras 61 e 62, podemos usar as propriedades de simetria da matriz de Mueller para amostras transparentes, isotrópicas e despolarizadas para completar os quatro elementos faltantes da matriz de Mueller. Os valores matriciais que precisam ser completados são aqueles referentes à última linha da matriz e, de acordo com a matriz de Mueller teórica, podemos completar a matriz experimental a partir das quantidades N, C, S e P, que foram definidas anteriormente como:

$$N = \cos(2\Psi), \quad (9.1a)$$

$$C = \sin(2\Psi) \cos(\Delta), \quad (9.1b)$$

$$S = \sin(2\Psi) \sin(\Delta), \quad (9.1c)$$

$$P = \sqrt{[N]^2 + [C]^2 + [S]^2}, \quad (9.1d)$$

tal que P é o grau de polarização [20].

Logo, podemos reparar que ao trabalhar na região do infravermelho próximo as simetrias na matriz de Mueller para esse tipo de amostra permitem reduzir o estudo dos 16 elementos de matriz para apenas 4 selecionados: M_{11} , M_{12} , M_{33} e M_{34} . Portanto, essa abordagem permite não somente a descoberta dos elementos faltantes da matriz, mas também uma simplificação do processo de análise espectroscópica das amostras, posto que a matriz resultante apresenta vários elementos nulos e ainda muitos dos valores restantes são obtidos a partir dos 4 elementos selecionados, que dependem diretamente das medidas padrões de Ψ e Δ da elipsometria. Para verificar o perfil spectral geral das amostras

$$\hat{M} = \begin{bmatrix} 1 & -P.N & 0 & 0 \\ -P.N & P & 0 & 0 \\ 0 & 0 & P.C & P.S \\ 0 & 0 & -P.S & P.C \end{bmatrix}$$

Figura 61 – Representação de uma Matriz de Mueller Teórica que é isotrópica e despolarizante.

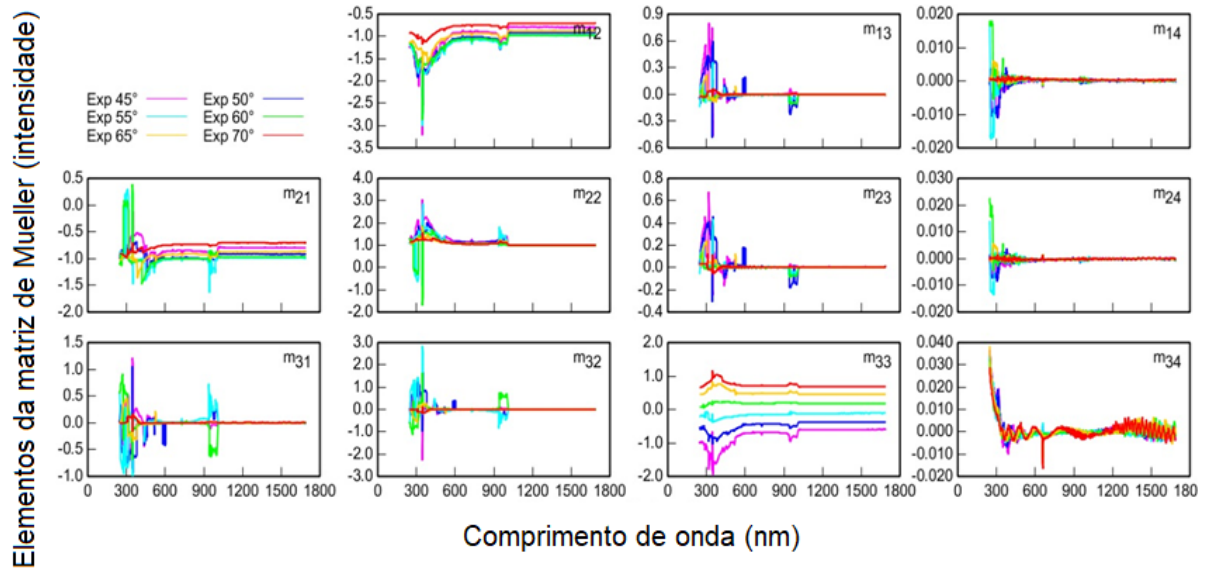


Figura 62 – Espectros da matriz de Mueller experimental para uma amostra de soro da base de dados para seis ângulos de incidência diferentes.

usadas neste trabalho, foram produzidos os espectros médios de cada elemento da matriz de Mueller para o ângulo de incidência de 70°. Os resultados desses espectros médios para as amostras negativas e reativas para malária estão, respectivamente, nas figuras 63 e 64, em que o sombreado ao redor do espectro médio representa o desvio padrão do

valor de intensidade do elemento para cada comprimento de onda. Comparando esses resultados com a matriz de Muller experimental da figura 62, é possível verificar que o comportamento de despolarização, transparência e isotropia no intervalo entre 1000 nm e 1700 nm permanece quando analisamos o espectro médio das amostras negativas e reativas. Portanto, as simetrias relacionadas à figura 61 podem ser generalizadas para o restante das amostras do banco de dados a depender da aplicação desejada. Um outro aspecto a ser analisado nesses espectros médios é que, embora os perfis espectrais da matriz de Mueller das amostras negativas e reativas pareçam semelhantes, há diferenças sutis e específicas de intensidades que podem ser suficientes para que um modelo robusto de aprendizagem de máquina seja capaz de diferenciar esses perfis espectrais.

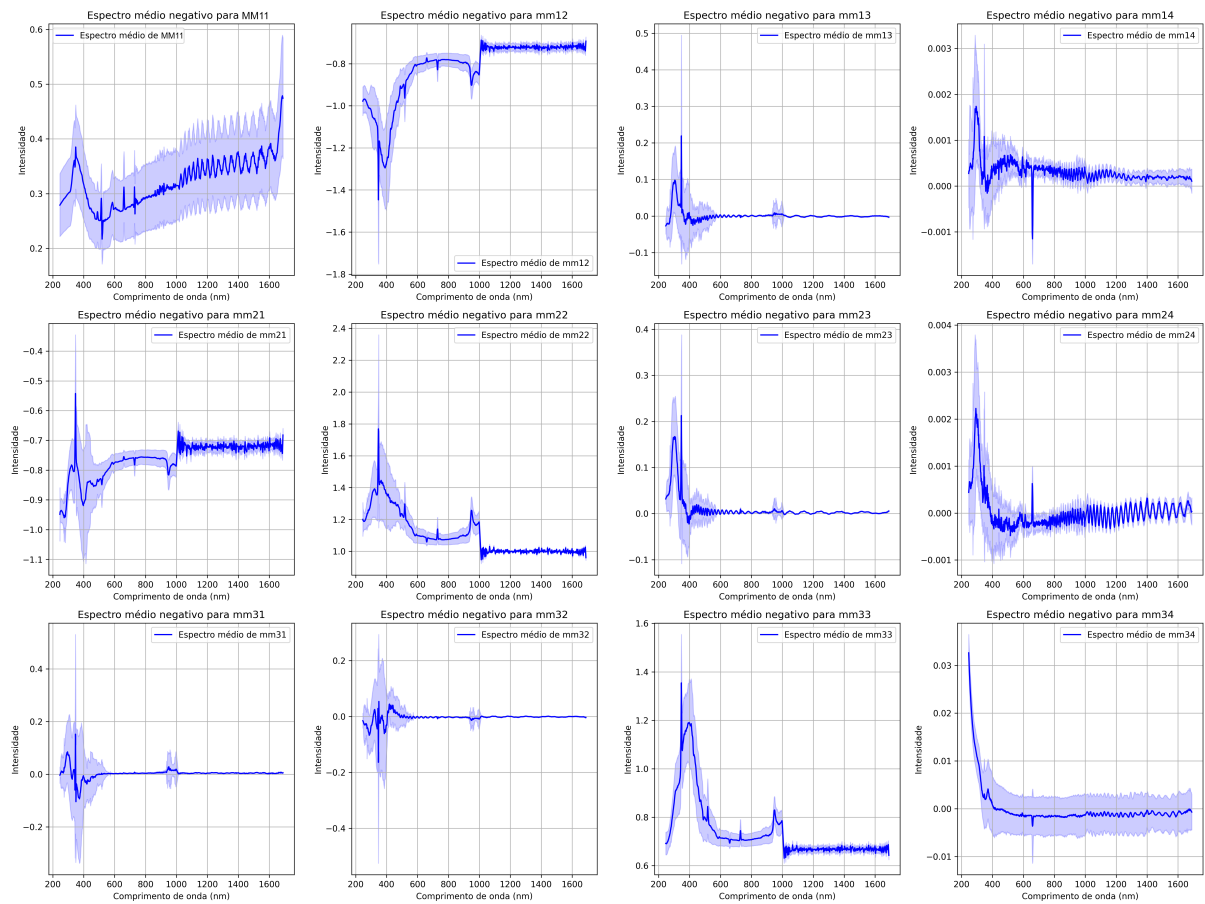


Figura 63 – Espectro médio de cada elemento da matriz de Mueller de amostras negativas para malária em um ângulo de incidência de 70°.

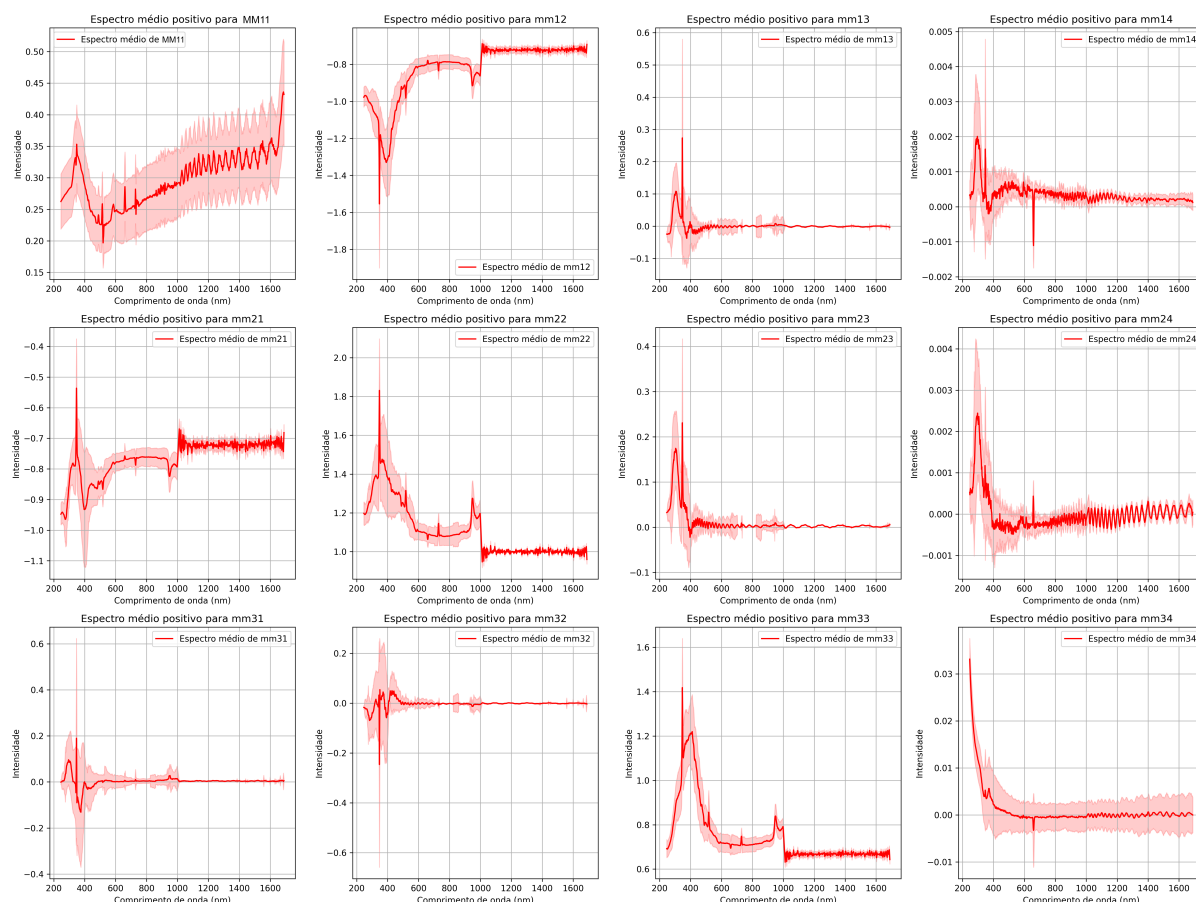


Figura 64 – Espectro médio de cada elemento da matriz de Mueller de amostras reativas para malária em um ângulo de incidência de 70° .

9.2 Modelos de diagnóstico por aprendizado de máquina

A partir da metodologia exposta, foram produzidos 425 modelos de aprendizado de máquina para diagnóstico de malária usando a SKFCV, sendo que cada um dos 5 grupos amostrais possui 85 modelos. Os resultados de todos os modelos criados para os 5 grupos amostrais estão nas tabelas do Apêndice B. Dentre todos os modelos criados, foram selecionados aqueles com a maior AUC de cada um dos grupos amostrais para aprofundar a análise de resultados de cada um. A tabela 2 mostra, para cada grupo amostral, o algoritmo, propriedade óptica, métricas, threshold ótimo e a faixa espectral otimizada para esses modelos com maiores pontuações de AUC por grupo amostral, e a figura 65 mostra as curvas ROC médias para cada modelo da tabela 2. Uma análise geral desses resultados mostra que todos os modelos selecionados tiveram uma pontuação de AUC acima de 0.9 e com erros (desvio padrão) na segunda casa decimal. Isso indica que os modelos obtiveram resultados satisfatórios de AUC, e o baixo erro mostra uma estabilidade das médias da métrica AUC para a SKFCV. Ao analisar as outras métricas da tabela 2 observa-se que

Tabela 2 – Resultados das métricas para os melhores modelos de cada grupo amostral usando a SKFCV

Grupo amostral	Algoritmo	Propriedade	AUC	Acc	F1	MCC	Espec.	Sens.	Threshold	Faixa espectral
Não endêmicos	SVM	MM_{11}	0.98 ± 0.02	0.86 ± 0.04	0.91 ± 0.03	0.7 ± 0.1	0.6 ± 0.1	0.99 ± 0.01	0.33 ± 0.05	(273 – 1512) nm
Reativos	SVM	MM_{11}	0.99 ± 0.01	0.97 ± 0.03	0.96 ± 0.04	0.94 ± 0.06	0.99 ± 0.01	0.93 ± 0.08	0.64 ± 0.05	(938 – 1375) nm
Endêmicos	LR	mm_{13}	0.91 ± 0.04	0.82 ± 0.05	0.84 ± 0.04	0.6 ± 0.1	0.7 ± 0.1	0.90 ± 0.06	0.32 ± 0.03	(870 – 1142) nm
Negativos	SVM	MM_{11}	0.99 ± 0.01	0.99 ± 0.02	0.99 ± 0.01	0.99 ± 0.03	1.0 ± 0.0	0.99 ± 0.03	0.50 ± 0.06	(559 – 905) nm
Todas as amostras	SVM	MM_{11}	0.95 ± 0.02	0.89 ± 0.03	0.91 ± 0.02	0.77 ± 0.05	0.70 ± 0.07	0.99 ± 0.01	0.40 ± 0.03	(292 – 1538) nm

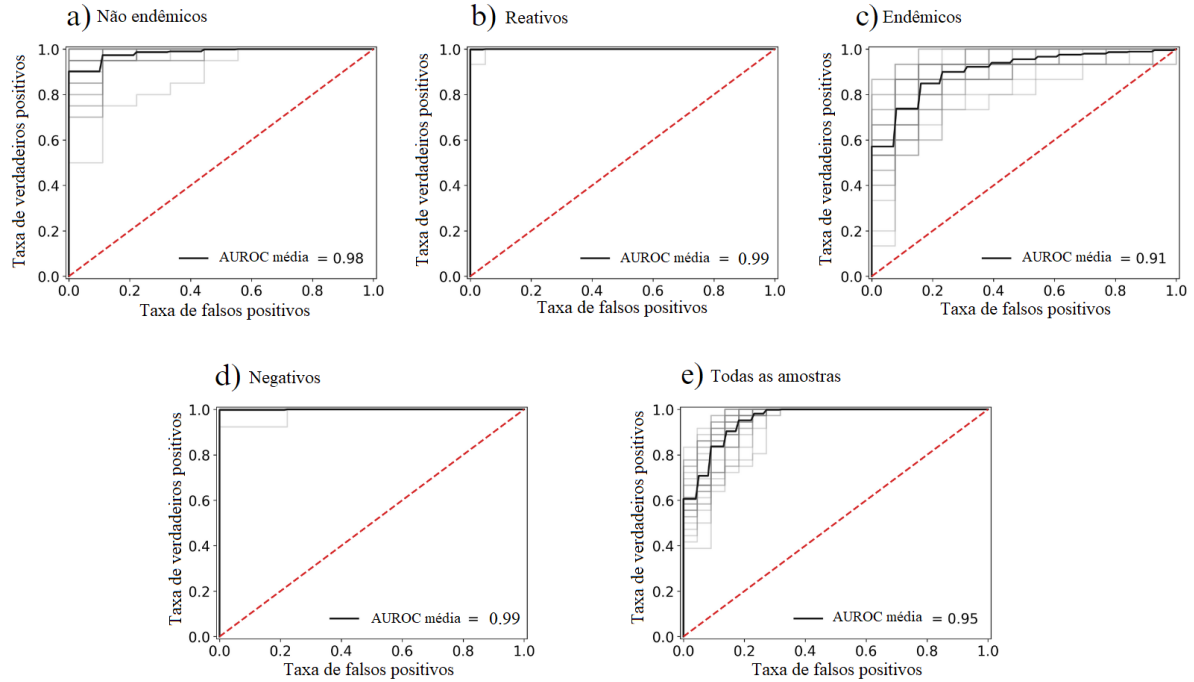


Figura 65 – Curvas ROC dos modelos da tabela 2.

elas tiveram valores acima de 0.9 e outros abaixo de 0.8, que são resultados que têm relação com uma base de dados que é pequena, mas ainda capaz de proporcionar métricas acima de 0.9. Isso pode ser um indicativo de que a otimização da AUC para o processo de obter um modelo ótimo tende a fornecer um modelo que, ao atingir um alto valor de AUC, tem também a capacidade estender bons resultados para o restante das métricas. Para os modelos em que há uma diferença perceptível entre os valores de especificidade e sensibilidade, observa-se que o valor de MCC é afetado. Isso é um resultado esperado, pois pela equação (6.12) vemos que um desequilíbrio entre especificidade e sensibilidade tende a penalizar o valor do MCC. Outro aspecto a ser notado nessa tabela é que dos 5 modelos selecionados, 4 deles usam o mesmo algoritmo e propriedade física: SVM- MM_{11} . Os bons resultados para esse algoritmo parece indicar que o processo de construção de hiperplanos feito pelo SVM é eficiente para discriminar amostras negativas das reativas para malária. O alto desempenho de MM_{11} é particularmente interessante do ponto de vista físico, pois essa propriedade está relacionada com a intensidade da luz não polarizada que é refletida pela amostra. A interpretação que podemos fornecer para os resultados satisfatórios dessa propriedade é que existe um padrão espectral de MM_{11} para cada classe

dos grupos amostrais da tabela 1, e um padrão bem distinto tem um potencial grande de ser percebido pelos algoritmos, possibilitando uma discriminação eficiente entre as duas classes.

Para o grupo amostral de dados não endêmicos foi obtido um AUC de 0.98, demonstrando um grande potencial do modelo, e os resultados das outras métricas foram bons de modo geral para SVM- MM_{11} . Embora a especificidade de 0.6 ± 0.1 não seja um resultado tão apropriado quando comparado ao restante das métricas, podemos ver pela tabela 1 que o número de amostras negativas não endêmicas é metade do número das reativas não endêmicas. Essa diferença de valores entre as duas classes pode ser um indicativo para esse valor de especificidade, principalmente considerando que a sensibilidade obtida foi de 0.99 ± 0.01 , mostrando que o modelo conseguiu compreender e classificar melhor as amostras reativas do que as negativas para a área endêmica. Há uma importância muito grande para o modelo do grupo amostral não endêmico ter um bom valor de sensibilidade, pois ele será capaz de identificar com boa precisão os indivíduos reativos para malária e que vivem em uma região não endêmica da doença. O fato da infecção não ser típica da região faz com que ela não seja uma suspeita clínica viável, e então o indivíduo contaminado por ela pode não receber o diagnóstico correto. Dessa forma, um modelo com alta sensibilidade consegue mapear bem os indivíduos reativos para a doença e tem um potencial enorme para detecção de indivíduos contaminados para iniciar um tratamento precoce contra a malária. Esse modelo se encaixa muito bem para pessoas que viajam por uma curta temporada para uma região endêmica de malária e contraem a doença, mas só manifestam os sintomas quando retornam para uma área com baixa recorrência de malária. Portanto, esse modelo tem uma importância para ser utilizado em áreas não endêmicas para diagnóstico de malária.

Ao analisarmos o grupo amostral de reativas, temos resultados adequados para as métricas e todos eles com um valor acima de 0.9. Esses valores altos dos resultados indicam que o modelo gerado pela combinação de SVM e MM_{11} foi eficiente na tarefa de diferenciar o que são amostras reativas de área endêmica das de área não endêmica. Na prática, esse alto desempenho tem consequências interessantes, pois os reativos de área endêmica podem ser indivíduos que contraíram malária em mais de uma ocasião, enquanto os reativos de área não endêmica tem um perfil associado a indivíduos que tiveram o primeiro contato com a doença. Partindo disso, há um potencial grande de desenvolver esse modelo para que ele possa diferenciar infecções primárias de malária das recorrentes.

Tomando os resultados do grupo amostral endêmico, temos algumas interpretações semelhantes ao caso não endêmico. Em geral, temos resultados acima de 0.8 e 0.9, mas a especificidade de 0.7 ± 0.1 ainda é um pouco menor e com maior erro quando comparada com a sensibilidade de 0.90 ± 0.06 . Porém, diferentemente do grupo amostral não endêmico, essas amostras endêmicas possuem uma base dados mais balanceada, em que 13 amostras

são negativas e 15 são positivas. Uma possível interpretação para essa diferença de desempenho entre especificidade e sensibilidade é que a quantidade e representatividade das amostras negativas endêmicas usadas não proporcionaram que o modelo captasse os padrões espectrais desse tipo de amostra, diferentemente do que aconteceu para amostras positivas de área endêmica, que tiveram bons resultados na sensibilidade. Então, para esse modelo de LR- mm_{13} , tem-se uma alta capacidade de identificar verdadeiros reativos para malária em grupos de amostras endêmicas. Com isso, esse modelo tem uma aplicabilidade para diagnóstico de malária em regiões endêmicas da doença.

No grupo amostral de negativas, o modelo usando SVM- MM_{11} para essas amostras teve um desempenho perfeito para especificidade, e todas as outras métricas ficaram com o valor de 0.99 e baixos erros. Esse grupo amostral não apresenta uma interpretação direta para compreender os detalhes dos resultados encontrados, mas o que observa-se é que o modelo foi muito eficiente em diferenciar amostras negativas não endêmicas das endêmicas. O que pode-se colocar em hipótese é que há alguma assinatura espectral das amostras negativas de área endêmica que permite diferenciá-las das medidas de área não endêmica.

Por fim, temos o grupo amostral que reúne todas as amostras negativas e reativas do banco de dados, independentemente se são de área endêmica ou não. A AUC de 0.95 ± 0.02 mostra um potencial muito grande para esse modelo de SVM- MM_{11} . A especificidade de 0.70 ± 0.07 e a sensibilidade de 0.99 ± 0.01 indicam que o modelo apresentou um desempenho superior na classificação de amostras reativas em comparação com as amostras negativas. Os valores de métricas indicam que esse modelo é promissor, e um número a mais de amostras para intensificar seu treinamento pode melhorar ainda mais seus resultados, principalmente com relação à especificidade. Logo, esse modelo tem potencial capacidade de diferenciar amostras negativas das reativas de malária independentemente da região geográfica, e é um modelo que pode ser estudado e adaptado para atuar em qualquer região do Brasil.

A tabela 3 contém os resultados de *Nested-CV* para os mesmos algoritmos e propriedades ópticas usadas em cada grupo amostral da tabela 2. Comparando as métricas das tabelas 2 e 3 para SKFCV e *Nested-CV* respectivamente, podemos ver que os resultados, de forma geral, estão todos próximos, embora alguns deles tenham sofrido modificações mais perceptíveis. Considerando a natureza da *Nested-CV*, os modelos possuem uma quantidade menor de dados disponíveis durante o processo de otimização. Uma consequência disso é que base de dados pequenas tendem a ser penalizadas nos resultados quando o número de dados para a otimização diminui, embora isso não seja uma regra intransponível. Podemos verificar o caso do grupo amostral endêmico que apresentou uma sensibilidade de 0.2 ± 0.3 , embora as outras métricas não tenham acompanhado essa considerável queda de desempenho. Por outro lado, o grupo amostral de negativos sofreu poucas alterações ao comparar os dois métodos de validação. Portanto, verifica-se que há semelhanças entre os

resultados de SKFCV e *Nested-CV*, e as mudanças que ocorreram são coerentes com a operação da *Nested-CV*.

Tabela 3 – Resultados das métricas para os melhores modelos de cada grupo amostral usando a *Nested-CV*

Grupo amostral	Algoritmo	Propriedade	AUC	Acc	F1	MCC	Espec.	Sens.	Threshold
Não endêmicos	SVM	MM_{11}	0.9 ± 0.1	0.8 ± 0.2	0.8 ± 0.2	0.5 ± 0.4	0.7 ± 0.4	0.8 ± 0.2	0.47 ± 0.04
Reativos	SVM	MM_{11}	0.9 ± 0.1	0.9 ± 0.1	0.9 ± 0.2	0.8 ± 0.2	1.0 ± 0.1	0.9 ± 0.3	0.6 ± 0.2
Endêmicos	LR	mm_{13}	0.9 ± 0.2	0.9 ± 0.2	0.8 ± 0.3	0.8 ± 0.4	0.8 ± 0.3	0.2 ± 0.3	0.5 ± 0.2
Negativos	SVM	MM_{11}	1	1 ± 0.1	0.96 ± 0.08	0.9 ± 0.2	0.9 ± 0.2	1	0.4 ± 0.1
Todas as amostras	SVM	MM_{11}	0.8 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.4 ± 0.2	0.75 ± 0.07	0.7 ± 0.1	0.5 ± 0.05

É importante destacar que, pela tabela 2, obtivemos o melhor desempenho de classificação na região do infravermelho próximo apenas para três grupos de amostras. Para os grupos de amostras não endêmicos, negativos e todas as amostras, o melhor desempenho na classificação foi encontrado em faixas espectrais que incluem o ultravioleta, o visível e o infravermelho próximo. Do ponto de vista físico, é interessante restringir os resultados à região do infravermelho, dado que as propriedades ópticas das amostras são mais simples nessa faixa, sem anisotropias ópticas, o que facilita a elaboração de modelos ópticos em trabalhos futuros. Do ponto de vista tecnológico, também é vantajoso obter os melhores resultados em uma única região espectral (ultravioleta, visível ou infravermelho próximo), pois cada faixa exige fontes de luz, detectores e monocromadores específicos, o que eleva significativamente os custos da instrumentação no caso de dispositivos que usem várias regiões espectrais.

Dessa forma, decidimos avaliar o desempenho dos modelos de ML apresentados na tabela 2 para os grupos não endêmicos, negativos e todas as amostras, restringindo a região espectral dos dados ao infravermelho próximo. Para o grupo Não Endêmicos, obtivemos $AUC = (0.99 \pm 0.01)$ na região espectral entre 1270 nm e 1648 nm. Para o grupo Negativos, o valor foi $AUC = (0.99 \pm 0.02)$ na região espectral entre 1388 nm e 1632 nm. Já para o grupo Todas as Amostras, obtivemos $AUC = (0.93 \pm 0.02)$ na região espectral entre 1071 nm e 1619 nm. As variações de desempenho observadas nos três grupos avaliados exclusivamente na faixa do infravermelho, em comparação com toda a faixa espectral medida, estão dentro das incertezas e podem ser desconsideradas. Assim, podemos restringir o método de diagnóstico de malária para cada um dos grupos de amostras estudados à região de comprimentos de onda do infravermelho próximo, sem perda de desempenho no processo classificatório.

9.3 Especificidade dos modelos contra outras doenças

Os modelos apresentados anteriormente obtiveram bons desempenhos para detectar amostras reativas de malária. Porém, deve-se dar um passo a mais e pensar o quanto esses modelos são específicos para o diagnóstico de malária. Isso significa que um modelo,

idealmente, não deve classificar uma amostra positiva para uma outra infecção como sendo uma amostra reativa de malária, pois o mais apropriado é que outras doenças sejam classificadas como negativas para malária. A fim de verificar o quanto os modelos construídos são específicos para malária, eles foram apresentados a amostras positivas de COVID-19 e dengue, que são infecções presentes em todo o território nacional. A tabela 4 informa a quantidade de amostras de COVID-19 e dengue que foram utilizadas nos testes.

Tabela 4 – Quantidade de amostras de COVID-19 e dengue

COVID-19	Dengue
65	18

Os parâmetros e hiperparâmetros ótimos que produziram os modelos da tabela 12 (Apêndice B), que é a tabela que contém os modelos produzidos para este trabalho usando todas as 17 propriedades e 5 algoritmos, foram usados para refazer esses modelos, mas agora usando todo o banco de dados de malária para treiná-los. Isso foi feito para criar vários modelos de produção para verificar a especificidade de cada modelo de produção contra a malária. Após isso, os dados de COVID-19 e dengue foram apresentados ao modelo para que fossem classificados como sendo negativos ou reativos para malária, e a acurácia com relação a essas classificações foi feita para cada doença. Nesse caso, uma alta acurácia contra COVID-19 ou dengue significa que o modelo classificou corretamente as amostras dessas doenças como sendo negativas para malária, resultando em uma alta especificidade contra COVID-19 e dengue. A tabela 5 mostra modelos com bons resultados de AUC e seus respectivos valores de especificidade contra COVID-19 e dengue. Foram selecionados três modelos que obtiveram resultados altos de especificidade contra pelo menos uma das duas infecções.

Tabela 5 – Especificidade dos modelos de malária contra COVID-19 e dengue

Algoritmo	Propriedade	AUC (Malária)	Especificidade (COVID-19)	Especificidade (Dengue)
SVM	MM_{11}	0.95	0.97	0.78
KNN	MM_{11}	0.81	0.89	0.98
LR	MM_{11}	0.85	0.99	0.29

Os três modelos da tabela 5 representam uma seleção com desempenhos interessantes para aplicação. O modelo SVM- MM_{11} apresenta uma especificidade de 0.97 contra COVID-19 e de 0.78 contra a dengue. Esse modelo tem uma especificidade alta contra COVID-19 e boa para dengue, significando que ele pode ter uma aplicação adequada e que confunde pouco essas doenças com malária. Para o modelo KNN- MM_{11} temos uma situação parecida com o caso de SVM- MM_{11} , mas agora o modelo tem acurácia de 0.89 contra COVID-19 e 0.98 para dengue. Como esse foi o modelo com maior especificidade contra dengue, ele pode ter uma aplicação direcionada aos casos em que dengue e malária sejam suspeitas clínicas. Tomando os resultados de LR- MM_{11} , temos uma especificidade contra COVID-19

de 0.99 e de 0.29 com relação à dengue. Embora o desempenho para dengue tenha sido baixo, a alta especificidade contra COVID-19 faz com que esse modelo tenha um potencial de uso nos casos em que um indivíduo tenha suspeita de malária ou COVID-19.

Logo, os três modelos da tabela 5 apresentam boas especificidades contra pelo menos uma das duas doenças testadas: COVID-19 e dengue. Como esses modelos têm bons valores de AUC, logo eles também apresentam um grande potencial para discriminar entre amostras negativas e reativas para malária. Portanto, os desempenhos e as altas especificidades de malária dos modelos permitem que eles classifiquem corretamente as amostras reativas para malária, evitando que amostras positivas para COVID-19 ou dengue sejam classificadas como reativas de malária.

10 Conclusão

A partir dos resultados e considerações feitas, temos uma abordagem sólida sobre o estudo e aplicação de um diagnóstico de amostras reativas para malária por aprendizado de máquina a partir da caracterização óptica de amostras de soro por elipsometria.

A análise espectral revela que há amostras com características de isotropia, despolarização e transparência para a região do espectro do infravermelho próximo. Logo, as amostras que apresentam essas características possuem simetrias tais que permitem obter os valores da última linha da matriz de Mueller a partir das três linhas anteriores. Esse é um resultado interessante, pois permite que a matriz de Mueller seja completamente preenchida para os casos em que o elipsômetro utilizado para as medidas não tem acesso aos valores da última linha.

Foram produzidos 425 modelos, sendo um valor de 85 para cada um dos 5 grupos amostrais. Para os modelos com maior AUC de cada grupo amostral os valores das métricas de desempenho, principalmente da acurácia e AUC, foram satisfatórios e revelam que os modelos de inteligência artificial criados se mostraram muito eficientes na tarefa de classificação.

Partindo dos cinco grupos amostrais, os modelos desenvolvidos para cada um deles desempenham papéis específicos, os quais foram verificados por meio do método de validação cruzada estratificada (SKFCV). Todos esses melhores modelos citados foram reavaliados de forma mais robusta pela *Nested-CV*, que validou várias das métrica e resultados de cada modelo selecionado a partir da SKFCV. Ao escolher o melhor modelo para cada grupo amostral, observou-se que todos apresentaram uma AUC acima de 0.9, indicando um alto potencial de desempenho. No grupo amostral reativo, os resultados foram satisfatórios para a classificação das amostras conforme a região de origem (endêmica ou não endêmica para malária), o que também se aplicou ao grupo de amostras negativas. Nos outros três grupos amostrais, houve uma boa classificação das amostras entre reativas e negativas para malária, considerando ou não a endemicidade da região geográfica. Os grupos amostrais de regiões endêmicas e não endêmicas apresentaram modelos com bom desempenho para diferenciar amostras negativas das reativas, de acordo com as respectivas características epidêmicas de cada grupo. Por fim, para o grupo que compreende todas as amostras do banco de dados, obteve-se um modelo capaz de discriminar amostras negativas e reativas para malária, independentemente da região de origem da amostra.

Esses resultados evidenciam um potencial método alternativo para diagnóstico rápido e barato de malária, visto que o tempo total de medida das propriedades mais relevantes para este trabalho (MM_{11} ou mm_{13}) pelo elipsômetro é de no máximo 20 segun-

dos, e o custo é restrito apenas ao valor das lamínulas de vidro utilizadas. Considerando os excelentes resultados com a propriedade M_{11} e fixando um ângulo de incidência, o tempo de diagnóstico pode ser drasticamente reduzido já que não é necessário medir todas as propriedades nem usar vários ângulos de incidência. Esses resultados, considerando que apenas precisamos medir MM_{11} para um ângulo fixo e na região do infravermelho próximo, também abrem oportunidades para um projeto de fabricação de um instrumento para diagnóstico de malária do tipo ponto de atendimento (point of care), que é inexistente no mercado até o momento, para que o diagnóstico clínico de malária fosse mais acessível.

Os modelos do grupo amostral não endêmico apresentam um papel de interesse fundamental neste trabalho, pois uma pessoa com malária dentro de uma região não endêmica da doença pode não ser diagnosticada com a infecção por ela não se tratar de uma suspeita clínica. Dessa forma, o diagnóstico pode ser demorado, e como o tratamento de malária precisa ser iniciado cedo, o indivíduo acometido pela doença pode ter complicações. Como o modelo não endêmico apresentado neste trabalho possui bom desempenho para diagnóstico de malária, baixos custos e resultados rápidos, espera-se que esse modelo possa auxiliar nesses casos em que a doença não é uma suspeita clínica.

Foram obtidos modelos com alta especificidade contra COVID-19 e dengue, que são infecções com alguns sintomas em comum com a malária, podendo resultar em diagnósticos errados. Dessa forma, alguns modelos selecionados podem diferenciar com alta precisão amostra de malária das amostras de COVID-19 e dengue. Essa capacidade dos modelos é de grande importância, pois a malária ser confundida com outras infecções atrasa o início do tratamento do indivíduo, podendo provocar complicações no quadro clínico.

Então, este trabalho tem como produto um total de 425 modelos, e cinco deles foram explorados com mais atenção e apresentam desempenhos adequados, avaliados por dois métodos de validação, para cada um dos cinco grupos amostrais. Os resultados mostram que é possível fazer classificação por aprendizado de máquina a partir de amostras negativas e reativas de malária usando caracterização espectroscópica de amostras de soro pelo uso da elipsometria. Os algoritmos conseguem processar os espectros das propriedades físicas de forma a reconhecer padrões que possibilitam diferenciar amostras negativas de reativas ou amostras endêmicas de não endêmicas.

Referências

- [1] Goodfellow, Ian, Yoshua Bengio e Aaron Courville: *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>. Citado 16 vezes nas páginas 17, 70, 71, 72, 73, 74, 76, 77, 78, 80, 81, 99, 101, 102, 103, and 105.
- [2] Ovid e Charles Martin: *Metamorphoses*. W.W. Norton, New York, 2004. Citado na página 17.
- [3] Sparkes, Brian: *The Red and the Black: Studies in Greek Pottery*. Routledge, London, 1ª edição, 1996. Citado na página 17.
- [4] Tandy, David W.: *Works and Days: A Translation and Commentary for the Social Sciences*. University of California Press, Berkeley, CA, 1997. Citado na página 17.
- [5] Harkup, Kathryn: *Making the Monster: The Science of Mary Shelley's Frankenstein*. Bloomsbury Sigma, 2018, ISBN 9781472933737. Citado na página 17.
- [6] Harkup, Kathryn: *Frankenstein: Anatomia de Monstro*. Darkside Books, 2023. Tradução do livro original "Making the Monster: The Science of Mary Shelley's Frankenstein". Citado na página 17.
- [7] Liu, Jie, Dong Zhang, Dong Yu *et al.*: *Machine learning powered ellipsometry*. Light: Science and Applications, 10(55), 2021. <https://doi.org/10.1038/s41377-021-00482-0>. Citado na página 17.
- [8] Mshani, Isack H., Deus J. Siria, Elias P. Mwanga *et al.*: *Key considerations, target product profiles, and research gaps in the application of infrared spectroscopy and artificial intelligence for malaria surveillance and diagnosis*. Malaria Journal, 22:346, 2023. <https://doi.org/10.1186/s12936-023-04780-3>. Citado na página 18.
- [9] França, Tanos, Marta Santos e José Figueroa-Villar: *Malária: aspectos históricos e quimioterapia*. Quím. Nova, 31:1271–1278, 2008. Citado 3 vezes nas páginas 20, 21, and 118.
- [10] Mito, L. D., L. C. Faccin Galhardi e M. K. Amarante: *Aspectos parasitológicos e imunológicos da malária*. Biosaúde, 14(1):42–55, 2016. <https://ojs.uel.br/revistas/uel/index.php/biosaude/article/view/24324>, Recuperado de <https://ojs.uel.br/revistas/uel/index.php/biosaude/article/view/24324>. Citado 5 vezes nas páginas 20, 21, 22, 23, and 24.

- [11] Saúde, Brasil. Ministério da: *Guia para profissionais de saúde sobre prevenção da malária em viajante*. https://bvsmis.saude.gov.br/bvs/publicacoes/guia_prevencao_malaria_viajantes.pdf. Acesso em 23/01/2023. Citado 3 vezes nas páginas 20, 23, and 24.
- [12] Moxon CA, Grau GE, Craig AG: *Malaria: modification of the red blood cell and consequences in the human host*. British Journal of Haematology, 154(6):670–9, 2011. Citado na página 21.
- [13] Murphy, Sean C. *et al.*: *Malaria diagnostics in clinical trials*. The American Journal of Tropical Medicine and Hygiene, 89(5):824–839, 2013. Citado na página 21.
- [14] Glória, Juliane Corrêa, Maria Edilene Martins de Almeida, Késsia Caroline Souza Alves, Felipe Araújo, Ricardo Andrez Machado de Ávila, Paulo Afonso Nogueira e Luís André Morais Mariúba: *Métodos para diagnóstico de Malária: Atualização e desafios*. Scientia Amazonia, 7(2):B24–B31, 2018. <https://scientia-amazonia.org/wp-content/uploads/2018/05/v7-n2-b24-b31-2018.pdf>, ISSN: 2238-1910. Citado na página 21.
- [15] Ferreira, Antonio Walter e Maria Carmen Arroyo Sanchez: *Malária humana: padronização e otimização de testes sorológicos para diagnóstico individual e inquéritos soroepidemiológicos*. Revista do Instituto de Medicina Tropical de São Paulo, 30(3):137–146, June 1988. <https://doi.org/10.1590/S0036-46651988000300003>. Citado na página 22.
- [16] Saúde, Brasil. Ministério da: *Situação Epidemiológica da Malária*. <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/m/malaria/situacao-epidemiologica-da-malaria>. Acesso em 23/01/2023. Citado na página 23.
- [17] El Saftawy, E., M. F. Farag, H. H. Gebreil, M. Abdelfatah, B. E. Aboulhoda, M. Alghamdi, E. A. Albadawi e M. A. Abd Elkhalek: *Malaria: biochemical, physiological, diagnostic, and therapeutic updates*. PeerJ, 12:e17084, 2024. <https://doi.org/10.7717/peerj.17084>. Citado 2 vezes nas páginas 25 e 120.
- [18] Fujiwara, Hiroyuki: *Spectroscopic Ellipsometry Principles and Applications*. John Wiley and Sons Ltd, Chichester, West Sussex, England, 2007. Citado 37 vezes nas páginas 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 65, 66 e 67.
- [19] Griffiths, David J.: *Eletrodinâmica*. Pearson Universidades, 3ª edição, dec 2010. Citado na página 32.

- [20] Hilfiker, James N., Nina Hong e Stefan Schoeche: *Mueller matrix spectroscopic ellipsometry*. Advanced Optical Technologies, 1(3):59–91, 2022. Citado 9 vezes nas páginas 59, 60, 61, 62, 63, 64, 68, 69 e 127.
- [21] J. A. Woollam Co., Inc.: *CompleteEASE Data Analysis Manual*. J. A. Woollam Co., Inc., 2011. Citado na página 68.
- [22] Tumkur, T., Y. Barnakov, S. T. Kee, M. A. Noginov e V. Liberman: *Permittivity evaluation of multilayered hyperbolic metamaterials: Ellipsometry vs. reflectometry*. Journal of Applied Physics, 117(10):103104(1)–103104(9), 2015. Citado na página 68.
- [23] Boden, Margaret A.: *Inteligência artificial: Uma brevíssima introdução*. Editora Unesp, São Paulo, 1ª edição, 2020. Citado na página 70.
- [24] Brownlee, Jason: *Master Machine Learning Algorithms*. v1.1 edição, 2016. <http://MachineLearningMastery.com>, Virtual Edition. Citado 7 vezes nas páginas 71, 73, 93, 94, 95, 112 e 113.
- [25] Choudhury, A.: *The Role of Machine Learning Algorithms in Materials Science: A State of Art Review on Industry 4.0*. Archives of Computational Methods in Engineering, 28:3361–3381, 2021. <https://doi.org/10.1007/s11831-020-09503-4>. Citado na página 76.
- [26] Rainio, O., J. Teuvo e R. Klén: *Evaluation Metrics and Statistical Tests for Machine Learning*. Scientific Reports, 14:6086, 2024. <https://doi.org/10.1038/s41598-024-56706-x>. Citado 2 vezes nas páginas 82 e 83.
- [27] Dehmer, Matthias e Subhash C. Basak: *Statistical and Machine Learning Approaches for Network Analysis*. Wiley, 2012. Citado na página 82.
- [28] Lantz, Brett: *Machine Learning with R: Learn Techniques for Building and Improving Machine Learning Models, from Data Preparation to Model Tuning, Evaluation, and Working with Big Data*. Packt Publishing, 2023. Citado na página 83.
- [29] Akobeng, Anthony K.: *Understanding diagnostic tests 3: Receiver operating characteristic curves*. Acta Paediatrica, 96(5):644–647, May 2007. Epub 2007 Mar 21. Citado na página 85.
- [30] Raschka, Sebastian: *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. arXiv preprint arXiv:1811.12808, 2018. <https://arxiv.org/abs/1811.12808>, Submitted on 13 Nov 2018, last revised 11 Nov 2020 (version v3). Citado na página 87.

- [31] Parvande, Saeid, Hung Wen Yeh, Martin P Paulus e Brett A McKinney: *Consensus features nested cross-validation*. Bioinformatics, 36(10):3093–3098, 2020. Citado 2 vezes nas páginas 90 e 91.
- [32] Bishop, Christopher M.: *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag New York, 2006, ISBN 978-0-387-31073-2. Citado 5 vezes nas páginas 95, 96, 107, 108 e 109.
- [33] Maalouf, Maher: *Logistic Regression in Data Analysis: An Overview*. International Journal of Data Analysis Techniques and Strategies, 3(3):281–299, 2011. Citado 2 vezes nas páginas 97 e 98.
- [34] Salman, H. A., A. Kalakech e A. Steiti: *Random Forest Algorithm Overview*. Babylonian Journal of Machine Learning, páginas 69–79, 2024. <https://doi.org/10.58496/BJML/2024/007>. Citado 2 vezes nas páginas 109 e 110.
- [35] Chen, Tianqi e Carlos Guestrin: *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, páginas 785–794, 2016. Citado na página 111.
- [36] Marsh, B.: *Multivariate Analysis of the Vector Boson Fusion Higgs Boson*, 2016. Citado na página 112.
- [37] Leeftang, M.M.G. e F. Allerberger: *How to: evaluate a diagnostic test*. Clinical Microbiology and Infection, 25(1):54–59, 2019. Citado 2 vezes nas páginas 115 e 116.
- [38] Woollam, J. A.: *About CompleteEASE*, 2023. <https://www.jawoollam.com/ellipsometry-software/completeease>. Citado na página 119.
- [39] Akiba, T., S. Sano, T. Yanase, T. Ohta e M. Koyama: *Optuna: A next-generation hyperparameter optimization framework*. Em *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. Citado na página 122.
- [40] Ahsan, M. M., M. A. P. Mahmud, P. K. Saha, K. D. Gupta e Z. Siddique: *Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance*. Technologies, 9(3):52, 2021. <https://doi.org/10.3390/technologies9030052>. Citado na página 122.
- [41] Zhao, Yue, Zain Nasrullah e Zheng Li: *PyOD: A Python Toolbox for Scalable Outlier Detection*. Journal of Machine Learning Research, 20(96):1–7, 2019. Citado na página 122.

-
- [42] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall e W. Philip Kegelmeyer: *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16:321–357, 2002. Citado na página 122.

Apêndices

APÊNDICE A – Tabelas de espaços de busca para parâmetros e hiperparâmetros

Tabela 6 – Parâmetros e espaço de busca utilizados durante a otimização do modelo.

Parâmetros	Espaço de busca
Ângulo	45, 50, 55, 60, 65, 70
Comprimento de onda mínimo ($wl\ min$)	245 - 1660
Comprimento de onda máximo ($wl\ max$)	$(wl\ min + 20) - 1700$
Número de comprimentos de onda	1 - (comprimentos de onda entre $wl\ max$ e $wl\ min$)

Tabela 7 – Hiperparâmetros e espaço de busca utilizados durante a otimização do modelo.

Modelos	Hyperparametros	Espaço de busca
SVM	C <i>Gamma</i> <i>weight</i>	1 - 10000 0.0001 - 0.01 0.1 - 0.9
MLP	hln1 (neurônios na camada 1) hln2 (neurônios na camada oculta 2) <i>solver</i> <i>activation</i> <i>momentum</i> <i>learning rate initial</i>	1 - $2 \times n.^o$ de comprimentos de onda 1 - $2 \times n.^o$ de comprimentos de onda sgd, adam <i>logistic</i> , tanh, relu 0.001 - 0.1 0.001 - 0.1
LR	C <i>penalty</i> <i>class weight</i>	0.01 - 10 11, 12 0.05 - 0.95
KNN	<i>n. neighbors</i> <i>weights</i> p	2 - 3 <i>uniform</i> , <i>distance</i> 1, 2, 3
XGBoost	<i>n. estimators</i> <i>learning rate</i> <i>max. depth</i> <i>subsample</i> <i>colsample bytree</i> <i>colsample bylevel</i> <i>colsample bynode</i> <i>scale pos weight</i>	1 - 1000 0.001 - 0.9 1 - 10 0.1 - 1.0 0.1 - 1.0 0.1 - 1.0 0.1 - 1.0 0.1 - 2.0

APÊNDICE B – Tabelas de resultados de todos modelos dos grupos amostrais

Tabela 8 – Resultados dos modelos não endêmicos

Legenda: AUC: *Area under ROC curve*, Acr: Acurácia, MCC: *Matthews correlation coefficient*, Espc: Especificidade, Sens: Sensibilidade.

Grupo amostral	Algoritmo	Propriedade	AUC	Acr	F1	MCC	Espc.	Sens.	Threshold
Não endêmicos	KNN	AnE	0.86 ± 0.03	0.79 ± 0.05	0.84 ± 0.04	0.5 ± 0.1	0.8 ± 0.1	0.80 ± 0.04	0.31 ± 0.01
Não endêmicos	KNN	Aps	0.83 ± 0.04	0.75 ± 0.05	0.80 ± 0.04	0.49 ± 0.08	0.77 ± 0.07	0.74 ± 0.07	0.33 ± 0.01
Não endêmicos	KNN	Asp	0.70 ± 0.08	0.33 ± 0.02	0.05 ± 0.05	0.05 ± 0.09	0.99 ± 0.03	0.03 ± 0.03	0.48 ± 0.01
Não endêmicos	KNN	dPolE	0.90 ± 0.04	0.79 ± 0.05	0.86 ± 0.03	0.5 ± 0.1	0.5 ± 0.1	0.94 ± 0.04	0.36 ± 0.05
Não endêmicos	KNN	Psi	0.92 ± 0.03	0.83 ± 0.05	0.87 ± 0.04	0.6 ± 0.1	0.75 ± 0.09	0.86 ± 0.06	0.31 ± 0.01
Não endêmicos	KNN	mm12	0.95 ± 0.03	0.76 ± 0.05	0.79 ± 0.06	0.61 ± 0.07	1.00 ± 0.02	0.66 ± 0.08	0.65 ± 0.03
Não endêmicos	KNN	mm13	0.82 ± 0.07	0.75 ± 0.06	0.83 ± 0.04	0.4 ± 0.2	0.5 ± 0.1	0.89 ± 0.05	0.5 ± 0.0
Não endêmicos	KNN	mm14	0.88 ± 0.05	0.54 ± 0.05	0.50 ± 0.08	0.35 ± 0.07	0.99 ± 0.04	0.34 ± 0.07	0.33 ± 0.01
Não endêmicos	KNN	mm21	0.94 ± 0.03	0.80 ± 0.05	0.83 ± 0.05	0.65 ± 0.07	0.98 ± 0.04	0.72 ± 0.08	0.62 ± 0.05
Não endêmicos	KNN	mm22	0.91 ± 0.03	0.80 ± 0.05	0.87 ± 0.03	0.5 ± 0.1	0.5 ± 0.1	0.95 ± 0.03	0.35 ± 0.03
Não endêmicos	KNN	mm23	0.84 ± 0.05	0.76 ± 0.06	0.83 ± 0.04	0.4 ± 0.2	0.6 ± 0.2	0.85 ± 0.04	0.5 ± 0.0
Não endêmicos	KNN	mm24	0.87 ± 0.05	0.78 ± 0.05	0.83 ± 0.05	0.6 ± 0.1	0.82 ± 0.08	0.76 ± 0.07	0.32 ± 0.01
Não endêmicos	KNN	mm31	0.85 ± 0.04	0.77 ± 0.05	0.83 ± 0.03	0.5 ± 0.1	0.7 ± 0.1	0.80 ± 0.04	0.46 ± 0.01
Não endêmicos	KNN	mm32	0.85 ± 0.04	0.73 ± 0.04	0.83 ± 0.03	0.3 ± 0.1	0.3 ± 0.1	0.93 ± 0.03	0.31 ± 0.02
Não endêmicos	KNN	mm33	0.93 ± 0.03	0.80 ± 0.05	0.87 ± 0.03	0.5 ± 0.1	0.5 ± 0.1	0.94 ± 0.02	0.32 ± 0.02
Não endêmicos	KNN	mm34	0.57 ± 0.08	0.57 ± 0.09	0.66 ± 0.08	0.1 ± 0.2	0.5 ± 0.1	0.6 ± 0.1	0.45 ± 0.01
Não endêmicos	KNN	MM ₁₁	0.95 ± 0.04	0.87 ± 0.05	0.91 ± 0.04	0.7 ± 0.1	0.7 ± 0.2	0.93 ± 0.07	0.50 ± 0.06
Não endêmicos	LR	AnE	0.95 ± 0.02	0.73 ± 0.03	0.83 ± 0.01	0.3 ± 0.1	0.24 ± 0.08	0.95 ± 0.01	0.30 ± 0.02
Não endêmicos	LR	Aps	0.91 ± 0.02	0.74 ± 0.04	0.84 ± 0.02	0.3 ± 0.2	0.2 ± 0.1	1.0 ± 0.0	0.11 ± 0.03
Não endêmicos	LR	Asp	0.80 ± 0.08	0.69 ± 0.01	0.82 ± 0.01	0.02 ± 0.07	0.01 ± 0.03	1.00 ± 0.02	0.23 ± 0.05
Não endêmicos	LR	dPolE	0.92 ± 0.04	0.86 ± 0.03	0.91 ± 0.03	0.68 ± 0.08	0.67 ± 0.06	0.95 ± 0.05	0.33 ± 0.03
Não endêmicos	LR	Psi	0.94 ± 0.01	0.78 ± 0.04	0.86 ± 0.02	0.4 ± 0.1	0.4 ± 0.1	0.9 ± 0.0	0.32 ± 0.02
Não endêmicos	LR	mm12	0.96 ± 0.03	0.87 ± 0.06	0.91 ± 0.04	0.7 ± 0.2	0.7 ± 0.2	0.96 ± 0.03	0.34 ± 0.04
Não endêmicos	LR	mm13	0.89 ± 0.05	0.78 ± 0.03	0.86 ± 0.02	0.46 ± 0.09	0.36 ± 0.06	0.97 ± 0.03	0.27 ± 0.02
Não endêmicos	LR	mm14	0.86 ± 0.03	0.82 ± 0.01	0.89 ± 0.01	0.58 ± 0.04	0.43 ± 0.03	1.00 ± 0.01	0.21 ± 0.05
Não endêmicos	LR	mm21	0.89 ± 0.02	0.80 ± 0.02	0.87 ± 0.02	0.51 ± 0.07	0.44 ± 0.05	0.96 ± 0.03	0.17 ± 0.02
Não endêmicos	LR	mm22	0.90 ± 0.02	0.82 ± 0.04	0.88 ± 0.03	0.6 ± 0.1	0.5 ± 0.1	0.97 ± 0.02	0.31 ± 0.04
Não endêmicos	LR	mm23	0.84 ± 0.05	0.74 ± 0.04	0.83 ± 0.03	0.3 ± 0.1	0.3 ± 0.1	0.92 ± 0.04	0.28 ± 0.02
Não endêmicos	LR	mm24	0.95 ± 0.04	0.89 ± 0.03	0.92 ± 0.02	0.73 ± 0.07	0.6 ± 0.1	1.0 ± 0.0	0.33 ± 0.05
Não endêmicos	LR	mm31	0.82 ± 0.03	0.75 ± 0.04	0.83 ± 0.03	0.4 ± 0.1	0.4 ± 0.1	0.88 ± 0.04	0.23 ± 0.03
Não endêmicos	LR	mm32	0.85 ± 0.04	0.76 ± 0.05	0.82 ± 0.04	0.5 ± 0.1	0.8 ± 0.1	0.77 ± 0.07	0.61 ± 0.03
Não endêmicos	LR	mm33	0.85 ± 0.03	0.81 ± 0.05	0.86 ± 0.03	0.6 ± 0.1	0.7 ± 0.1	0.87 ± 0.05	0.47 ± 0.03
Não endêmicos	LR	mm34	0.73 ± 0.05	0.76 ± 0.04	0.85 ± 0.03	0.4 ± 0.2	0.3 ± 0.1	0.98 ± 0.03	0.37 ± 0.03
Não endêmicos	LR	MM ₁₁	0.97 ± 0.02	0.92 ± 0.04	0.94 ± 0.03	0.81 ± 0.08	0.83 ± 0.06	0.95 ± 0.05	0.51 ± 0.03
Não endêmicos	MLP	AnE	0.90 ± 0.03	0.86 ± 0.04	0.90 ± 0.03	0.66 ± 0.09	0.7 ± 0.1	0.92 ± 0.04	0.50 ± 0.02
Não endêmicos	MLP	Aps	0.89 ± 0.04	0.72 ± 0.03	0.83 ± 0.02	0.2 ± 0.2	0.1 ± 0.1	1.00 ± 0.01	0.18 ± 0.05
Não endêmicos	MLP	Asp	0.75 ± 0.05	0.69 ± 0.02	0.82 ± 0.01	0.0 ± 0.1	0.03 ± 0.05	0.99 ± 0.02	0.15 ± 0.05
Não endêmicos	MLP	dPolE	0.87 ± 0.04	0.78 ± 0.04	0.84 ± 0.03	0.5 ± 0.1	0.7 ± 0.1	0.83 ± 0.05	0.49 ± 0.02
Não endêmicos	MLP	Psi	0.89 ± 0.04	0.78 ± 0.05	0.86 ± 0.03	0.4 ± 0.2	0.3 ± 0.2	0.98 ± 0.03	0.34 ± 0.09
Não endêmicos	MLP	mm12	0.84 ± 0.03	0.84 ± 0.04	0.89 ± 0.03	0.6 ± 0.1	0.5 ± 0.1	0.96 ± 0.03	0.43 ± 0.06
Não endêmicos	MLP	mm13	0.87 ± 0.04	0.75 ± 0.02	0.84 ± 0.01	0.37 ± 0.08	0.26 ± 0.09	0.97 ± 0.03	0.32 ± 0.08
Não endêmicos	MLP	mm14	0.95 ± 0.04	0.69 ± 0.01	0.82 ± 0.01	0.03 ± 0.08	0.01 ± 0.03	1.0 ± 0.0	0.15 ± 0.03
Não endêmicos	MLP	mm21	0.91 ± 0.03	0.73 ± 0.04	0.83 ± 0.02	0.2 ± 0.2	0.1 ± 0.1	0.99 ± 0.02	0.11 ± 0.05
Não endêmicos	MLP	mm22	0.95 ± 0.02	0.90 ± 0.02	0.93 ± 0.01	0.76 ± 0.05	0.77 ± 0.05	0.96 ± 0.02	0.42 ± 0.03
Não endêmicos	MLP	mm23	0.83 ± 0.04	0.76 ± 0.04	0.83 ± 0.03	0.4 ± 0.1	0.5 ± 0.1	0.87 ± 0.05	0.44 ± 0.04
Não endêmicos	MLP	mm24	0.93 ± 0.04	0.79 ± 0.06	0.87 ± 0.03	0.5 ± 0.2	0.4 ± 0.2	0.99 ± 0.02	0.34 ± 0.04
Não endêmicos	MLP	mm31	0.86 ± 0.03	0.71 ± 0.03	0.82 ± 0.02	0.2 ± 0.1	0.12 ± 0.06	0.98 ± 0.03	0.15 ± 0.05
Não endêmicos	MLP	mm32	0.79 ± 0.03	0.74 ± 0.04	0.80 ± 0.04	0.43 ± 0.09	0.7 ± 0.1	0.77 ± 0.06	0.50 ± 0.02
Não endêmicos	MLP	mm33	0.92 ± 0.03	0.84 ± 0.05	0.89 ± 0.03	0.6 ± 0.1	0.7 ± 0.1	0.91 ± 0.04	0.46 ± 0.03
Não endêmicos	MLP	mm34	0.74 ± 0.06	0.74 ± 0.03	0.84 ± 0.02	0.34 ± 0.09	0.27 ± 0.09	0.96 ± 0.02	0.32 ± 0.03
Não endêmicos	MLP	MM ₁₁	0.99 ± 0.01	0.93 ± 0.02	0.95 ± 0.02	0.84 ± 0.05	0.77 ± 0.08	1.0 ± 0.0	0.44 ± 0.04
Não endêmicos	SVM	AnE	0.93 ± 0.03	0.69 ± 0.03	0.82 ± 0.01	0.1 ± 0.2	0.07 ± 0.09	0.98 ± 0.03	0.33 ± 0.06
Não endêmicos	SVM	Aps	0.89 ± 0.03	0.81 ± 0.04	0.87 ± 0.03	0.5 ± 0.1	0.6 ± 0.2	0.92 ± 0.06	0.52 ± 0.06
Não endêmicos	SVM	Asp	0.76 ± 0.06	0.69 ± 0.02	0.82 ± 0.01	0.1 ± 0.1	0.04 ± 0.05	0.99 ± 0.02	0.36 ± 0.04
Não endêmicos	SVM	dPolE	0.96 ± 0.04	0.83 ± 0.04	0.89 ± 0.02	0.6 ± 0.1	0.5 ± 0.1	1.00 ± 0.01	0.37 ± 0.04
Não endêmicos	SVM	Psi	0.90 ± 0.05	0.75 ± 0.06	0.85 ± 0.03	0.3 ± 0.2	0.2 ± 0.2	0.98 ± 0.02	0.39 ± 0.04
Não endêmicos	SVM	mm12	0.90 ± 0.03	0.83 ± 0.04	0.89 ± 0.03	0.6 ± 0.1	0.5 ± 0.1	0.98 ± 0.03	0.40 ± 0.04
Não endêmicos	SVM	mm13	0.86 ± 0.04	0.73 ± 0.03	0.84 ± 0.01	0.3 ± 0.1	0.14 ± 0.08	1.00 ± 0.01	0.35 ± 0.04
Não endêmicos	SVM	mm14	0.92 ± 0.03	0.86 ± 0.04	0.91 ± 0.02	0.68 ± 0.09	0.6 ± 0.1	0.99 ± 0.02	0.47 ± 0.05
Não endêmicos	SVM	mm21	0.94 ± 0.03	0.81 ± 0.05	0.87 ± 0.03	0.5 ± 0.1	0.4 ± 0.1	0.98 ± 0.04	0.37 ± 0.02
Não endêmicos	SVM	mm22	0.94 ± 0.02	0.78 ± 0.05	0.86 ± 0.03	0.4 ± 0.1	0.3 ± 0.2	0.98 ± 0.02	0.39 ± 0.05
Não endêmicos	SVM	mm23	0.90 ± 0.03	0.75 ± 0.03	0.84 ± 0.02	0.3 ± 0.1	0.2 ± 0.1	0.98 ± 0.02	0.34 ± 0.04
Não endêmicos	SVM	mm24	0.96 ± 0.03	0.78 ± 0.03	0.86 ± 0.02	0.46 ± 0.09	0.3 ± 0.1	1.0 ± 0.0	0.43 ± 0.06
Não endêmicos	SVM	mm31	0.92 ± 0.04	0.75 ± 0.03	0.85 ± 0.01	0.4 ± 0.1	0.19 ± 0.09	1.00 ± 0.01	0.33 ± 0.03
Não endêmicos	SVM	mm32	0.94 ± 0.03	0.75 ± 0.03	0.84 ± 0.02	0.3 ± 0.2	0.2 ± 0.1	1.0 ± 0.0	0.31 ± 0.04
Não endêmicos	SVM	mm33	0.86 ± 0.05	0.79 ± 0.04	0.86 ± 0.03	0.5 ± 0.1	0.53 ± 0.09	0.91 ± 0.05	0.45 ± 0.01
Não endêmicos	SVM	mm34	0.67 ± 0.05	0.69 ± 0.01	0.82 ± 0.01	0.0 ± 0.1	0.02 ± 0.04	1.00 ± 0.01	0.27 ± 0.03
Não endêmicos	SVM	MM ₁₁	0.98 ± 0.02	0.86 ± 0.04	0.91 ± 0.03	0.7 ± 0.1	0.6 ± 0.1	1.00 ± 0.01	0.33 ± 0.05
Não endêmicos	XGBoost	AnE	0.92 ± 0.03	0.84 ± 0.04	0.89 ± 0.02	0.6 ± 0.1	0.6 ± 0.1	0.95 ± 0.04	0.50 ± 0.03
Não endêmicos	XGBoost	Aps	0.86 ± 0.04	0.72 ± 0.02	0.83 ± 0.01	0.2 ± 0.2	0.11 ± 0.09	1.00 ± 0.02	0.33 ± 0.04
Não endêmicos	XGBoost	Asp	0.85 ± 0.03	0.70 ± 0.02	0.82 ± 0.01	0.1 ± 0.1	0.05 ± 0.07	1.0 ± 0.0	0.31 ± 0.04
Não endêmicos	XGBoost	dPolE	0.90 ± 0.02	0.78 ± 0.05	0.85 ± 0.03	0.5 ± 0.2	0.5 ± 0.2	0.90 ± 0.02	0.34 ± 0.05
Não endêmicos	XGBoost	Psi	0.93 ± 0.03	0.83 ± 0.04	0.88 ± 0.03	0.6 ± 0.1	0.6 ± 0.1	0.95 ± 0.04	0.47 ± 0.03
Não endêmicos	XGBoost	mm12	0.88 ± 0.04	0.79 ± 0.05	0.85 ± 0.04	0.5 ± 0.1	0.5 ± 0.1	0.91 ± 0.06	0.46 ± 0.05
Não endêmicos	XGBoost	mm13	0.90 ± 0.03	0.80 ± 0.04	0.87 ± 0.02	0.5 ± 0.1	0.4 ± 0.1	0.99 ± 0.02	0.31 ± 0.05
Não endêmicos	XGBoost	mm14	0.84 ± 0.05	0.80 ± 0.03	0.87 ± 0.01	0.51 ± 0.07	0.4 ± 0.1	0.98 ± 0.03	0.43 ± 0.08
Não endêmicos	XGBoost	mm21	0.89 ± 0.03	0.77 ± 0.04	0.84 ± 0.03	0.4 ± 0.1	0.43 ± 0.07	0.92 ± 0.06	0.36 ± 0.06
Não endêmicos	XGBoost	mm22	0.90 ± 0.03	0.80 ± 0.04	0.86 ± 0.02	0.5 ± 0.1	0.6 ± 0.1	0.91 ± 0.05	0.42 ± 0.08
Não endêmicos	XGBoost	mm23	0.83 ± 0.04	0.75 ± 0.04	0.84 ± 0.02	0.4 ± 0.1	0.3 ± 0.1	0.97 ± 0.04	0.35 ± 0.04
Não endêmicos	XGBoost	mm24	0.93 ± 0.04	0.86 ± 0.04	0.91 ± 0.02	0.7 ± 0.1	0.6 ± 0.1	0.99 ± 0.01	0.57 ± 0.06
Não endêmicos	XGBoost	mm31	0.88 ± 0.02	0.77 ± 0.04	0.85 ± 0.02	0.4 ± 0.1	0.4 ± 0.1	0.92 ± 0.04	0.37 ± 0.04
Não endêmicos	XGBoost	mm32	0.87 ± 0.03	0.77 ± 0.05	0.86 ± 0.03	0.4 ± 0.1	0.3 ± 0.1	0.97 ± 0.03	0.33 ± 0.04
Não endêmicos	XGBoost	mm33	0.86 ± 0.04	0.74 ± 0.06	0.83 ± 0.04	0.4 ± 0.2	0.5 ± 0.2	0.87 ± 0.04	0.50 ± 0.04
Não endêmicos	XGBoost	mm34	0.72 ± 0.05	0.73 ± 0.03	0.82 ± 0.02	0.31 ± 0.09	0.33 ± 0.07	0.91 ± 0.04	0.45 ± 0.05
Não endêmicos	XGBoost	MM ₁₁	0.93 ± 0.02	0.83 ± 0.03	0.89 ± 0.02	0.59 ± 0.07	0.49 ± 0.06	0.98 ± 0.02	0.32 ± 0.04

Tabela 9 – Resultados dos modelos reativos

Grupo amostral	Algoritmo	Propriedade	AUC	Acr	F1	MCC	Espec.	Sens.	Threshold
reativos	KNN	AnE	0.69 ± 0.07	0.63 ± 0.06	0.55 ± 0.07	0.2 ± 0.1	0.69 ± 0.08	0.54 ± 0.08	0.56 ± 0.01
reativos	KNN	Aps	0.91 ± 0.03	0.80 ± 0.06	0.80 ± 0.05	0.63 ± 0.09	0.7 ± 0.1	0.92 ± 0.06	0.53 ± 0.03
reativos	KNN	Asp	0.68 ± 0.06	0.47 ± 0.04	0.61 ± 0.02	0.1 ± 0.1	0.09 ± 0.06	0.98 ± 0.03	0.5 ± 0.0
reativos	KNN	dPolE	0.94 ± 0.01	0.91 ± 0.03	0.90 ± 0.03	0.82 ± 0.06	0.90 ± 0.01	0.93 ± 0.06	0.52 ± 0.04
reativos	KNN	Psi	0.70 ± 0.07	0.62 ± 0.07	0.54 ± 0.07	0.2 ± 0.1	0.7 ± 0.1	0.53 ± 0.09	0.56 ± 0.01
reativos	KNN	mm12	0.97 ± 0.01	0.91 ± 0.02	0.90 ± 0.02	0.82 ± 0.04	0.91 ± 0.02	0.91 ± 0.04	0.52 ± 0.03
reativos	KNN	mm13	0.90 ± 0.03	0.82 ± 0.04	0.78 ± 0.05	0.64 ± 0.08	0.89 ± 0.04	0.74 ± 0.07	0.55 ± 0.02
reativos	KNN	mm14	0.61 ± 0.09	0.54 ± 0.07	0.60 ± 0.05	0.2 ± 0.1	0.3 ± 0.2	0.8 ± 0.1	0.54 ± 0.03
reativos	KNN	mm21	0.96 ± 0.01	0.90 ± 0.03	0.88 ± 0.04	0.80 ± 0.06	0.95 ± 0.01	0.84 ± 0.07	0.54 ± 0.03
reativos	KNN	mm22	0.95 ± 0.01	0.91 ± 0.03	0.89 ± 0.04	0.81 ± 0.07	0.90 ± 0.02	0.92 ± 0.08	0.54 ± 0.04
reativos	KNN	mm23	0.89 ± 0.03	0.81 ± 0.04	0.76 ± 0.05	0.62 ± 0.09	0.90 ± 0.05	0.70 ± 0.07	0.57 ± 0.02
reativos	KNN	mm24	0.87 ± 0.04	0.74 ± 0.07	0.74 ± 0.05	0.5 ± 0.1	0.6 ± 0.2	0.86 ± 0.09	0.56 ± 0.04
reativos	KNN	mm31	0.84 ± 0.04	0.78 ± 0.05	0.74 ± 0.05	0.5 ± 0.1	0.80 ± 0.06	0.74 ± 0.07	0.5 ± 0.0
reativos	KNN	mm32	0.90 ± 0.03	0.74 ± 0.06	0.6 ± 0.1	0.5 ± 0.1	0.97 ± 0.04	0.4 ± 0.2	0.65 ± 0.04
reativos	KNN	mm33	0.92 ± 0.02	0.86 ± 0.04	0.84 ± 0.04	0.72 ± 0.08	0.84 ± 0.07	0.88 ± 0.05	0.56 ± 0.02
reativos	KNN	mm34	0.74 ± 0.06	0.63 ± 0.05	0.4 ± 0.1	0.2 ± 0.1	0.89 ± 0.07	0.3 ± 0.1	0.65 ± 0.02
reativos	KNN	MM_{11}	0.95 ± 0.01	0.84 ± 0.04	0.78 ± 0.06	0.67 ± 0.07	0.9 ± 0.0	0.68 ± 0.09	0.63 ± 0.03
reativos	LR	AnE	0.82 ± 0.03	0.77 ± 0.06	0.73 ± 0.05	0.5 ± 0.1	0.8 ± 0.1	0.70 ± 0.05	0.43 ± 0.04
reativos	LR	Aps	0.99 ± 0.01	0.86 ± 0.05	0.81 ± 0.08	0.74 ± 0.09	0.98 ± 0.02	0.7 ± 0.1	0.81 ± 0.03
reativos	LR	Asp	0.79 ± 0.05	0.58 ± 0.01	0.03 ± 0.05	0.05 ± 0.08	1.0 ± 0.0	0.02 ± 0.03	0.81 ± 0.05
reativos	LR	dPolE	0.98 ± 0.01	0.88 ± 0.05	0.88 ± 0.04	0.78 ± 0.08	0.83 ± 0.08	0.96 ± 0.05	0.20 ± 0.02
reativos	LR	Psi	0.82 ± 0.03	0.72 ± 0.04	0.6 ± 0.1	0.43 ± 0.09	0.91 ± 0.05	0.5 ± 0.1	0.61 ± 0.04
reativos	LR	mm12	0.97 ± 0.01	0.92 ± 0.02	0.90 ± 0.03	0.83 ± 0.05	0.91 ± 0.03	0.92 ± 0.03	0.61 ± 0.02
reativos	LR	mm13	0.99 ± 0.01	0.85 ± 0.05	0.85 ± 0.04	0.74 ± 0.07	0.74 ± 0.09	1.0 ± 0.0	0.16 ± 0.02
reativos	LR	mm14	0.90 ± 0.04	0.62 ± 0.05	0.2 ± 0.2	0.2 ± 0.2	0.99 ± 0.02	0.1 ± 0.1	0.75 ± 0.06
reativos	LR	mm21	0.96 ± 0.01	0.89 ± 0.03	0.87 ± 0.03	0.78 ± 0.05	0.91 ± 0.04	0.87 ± 0.05	0.49 ± 0.03
reativos	LR	mm22	0.98 ± 0.01	0.90 ± 0.04	0.87 ± 0.06	0.81 ± 0.07	0.98 ± 0.02	0.80 ± 0.09	0.74 ± 0.04
reativos	LR	mm23	0.98 ± 0.01	0.86 ± 0.04	0.81 ± 0.06	0.72 ± 0.07	0.9 ± 0.0	0.74 ± 0.09	0.74 ± 0.03
reativos	LR	mm24	0.96 ± 0.02	0.69 ± 0.07	0.4 ± 0.2	0.4 ± 0.1	1.0 ± 0.0	0.3 ± 0.2	0.73 ± 0.05
reativos	LR	mm31	0.97 ± 0.01	0.73 ± 0.06	0.5 ± 0.1	0.5 ± 0.1	1.0 ± 0.0	0.4 ± 0.1	0.78 ± 0.03
reativos	LR	mm32	0.91 ± 0.02	0.76 ± 0.07	0.78 ± 0.04	0.60 ± 0.09	0.6 ± 0.1	0.99 ± 0.02	0.13 ± 0.02
reativos	LR	mm33	0.98 ± 0.01	0.90 ± 0.03	0.88 ± 0.04	0.80 ± 0.06	0.92 ± 0.03	0.88 ± 0.08	0.64 ± 0.03
reativos	LR	mm34	0.66 ± 0.06	0.54 ± 0.06	0.60 ± 0.04	0.2 ± 0.1	0.3 ± 0.1	0.81 ± 0.08	0.08 ± 0.01
reativos	LR	MM_{11}	0.99 ± 0.01	0.93 ± 0.03	0.92 ± 0.04	0.86 ± 0.06	0.97 ± 0.02	0.88 ± 0.06	0.54 ± 0.03
reativos	MLP	AnE	0.82 ± 0.04	0.75 ± 0.06	0.73 ± 0.05	0.5 ± 0.1	0.7 ± 0.1	0.78 ± 0.08	0.52 ± 0.03
reativos	MLP	Aps	0.96 ± 0.03	0.72 ± 0.09	0.5 ± 0.3	0.4 ± 0.2	0.98 ± 0.03	0.4 ± 0.2	0.79 ± 0.08
reativos	MLP	Asp	0.74 ± 0.06	0.58 ± 0.01	0.03 ± 0.06	0.04 ± 0.09	1.0 ± 0.0	0.02 ± 0.03	0.83 ± 0.06
reativos	MLP	dPolE	0.97 ± 0.01	0.92 ± 0.02	0.92 ± 0.03	0.85 ± 0.05	0.89 ± 0.04	0.97 ± 0.04	0.48 ± 0.06
reativos	MLP	Psi	0.80 ± 0.04	0.72 ± 0.05	0.69 ± 0.05	0.44 ± 0.09	0.69 ± 0.08	0.75 ± 0.09	0.51 ± 0.03
reativos	MLP	mm12	0.97 ± 0.01	0.92 ± 0.02	0.90 ± 0.02	0.83 ± 0.04	0.92 ± 0.03	0.92 ± 0.04	0.48 ± 0.04
reativos	MLP	mm13	0.98 ± 0.01	0.86 ± 0.05	0.80 ± 0.09	0.74 ± 0.09	0.99 ± 0.02	0.7 ± 0.1	0.74 ± 0.06
reativos	MLP	mm14	0.99 ± 0.01	0.87 ± 0.07	0.8 ± 0.1	0.8 ± 0.1	0.98 ± 0.05	0.7 ± 0.2	0.58 ± 0.05
reativos	MLP	mm21	0.95 ± 0.02	0.92 ± 0.02	0.90 ± 0.03	0.83 ± 0.05	0.95 ± 0.02	0.88 ± 0.05	0.55 ± 0.03
reativos	MLP	mm22	0.97 ± 0.01	0.92 ± 0.02	0.91 ± 0.02	0.84 ± 0.04	0.90 ± 0.03	0.94 ± 0.03	0.44 ± 0.04
reativos	MLP	mm23	0.98 ± 0.01	0.83 ± 0.05	0.8 ± 0.1	0.7 ± 0.1	0.98 ± 0.03	0.6 ± 0.1	0.69 ± 0.07
reativos	MLP	mm24	0.99 ± 0.01	0.94 ± 0.04	0.92 ± 0.06	0.88 ± 0.08	0.96 ± 0.04	0.9 ± 0.1	0.55 ± 0.04
reativos	MLP	mm31	0.95 ± 0.03	0.88 ± 0.05	0.84 ± 0.07	0.8 ± 0.1	0.93 ± 0.07	0.8 ± 0.1	0.54 ± 0.05
reativos	MLP	mm32	0.86 ± 0.02	0.78 ± 0.03	0.73 ± 0.05	0.55 ± 0.07	0.84 ± 0.04	0.70 ± 0.07	0.60 ± 0.04
reativos	MLP	mm33	0.95 ± 0.01	0.89 ± 0.03	0.88 ± 0.04	0.79 ± 0.06	0.88 ± 0.04	0.91 ± 0.07	0.52 ± 0.08
reativos	MLP	mm34	0.70 ± 0.05	0.67 ± 0.05	0.4 ± 0.1	0.3 ± 0.1	0.95 ± 0.06	0.3 ± 0.1	0.67 ± 0.08
reativos	MLP	MM_{11}	0.98 ± 0.01	0.92 ± 0.04	0.91 ± 0.04	0.84 ± 0.08	0.91 ± 0.06	0.93 ± 0.07	0.47 ± 0.08
reativos	SVM	AnE	0.77 ± 0.05	0.63 ± 0.03	0.3 ± 0.1	0.3 ± 0.1	0.99 ± 0.02	0.16 ± 0.07	0.64 ± 0.03
reativos	SVM	Aps	0.96 ± 0.02	0.84 ± 0.04	0.79 ± 0.07	0.69 ± 0.08	0.95 ± 0.04	0.7 ± 0.1	0.59 ± 0.07
reativos	SVM	Asp	0.84 ± 0.03	0.60 ± 0.02	0.13 ± 0.09	0.2 ± 0.1	0.99 ± 0.02	0.08 ± 0.06	0.65 ± 0.03
reativos	SVM	dPolE	0.97 ± 0.01	0.92 ± 0.02	0.90 ± 0.03	0.83 ± 0.04	0.94 ± 0.02	0.89 ± 0.06	0.59 ± 0.05
reativos	SVM	Psi	0.74 ± 0.05	0.59 ± 0.04	0.2 ± 0.1	0.1 ± 0.1	0.95 ± 0.03	0.10 ± 0.08	0.63 ± 0.03
reativos	SVM	mm12	0.97 ± 0.01	0.94 ± 0.03	0.93 ± 0.03	0.87 ± 0.06	0.91 ± 0.03	0.97 ± 0.05	0.49 ± 0.06
reativos	SVM	mm13	0.97 ± 0.02	0.82 ± 0.04	0.74 ± 0.08	0.67 ± 0.08	0.99 ± 0.01	0.6 ± 0.1	0.59 ± 0.04
reativos	SVM	mm14	0.96 ± 0.02	0.71 ± 0.05	0.5 ± 0.1	0.5 ± 0.1	1.00 ± 0.01	0.3 ± 0.1	0.61 ± 0.03
reativos	SVM	mm21	0.98 ± 0.01	0.92 ± 0.03	0.91 ± 0.03	0.84 ± 0.06	0.93 ± 0.04	0.91 ± 0.06	0.54 ± 0.03
reativos	SVM	mm22	0.98 ± 0.01	0.92 ± 0.03	0.90 ± 0.04	0.83 ± 0.06	0.94 ± 0.03	0.89 ± 0.06	0.59 ± 0.05
reativos	SVM	mm23	0.94 ± 0.02	0.76 ± 0.05	0.6 ± 0.1	0.55 ± 0.09	0.97 ± 0.04	0.5 ± 0.1	0.61 ± 0.06
reativos	SVM	mm24	0.98 ± 0.01	0.83 ± 0.05	0.7 ± 0.1	0.67 ± 0.09	0.99 ± 0.01	0.6 ± 0.1	0.55 ± 0.04
reativos	SVM	mm31	0.96 ± 0.02	0.82 ± 0.04	0.74 ± 0.08	0.67 ± 0.07	1.0 ± 0.0	0.59 ± 0.09	0.61 ± 0.03
reativos	SVM	mm32	0.91 ± 0.02	0.77 ± 0.04	0.6 ± 0.1	0.55 ± 0.07	0.96 ± 0.03	0.5 ± 0.1	0.62 ± 0.05
reativos	SVM	mm33	0.98 ± 0.01	0.91 ± 0.03	0.90 ± 0.04	0.83 ± 0.06	0.94 ± 0.02	0.88 ± 0.08	0.52 ± 0.05
reativos	SVM	mm34	0.82 ± 0.06	0.72 ± 0.06	0.6 ± 0.1	0.4 ± 0.1	0.85 ± 0.07	0.5 ± 0.1	0.52 ± 0.02
reativos	SVM	MM_{11}	1.00 ± 0.01	0.97 ± 0.03	0.96 ± 0.04	0.94 ± 0.06	1.00 ± 0.01	0.93 ± 0.08	0.64 ± 0.05
reativos	XGBoost	AnE	0.74 ± 0.05	0.65 ± 0.05	0.4 ± 0.2	0.3 ± 0.1	0.94 ± 0.06	0.3 ± 0.1	0.70 ± 0.06
reativos	XGBoost	Aps	0.95 ± 0.01	0.86 ± 0.04	0.81 ± 0.07	0.71 ± 0.09	0.94 ± 0.02	0.8 ± 0.1	0.37 ± 0.04
reativos	XGBoost	Asp	0.77 ± 0.05	0.63 ± 0.04	0.3 ± 0.1	0.2 ± 0.1	0.95 ± 0.05	0.2 ± 0.1	0.36 ± 0.03
reativos	XGBoost	dPolE	0.95 ± 0.01	0.85 ± 0.04	0.80 ± 0.06	0.69 ± 0.07	0.94 ± 0.03	0.72 ± 0.09	0.63 ± 0.07
reativos	XGBoost	Psi	0.71 ± 0.05	0.62 ± 0.04	0.2 ± 0.2	0.2 ± 0.2	0.99 ± 0.02	0.1 ± 0.1	0.78 ± 0.06
reativos	XGBoost	mm12	0.94 ± 0.02	0.81 ± 0.04	0.73 ± 0.07	0.62 ± 0.08	0.94 ± 0.05	0.6 ± 0.1	0.72 ± 0.05
reativos	XGBoost	mm13	0.91 ± 0.02	0.82 ± 0.02	0.77 ± 0.03	0.64 ± 0.05	0.91 ± 0.04	0.70 ± 0.06	0.62 ± 0.05
reativos	XGBoost	mm14	0.88 ± 0.04	0.76 ± 0.08	0.6 ± 0.2	0.5 ± 0.2	0.89 ± 0.08	0.6 ± 0.2	0.55 ± 0.07
reativos	XGBoost	mm21	0.93 ± 0.02	0.79 ± 0.04	0.71 ± 0.08	0.58 ± 0.08	0.93 ± 0.04	0.6 ± 0.1	0.62 ± 0.04
reativos	XGBoost	mm22	0.93 ± 0.02	0.83 ± 0.03	0.77 ± 0.05	0.65 ± 0.06	0.94 ± 0.02	0.68 ± 0.07	0.70 ± 0.05
reativos	XGBoost	mm23	0.95 ± 0.03	0.71 ± 0.05	0.5 ± 0.1	0.5 ± 0.1	1.0 ± 0.0	0.3 ± 0.1	0.68 ± 0.05
reativos	XGBoost	mm24	0.96 ± 0.03	0.79 ± 0.08	0.7 ± 0.2	0.6 ± 0.2	0.97 ± 0.04	0.5 ± 0.2	0.62 ± 0.08
reativos	XGBoost	mm31	0.93 ± 0.03	0.74 ± 0.05	0.6 ± 0.1	0.50 ± 0.09	0.97 ± 0.03	0.4 ± 0.1	0.67 ± 0.05
reativos	XGBoost	mm32	0.91 ± 0.02	0.84 ± 0.05	0.80 ± 0.07	0.7 ± 0.1	0.90 ± 0.05	0.8 ± 0.1	0.56 ± 0.06
reativos	XGBoost	mm33	0.92 ± 0.02	0.78 ± 0.08	0.7 ± 0.1	0.5 ± 0.2	0.91 ± 0.04	0.6 ± 0.2	0.67 ± 0.05
reativos	XGBoost	mm34	0.70 ± 0.07	0.63 ± 0.07	0.4 ± 0.2	0.2 ± 0.2	0.9 ± 0.1	0.3 ± 0.2	0.57 ± 0.08
reativos	XGBoost	MM_{11}	0.97 ± 0.01	0.91 ± 0.03	0.89 ± 0.04	0.82 ± 0.06	0.94 ± 0.02	0.87 ± 0.07	0.49 ± 0.07

Tabela 10 – Resultados dos modelos endêmicos

Grupo amostral	Algoritmo	Propriedade	AUC	Acr	F1	MCC	Espc.	Sens.	Threshold
endêmicos	KNN	AnE	0.80 ± 0.05	0.73 ± 0.05	0.72 ± 0.05	0.5 ± 0.1	0.81 ± 0.08	0.66 ± 0.07	0.5 ± 0.0
endêmicos	KNN	Aps	0.85 ± 0.05	0.77 ± 0.06	0.79 ± 0.06	0.5 ± 0.1	0.71 ± 0.08	0.82 ± 0.09	0.49 ± 0.01
endêmicos	KNN	Asp	0.63 ± 0.06	0.53 ± 0.04	0.30 ± 0.09	0.2 ± 0.1	0.93 ± 0.05	0.19 ± 0.07	0.37 ± 0.04
endêmicos	KNN	dPolE	0.74 ± 0.05	0.72 ± 0.06	0.73 ± 0.07	0.4 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.50 ± 0.03
endêmicos	KNN	Psi	0.79 ± 0.06	0.70 ± 0.05	0.69 ± 0.06	0.4 ± 0.1	0.75 ± 0.07	0.65 ± 0.08	0.5 ± 0.0
endêmicos	KNN	mm12	0.83 ± 0.04	0.78 ± 0.06	0.80 ± 0.05	0.6 ± 0.1	0.7 ± 0.1	0.84 ± 0.07	0.49 ± 0.02
endêmicos	KNN	mm13	0.82 ± 0.07	0.76 ± 0.08	0.76 ± 0.08	0.5 ± 0.1	0.77 ± 0.07	0.7 ± 0.1	0.5 ± 0.0
endêmicos	KNN	mm14	0.69 ± 0.08	0.62 ± 0.08	0.5 ± 0.1	0.3 ± 0.2	0.8 ± 0.1	0.4 ± 0.1	0.32 ± 0.01
endêmicos	KNN	mm21	0.75 ± 0.06	0.71 ± 0.06	0.71 ± 0.07	0.4 ± 0.1	0.76 ± 0.08	0.67 ± 0.09	0.5 ± 0.0
endêmicos	KNN	mm22	0.76 ± 0.05	0.75 ± 0.04	0.77 ± 0.04	0.49 ± 0.09	0.68 ± 0.08	0.80 ± 0.07	0.50 ± 0.01
endêmicos	KNN	mm23	0.84 ± 0.05	0.74 ± 0.06	0.77 ± 0.05	0.5 ± 0.1	0.6 ± 0.1	0.84 ± 0.07	0.48 ± 0.02
endêmicos	KNN	mm24	0.68 ± 0.07	0.52 ± 0.03	0.20 ± 0.09	0.2 ± 0.1	0.99 ± 0.02	0.11 ± 0.05	0.48 ± 0.01
endêmicos	KNN	mm31	0.71 ± 0.08	0.67 ± 0.09	0.70 ± 0.09	0.3 ± 0.2	0.6 ± 0.1	0.7 ± 0.1	0.48 ± 0.01
endêmicos	KNN	mm32	0.72 ± 0.05	0.69 ± 0.06	0.73 ± 0.05	0.4 ± 0.1	0.5 ± 0.1	0.81 ± 0.08	0.48 ± 0.01
endêmicos	KNN	mm33	0.79 ± 0.07	0.74 ± 0.06	0.76 ± 0.06	0.5 ± 0.1	0.74 ± 0.09	0.75 ± 0.08	0.5 ± 0.0
endêmicos	KNN	mm34	0.82 ± 0.07	0.61 ± 0.07	0.4 ± 0.2	0.3 ± 0.1	0.96 ± 0.08	0.3 ± 0.1	0.63 ± 0.04
endêmicos	KNN	MM ₁₁	0.80 ± 0.05	0.69 ± 0.05	0.72 ± 0.04	0.4 ± 0.1	0.66 ± 0.09	0.72 ± 0.06	0.50 ± 0.03
endêmicos	LR	AnE	0.85 ± 0.04	0.72 ± 0.07	0.7 ± 0.1	0.5 ± 0.1	0.8 ± 0.1	0.7 ± 0.2	0.55 ± 0.05
endêmicos	LR	Aps	0.81 ± 0.05	0.73 ± 0.07	0.77 ± 0.06	0.5 ± 0.1	0.6 ± 0.2	0.8 ± 0.1	0.45 ± 0.07
endêmicos	LR	Asp	0.81 ± 0.03	0.66 ± 0.08	0.6 ± 0.2	0.4 ± 0.1	0.8 ± 0.1	0.5 ± 0.2	0.63 ± 0.04
endêmicos	LR	dPolE	0.85 ± 0.04	0.81 ± 0.06	0.82 ± 0.06	0.6 ± 0.1	0.81 ± 0.08	0.8 ± 0.1	0.68 ± 0.03
endêmicos	LR	Psi	0.89 ± 0.03	0.81 ± 0.05	0.82 ± 0.06	0.6 ± 0.1	0.75 ± 0.07	0.9 ± 0.1	0.45 ± 0.03
endêmicos	LR	mm12	0.80 ± 0.03	0.73 ± 0.05	0.74 ± 0.05	0.5 ± 0.1	0.7 ± 0.1	0.74 ± 0.09	0.42 ± 0.05
endêmicos	LR	mm13	0.91 ± 0.04	0.82 ± 0.05	0.84 ± 0.04	0.6 ± 0.1	0.7 ± 0.1	0.89 ± 0.06	0.33 ± 0.03
endêmicos	LR	mm14	0.85 ± 0.05	0.61 ± 0.06	0.73 ± 0.03	0.3 ± 0.2	0.2 ± 0.1	1.00 ± 0.02	0.21 ± 0.06
endêmicos	LR	mm21	0.77 ± 0.06	0.72 ± 0.07	0.76 ± 0.06	0.4 ± 0.2	0.6 ± 0.2	0.80 ± 0.09	0.22 ± 0.03
endêmicos	LR	mm22	0.83 ± 0.03	0.75 ± 0.05	0.75 ± 0.05	0.50 ± 0.09	0.79 ± 0.09	0.70 ± 0.07	0.48 ± 0.05
endêmicos	LR	mm23	0.79 ± 0.04	0.65 ± 0.08	0.73 ± 0.04	0.3 ± 0.2	0.4 ± 0.2	0.88 ± 0.04	0.28 ± 0.04
endêmicos	LR	mm24	0.72 ± 0.05	0.63 ± 0.06	0.74 ± 0.04	0.3 ± 0.1	0.3 ± 0.1	0.95 ± 0.05	0.27 ± 0.05
endêmicos	LR	mm31	0.80 ± 0.07	0.69 ± 0.08	0.74 ± 0.06	0.4 ± 0.2	0.6 ± 0.2	0.81 ± 0.09	0.56 ± 0.03
endêmicos	LR	mm32	0.74 ± 0.07	0.66 ± 0.07	0.75 ± 0.04	0.4 ± 0.2	0.3 ± 0.2	0.96 ± 0.08	0.33 ± 0.05
endêmicos	LR	mm33	0.79 ± 0.03	0.73 ± 0.03	0.76 ± 0.04	0.46 ± 0.07	0.67 ± 0.05	0.78 ± 0.08	0.45 ± 0.05
endêmicos	LR	mm34	0.74 ± 0.01	0.69 ± 0.03	0.68 ± 0.05	0.39 ± 0.06	0.77 ± 0.02	0.61 ± 0.06	0.51 ± 0.01
endêmicos	LR	MM ₁₁	0.85 ± 0.03	0.78 ± 0.05	0.77 ± 0.05	0.58 ± 0.09	0.88 ± 0.06	0.69 ± 0.07	0.51 ± 0.03
endêmicos	MLP	AnE	0.85 ± 0.04	0.70 ± 0.07	0.7 ± 0.1	0.5 ± 0.1	0.88 ± 0.09	0.6 ± 0.2	0.56 ± 0.05
endêmicos	MLP	Aps	0.84 ± 0.05	0.72 ± 0.05	0.76 ± 0.06	0.5 ± 0.1	0.5 ± 0.2	0.9 ± 0.1	0.44 ± 0.09
endêmicos	MLP	Asp	0.84 ± 0.04	0.73 ± 0.06	0.77 ± 0.05	0.5 ± 0.1	0.6 ± 0.2	0.9 ± 0.1	0.43 ± 0.08
endêmicos	MLP	dPolE	0.80 ± 0.06	0.71 ± 0.07	0.7 ± 0.1	0.4 ± 0.1	0.7 ± 0.1	0.7 ± 0.2	0.6 ± 0.1
endêmicos	MLP	Psi	0.83 ± 0.06	0.72 ± 0.08	0.76 ± 0.06	0.5 ± 0.2	0.6 ± 0.2	0.8 ± 0.1	0.43 ± 0.09
endêmicos	MLP	mm12	0.81 ± 0.04	0.72 ± 0.06	0.75 ± 0.08	0.5 ± 0.1	0.7 ± 0.2	0.8 ± 0.2	0.5 ± 0.1
endêmicos	MLP	mm13	0.78 ± 0.05	0.65 ± 0.08	0.74 ± 0.04	0.3 ± 0.2	0.3 ± 0.2	0.95 ± 0.08	0.3 ± 0.1
endêmicos	MLP	mm14	0.74 ± 0.07	0.65 ± 0.08	0.70 ± 0.07	0.3 ± 0.2	0.5 ± 0.2	0.8 ± 0.1	0.46 ± 0.04
endêmicos	MLP	mm21	0.73 ± 0.05	0.67 ± 0.07	0.7 ± 0.1	0.3 ± 0.2	0.6 ± 0.2	0.7 ± 0.2	0.4 ± 0.1
endêmicos	MLP	mm22	0.82 ± 0.06	0.66 ± 0.06	0.75 ± 0.03	0.4 ± 0.1	0.4 ± 0.2	0.92 ± 0.09	0.3 ± 0.1
endêmicos	MLP	mm23	0.82 ± 0.04	0.72 ± 0.05	0.75 ± 0.06	0.4 ± 0.1	0.6 ± 0.1	0.8 ± 0.1	0.51 ± 0.08
endêmicos	MLP	mm24	0.77 ± 0.06	0.65 ± 0.08	0.71 ± 0.07	0.3 ± 0.2	0.5 ± 0.2	0.79 ± 0.09	0.45 ± 0.04
endêmicos	MLP	mm31	0.79 ± 0.06	0.58 ± 0.07	0.69 ± 0.05	0.1 ± 0.2	0.2 ± 0.2	0.9 ± 0.1	0.3 ± 0.1
endêmicos	MLP	mm32	0.75 ± 0.07	0.62 ± 0.07	0.71 ± 0.04	0.2 ± 0.2	0.3 ± 0.2	0.9 ± 0.1	0.3 ± 0.1
endêmicos	MLP	mm33	0.78 ± 0.04	0.69 ± 0.06	0.72 ± 0.09	0.4 ± 0.1	0.6 ± 0.2	0.8 ± 0.1	0.4 ± 0.1
endêmicos	MLP	mm34	0.78 ± 0.01	0.68 ± 0.03	0.63 ± 0.05	0.41 ± 0.06	0.87 ± 0.06	0.52 ± 0.07	0.58 ± 0.05
endêmicos	MLP	MM ₁₁	0.85 ± 0.05	0.77 ± 0.07	0.7 ± 0.1	0.6 ± 0.1	0.93 ± 0.06	0.6 ± 0.1	0.74 ± 0.06
endêmicos	SVM	AnE	0.72 ± 0.07	0.68 ± 0.06	0.69 ± 0.08	0.4 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.50 ± 0.01
endêmicos	SVM	Aps	0.88 ± 0.03	0.82 ± 0.05	0.84 ± 0.04	0.64 ± 0.09	0.8 ± 0.1	0.85 ± 0.07	0.56 ± 0.04
endêmicos	SVM	Asp	0.82 ± 0.05	0.75 ± 0.06	0.77 ± 0.06	0.5 ± 0.1	0.7 ± 0.1	0.8 ± 0.1	0.53 ± 0.04
endêmicos	SVM	dPolE	0.86 ± 0.06	0.74 ± 0.05	0.79 ± 0.04	0.5 ± 0.1	0.6 ± 0.1	0.90 ± 0.06	0.46 ± 0.03
endêmicos	SVM	Psi	0.76 ± 0.05	0.68 ± 0.05	0.75 ± 0.03	0.4 ± 0.1	0.4 ± 0.1	0.92 ± 0.05	0.48 ± 0.02
endêmicos	SVM	mm12	0.83 ± 0.06	0.76 ± 0.07	0.79 ± 0.07	0.5 ± 0.2	0.7 ± 0.1	0.8 ± 0.1	0.53 ± 0.02
endêmicos	SVM	mm13	0.83 ± 0.04	0.75 ± 0.06	0.78 ± 0.05	0.5 ± 0.1	0.6 ± 0.1	0.87 ± 0.08	0.50 ± 0.04
endêmicos	SVM	mm14	0.83 ± 0.06	0.67 ± 0.05	0.75 ± 0.03	0.4 ± 0.1	0.4 ± 0.1	0.93 ± 0.07	0.45 ± 0.03
endêmicos	SVM	mm21	0.85 ± 0.04	0.79 ± 0.07	0.80 ± 0.07	0.6 ± 0.1	0.8 ± 0.1	0.8 ± 0.1	0.51 ± 0.03
endêmicos	SVM	mm22	0.80 ± 0.05	0.71 ± 0.05	0.72 ± 0.06	0.4 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.51 ± 0.03
endêmicos	SVM	mm23	0.82 ± 0.06	0.71 ± 0.05	0.77 ± 0.03	0.45 ± 0.09	0.4 ± 0.1	0.93 ± 0.06	0.45 ± 0.04
endêmicos	SVM	mm24	0.75 ± 0.04	0.75 ± 0.05	0.79 ± 0.06	0.5 ± 0.1	0.6 ± 0.1	0.9 ± 0.1	0.50 ± 0.04
endêmicos	SVM	mm31	0.78 ± 0.07	0.64 ± 0.05	0.73 ± 0.03	0.3 ± 0.1	0.3 ± 0.2	0.91 ± 0.09	0.47 ± 0.03
endêmicos	SVM	mm32	0.78 ± 0.05	0.70 ± 0.07	0.74 ± 0.06	0.4 ± 0.1	0.6 ± 0.2	0.8 ± 0.1	0.51 ± 0.03
endêmicos	SVM	mm33	0.80 ± 0.05	0.73 ± 0.04	0.77 ± 0.04	0.48 ± 0.09	0.6 ± 0.1	0.9 ± 0.1	0.48 ± 0.05
endêmicos	SVM	mm34	0.80 ± 0.04	0.68 ± 0.06	0.75 ± 0.04	0.4 ± 0.1	0.5 ± 0.1	0.88 ± 0.05	0.41 ± 0.04
endêmicos	SVM	MM ₁₁	0.88 ± 0.04	0.83 ± 0.04	0.82 ± 0.05	0.67 ± 0.08	0.90 ± 0.07	0.76 ± 0.08	0.63 ± 0.06
endêmicos	XGBoost	AnE	0.83 ± 0.05	0.72 ± 0.05	0.77 ± 0.04	0.4 ± 0.1	0.6 ± 0.1	0.86 ± 0.08	0.50 ± 0.06
endêmicos	XGBoost	Aps	0.83 ± 0.06	0.72 ± 0.04	0.78 ± 0.03	0.47 ± 0.08	0.5 ± 0.1	0.92 ± 0.04	0.21 ± 0.03
endêmicos	XGBoost	Asp	0.64 ± 0.09	0.61 ± 0.09	0.6 ± 0.1	0.2 ± 0.2	0.5 ± 0.3	0.7 ± 0.3	0.5 ± 0.1
endêmicos	XGBoost	dPolE	0.74 ± 0.04	0.67 ± 0.06	0.69 ± 0.06	0.3 ± 0.1	0.6 ± 0.1	0.7 ± 0.1	0.57 ± 0.07
endêmicos	XGBoost	Psi	0.80 ± 0.04	0.73 ± 0.05	0.78 ± 0.04	0.5 ± 0.1	0.55 ± 0.08	0.88 ± 0.06	0.24 ± 0.01
endêmicos	XGBoost	mm12	0.75 ± 0.07	0.65 ± 0.08	0.7 ± 0.1	0.3 ± 0.2	0.5 ± 0.2	0.8 ± 0.2	0.5 ± 0.1
endêmicos	XGBoost	mm13	0.84 ± 0.06	0.68 ± 0.06	0.76 ± 0.04	0.4 ± 0.1	0.4 ± 0.1	0.94 ± 0.07	0.25 ± 0.04
endêmicos	XGBoost	mm14	0.68 ± 0.09	0.58 ± 0.08	0.5 ± 0.2	0.2 ± 0.2	0.7 ± 0.2	0.5 ± 0.3	0.52 ± 0.04
endêmicos	XGBoost	mm21	0.73 ± 0.07	0.64 ± 0.05	0.73 ± 0.05	0.3 ± 0.1	0.4 ± 0.1	0.9 ± 0.1	0.42 ± 0.08
endêmicos	XGBoost	mm22	0.79 ± 0.05	0.70 ± 0.06	0.74 ± 0.04	0.4 ± 0.1	0.6 ± 0.2	0.8 ± 0.1	0.55 ± 0.08
endêmicos	XGBoost	mm23	0.86 ± 0.04	0.78 ± 0.07	0.79 ± 0.06	0.6 ± 0.1	0.8 ± 0.2	0.8 ± 0.1	0.51 ± 0.07
endêmicos	XGBoost	mm24	0.69 ± 0.09	0.60 ± 0.07	0.6 ± 0.2	0.2 ± 0.1	0.6 ± 0.3	0.6 ± 0.3	0.5 ± 0.2
endêmicos	XGBoost	mm31	0.76 ± 0.05	0.68 ± 0.05	0.75 ± 0.04	0.4 ± 0.1	0.4 ± 0.1	0.89 ± 0.06	0.39 ± 0.06
endêmicos	XGBoost	mm32	0.70 ± 0.07	0.62 ± 0.06	0.66 ± 0.07	0.2 ± 0.1	0.5 ± 0.2	0.7 ± 0.2	0.52 ± 0.09
endêmicos	XGBoost	mm33	0.70 ± 0.08	0.65 ± 0.07	0.72 ± 0.09	0.3 ± 0.2	0.4 ± 0.2	0.9 ± 0.2	0.5 ± 0.1
endêmicos	XGBoost	mm34	0.73 ± 0.04	0.67 ± 0.04	0.75 ± 0.03	0.4 ± 0.1	0.3 ± 0.1	0.95 ± 0.06	0.39 ± 0.04
endêmicos	XGBoost	MM ₁₁	0.82 ± 0.04	0.72 ± 0.06	0.74 ± 0.07	0.5 ± 0.1	0.6 ± 0.2	0.8 ± 0.2	0.27 ± 0.04

Tabela 11 – Resultados dos modelos negativos

Grupo amostral	Algoritmo	Propriedade	AUC	Acr	F1	MCC	Espc.	Sens.	Threshold
negativos	KNN	AnE	0.92 ± 0.05	0.7 ± 0.1	0.6 ± 0.2	0.5 ± 0.1	0.98 ± 0.06	0.5 ± 0.2	0.50 ± 0.06
negativos	KNN	Aps	1.00 ± 0.01	0.95 ± 0.04	0.95 ± 0.04	0.90 ± 0.07	0.99 ± 0.04	0.92 ± 0.07	0.49 ± 0.06
negativos	KNN	Asp	0.93 ± 0.04	0.8 ± 0.1	0.7 ± 0.2	0.6 ± 0.2	0.94 ± 0.09	0.6 ± 0.2	0.45 ± 0.06
negativos	KNN	dPolE	1.00 ± 0.01	0.95 ± 0.04	0.96 ± 0.03	0.90 ± 0.07	0.88 ± 0.09	1.00 ± 0.01	0.38 ± 0.05
negativos	KNN	Psi	0.90 ± 0.07	0.7 ± 0.1	0.6 ± 0.2	0.4 ± 0.2	0.95 ± 0.08	0.4 ± 0.2	0.61 ± 0.05
negativos	KNN	mm12	1.00 ± 0.01	0.95 ± 0.03	0.96 ± 0.03	0.89 ± 0.07	0.88 ± 0.09	0.99 ± 0.02	0.41 ± 0.06
negativos	KNN	mm13	1.00 ± 0.01	0.97 ± 0.03	0.98 ± 0.02	0.94 ± 0.06	0.95 ± 0.06	0.98 ± 0.03	0.47 ± 0.04
negativos	KNN	mm14	0.96 ± 0.02	0.85 ± 0.07	0.86 ± 0.08	0.7 ± 0.1	0.91 ± 0.06	0.8 ± 0.1	0.47 ± 0.06
negativos	KNN	mm21	0.99 ± 0.01	0.92 ± 0.04	0.94 ± 0.03	0.84 ± 0.07	0.8 ± 0.1	1.0 ± 0.0	0.39 ± 0.05
negativos	KNN	mm22	1.0 ± 0.0	0.99 ± 0.03	0.99 ± 0.02	0.97 ± 0.05	0.97 ± 0.07	1.0 ± 0.0	0.50 ± 0.01
negativos	KNN	mm23	1.00 ± 0.01	0.98 ± 0.03	0.98 ± 0.03	0.96 ± 0.06	0.99 ± 0.03	0.98 ± 0.04	0.49 ± 0.02
negativos	KNN	mm24	0.97 ± 0.03	0.88 ± 0.06	0.90 ± 0.05	0.8 ± 0.1	0.8 ± 0.2	0.95 ± 0.09	0.40 ± 0.06
negativos	KNN	mm31	0.99 ± 0.01	0.96 ± 0.03	0.96 ± 0.02	0.92 ± 0.06	0.97 ± 0.05	0.95 ± 0.04	0.63 ± 0.04
negativos	KNN	mm32	1.00 ± 0.01	0.96 ± 0.04	0.97 ± 0.04	0.92 ± 0.08	0.98 ± 0.05	0.95 ± 0.06	0.64 ± 0.03
negativos	KNN	mm33	1.0 ± 0.0	0.94 ± 0.05	0.95 ± 0.04	0.9 ± 0.1	0.9 ± 0.1	1.0 ± 0.0	0.46 ± 0.05
negativos	KNN	mm34	0.82 ± 0.08	0.75 ± 0.07	0.79 ± 0.06	0.5 ± 0.2	0.7 ± 0.2	0.8 ± 0.1	0.40 ± 0.06
negativos	KNN	MM_{11}	1.00 ± 0.01	0.97 ± 0.04	0.97 ± 0.03	0.94 ± 0.07	0.98 ± 0.04	0.96 ± 0.06	0.66 ± 0.02
negativos	LR	AnE	0.90 ± 0.03	0.78 ± 0.05	0.82 ± 0.05	0.5 ± 0.1	0.69 ± 0.09	0.84 ± 0.09	0.27 ± 0.02
negativos	LR	Aps	1.0 ± 0.0	1.00 ± 0.01	1.00 ± 0.01	1.00 ± 0.02	1.0 ± 0.0	1.00 ± 0.01	0.51 ± 0.05
negativos	LR	Asp	0.98 ± 0.01	0.89 ± 0.05	0.91 ± 0.04	0.8 ± 0.1	0.8 ± 0.1	0.97 ± 0.04	0.41 ± 0.05
negativos	LR	dPolE	1.00 ± 0.01	0.93 ± 0.04	0.95 ± 0.03	0.87 ± 0.08	0.8 ± 0.1	1.0 ± 0.0	0.30 ± 0.04
negativos	LR	Psi	0.91 ± 0.03	0.80 ± 0.04	0.84 ± 0.04	0.6 ± 0.1	0.66 ± 0.07	0.91 ± 0.07	0.27 ± 0.02
negativos	LR	mm12	1.00 ± 0.01	0.99 ± 0.02	0.99 ± 0.02	0.98 ± 0.04	0.98 ± 0.05	1.00 ± 0.01	0.32 ± 0.03
negativos	LR	mm13	1.00 ± 0.01	0.98 ± 0.03	0.98 ± 0.02	0.96 ± 0.05	0.99 ± 0.03	0.97 ± 0.04	0.56 ± 0.05
negativos	LR	mm14	0.99 ± 0.01	0.91 ± 0.05	0.93 ± 0.03	0.82 ± 0.09	0.8 ± 0.1	0.98 ± 0.03	0.41 ± 0.05
negativos	LR	mm21	0.99 ± 0.01	0.89 ± 0.04	0.92 ± 0.03	0.79 ± 0.08	0.7 ± 0.1	0.99 ± 0.02	0.38 ± 0.07
negativos	LR	mm22	1.00 ± 0.01	0.95 ± 0.03	0.96 ± 0.02	0.90 ± 0.06	0.89 ± 0.07	0.99 ± 0.03	0.54 ± 0.05
negativos	LR	mm23	1.0 ± 0.0	1.00 ± 0.01	1.00 ± 0.01	1.00 ± 0.02	1.0 ± 0.0	1.00 ± 0.01	0.32 ± 0.04
negativos	LR	mm24	0.99 ± 0.02	0.91 ± 0.03	0.93 ± 0.02	0.82 ± 0.06	0.77 ± 0.08	1.0 ± 0.0	0.38 ± 0.06
negativos	LR	mm31	0.99 ± 0.01	0.93 ± 0.03	0.94 ± 0.03	0.86 ± 0.07	0.89 ± 0.08	0.95 ± 0.05	0.51 ± 0.06
negativos	LR	mm32	0.98 ± 0.01	0.83 ± 0.05	0.87 ± 0.04	0.7 ± 0.1	0.6 ± 0.1	1.0 ± 0.0	0.37 ± 0.05
negativos	LR	mm33	1.00 ± 0.01	0.97 ± 0.03	0.98 ± 0.03	0.94 ± 0.07	0.94 ± 0.08	0.99 ± 0.02	0.33 ± 0.03
negativos	LR	mm34	0.87 ± 0.06	0.75 ± 0.07	0.78 ± 0.08	0.5 ± 0.1	0.7 ± 0.2	0.8 ± 0.2	0.68 ± 0.05
negativos	LR	MM_{11}	1.0 ± 0.0	1.00 ± 0.01	1.00 ± 0.01	0.99 ± 0.02	1.0 ± 0.0	0.99 ± 0.02	0.27 ± 0.02
negativos	MLP	AnE	0.90 ± 0.05	0.78 ± 0.06	0.84 ± 0.04	0.6 ± 0.1	0.5 ± 0.2	0.95 ± 0.07	0.44 ± 0.03
negativos	MLP	Aps	1.00 ± 0.01	0.97 ± 0.05	0.97 ± 0.05	0.94 ± 0.09	1.0 ± 0.0	0.94 ± 0.08	0.87 ± 0.04
negativos	MLP	Asp	0.92 ± 0.06	0.74 ± 0.08	0.82 ± 0.05	0.5 ± 0.2	0.4 ± 0.2	0.97 ± 0.06	0.38 ± 0.07
negativos	MLP	dPolE	1.0 ± 0.0	0.99 ± 0.03	0.99 ± 0.03	0.97 ± 0.05	1.00 ± 0.02	0.98 ± 0.05	0.64 ± 0.09
negativos	MLP	Psi	0.92 ± 0.04	0.78 ± 0.08	0.84 ± 0.06	0.6 ± 0.2	0.5 ± 0.2	0.94 ± 0.09	0.44 ± 0.05
negativos	MLP	mm12	1.0 ± 0.0	0.95 ± 0.04	0.95 ± 0.04	0.91 ± 0.07	1.0 ± 0.0	0.92 ± 0.07	0.72 ± 0.09
negativos	MLP	mm13	1.00 ± 0.01	0.97 ± 0.03	0.97 ± 0.02	0.93 ± 0.05	0.96 ± 0.05	0.97 ± 0.04	0.50 ± 0.07
negativos	MLP	mm14	0.99 ± 0.01	0.93 ± 0.05	0.94 ± 0.04	0.9 ± 0.1	0.8 ± 0.1	1.00 ± 0.01	0.4 ± 0.1
negativos	MLP	mm21	0.99 ± 0.01	0.91 ± 0.06	0.93 ± 0.04	0.8 ± 0.1	0.8 ± 0.2	0.99 ± 0.03	0.4 ± 0.1
negativos	MLP	mm22	1.0 ± 0.0	1.00 ± 0.01	1.00 ± 0.01	0.99 ± 0.02	1.0 ± 0.0	0.99 ± 0.02	0.59 ± 0.07
negativos	MLP	mm23	1.0 ± 0.0	0.98 ± 0.03	0.98 ± 0.03	0.96 ± 0.06	0.99 ± 0.04	0.97 ± 0.05	0.5 ± 0.1
negativos	MLP	mm24	0.99 ± 0.02	0.93 ± 0.06	0.94 ± 0.05	0.9 ± 0.1	0.8 ± 0.1	0.99 ± 0.03	0.47 ± 0.04
negativos	MLP	mm31	0.97 ± 0.01	0.85 ± 0.03	0.89 ± 0.02	0.71 ± 0.06	0.70 ± 0.09	0.96 ± 0.05	0.3 ± 0.1
negativos	MLP	mm32	0.95 ± 0.03	0.80 ± 0.04	0.86 ± 0.03	0.62 ± 0.08	0.5 ± 0.1	1.00 ± 0.01	0.22 ± 0.09
negativos	MLP	mm33	0.99 ± 0.01	0.96 ± 0.04	0.96 ± 0.04	0.91 ± 0.09	0.94 ± 0.09	0.97 ± 0.04	0.34 ± 0.09
negativos	MLP	mm34	0.82 ± 0.04	0.80 ± 0.04	0.85 ± 0.04	0.6 ± 0.1	0.58 ± 0.04	0.95 ± 0.06	0.45 ± 0.05
negativos	MLP	MM_{11}	1.0 ± 0.0	1.00 ± 0.01	1.00 ± 0.01	1.00 ± 0.02	1.0 ± 0.0	1.00 ± 0.01	0.49 ± 0.04
negativos	SVM	AnE	0.87 ± 0.05	0.71 ± 0.06	0.80 ± 0.04	0.4 ± 0.1	0.4 ± 0.2	0.96 ± 0.06	0.43 ± 0.03
negativos	SVM	Aps	1.0 ± 0.0	0.98 ± 0.02	0.99 ± 0.02	0.97 ± 0.04	1.0 ± 0.0	0.97 ± 0.04	0.59 ± 0.06
negativos	SVM	Asp	0.98 ± 0.02	0.75 ± 0.04	0.83 ± 0.02	0.53 ± 0.08	0.40 ± 0.09	1.0 ± 0.0	0.42 ± 0.04
negativos	SVM	dPolE	1.00 ± 0.01	0.99 ± 0.02	0.99 ± 0.02	0.98 ± 0.04	0.99 ± 0.03	0.99 ± 0.03	0.50 ± 0.04
negativos	SVM	Psi	0.88 ± 0.05	0.75 ± 0.07	0.82 ± 0.04	0.5 ± 0.1	0.4 ± 0.2	0.97 ± 0.04	0.44 ± 0.03
negativos	SVM	mm12	1.0 ± 0.0	0.99 ± 0.02	1.00 ± 0.02	0.99 ± 0.04	0.99 ± 0.05	1.0 ± 0.0	0.46 ± 0.06
negativos	SVM	mm13	1.00 ± 0.01	0.98 ± 0.02	0.98 ± 0.02	0.95 ± 0.05	0.94 ± 0.06	1.00 ± 0.01	0.47 ± 0.06
negativos	SVM	mm14	1.00 ± 0.01	0.87 ± 0.07	0.91 ± 0.04	0.8 ± 0.1	0.7 ± 0.2	1.00 ± 0.01	0.47 ± 0.05
negativos	SVM	mm21	1.00 ± 0.01	0.99 ± 0.02	0.99 ± 0.02	0.98 ± 0.04	1.0 ± 0.0	0.98 ± 0.03	0.46 ± 0.06
negativos	SVM	mm22	0.99 ± 0.01	0.93 ± 0.04	0.94 ± 0.03	0.86 ± 0.08	0.9 ± 0.1	0.98 ± 0.03	0.49 ± 0.04
negativos	SVM	mm23	1.00 ± 0.01	0.97 ± 0.03	0.98 ± 0.03	0.94 ± 0.06	0.93 ± 0.08	1.0 ± 0.0	0.44 ± 0.04
negativos	SVM	mm24	0.99 ± 0.01	0.94 ± 0.05	0.95 ± 0.04	0.9 ± 0.1	0.9 ± 0.1	0.98 ± 0.04	0.56 ± 0.05
negativos	SVM	mm31	1.00 ± 0.01	0.97 ± 0.02	0.97 ± 0.02	0.94 ± 0.05	0.95 ± 0.06	0.98 ± 0.03	0.58 ± 0.05
negativos	SVM	mm32	1.0 ± 0.0	0.98 ± 0.03	0.99 ± 0.02	0.96 ± 0.06	0.97 ± 0.07	0.99 ± 0.02	0.56 ± 0.05
negativos	SVM	mm33	1.00 ± 0.01	0.95 ± 0.02	0.96 ± 0.02	0.91 ± 0.04	1.0 ± 0.0	0.92 ± 0.03	0.62 ± 0.03
negativos	SVM	mm34	0.69 ± 0.09	0.64 ± 0.07	0.75 ± 0.05	0.2 ± 0.2	0.2 ± 0.2	0.93 ± 0.09	0.41 ± 0.04
negativos	SVM	MM_{11}	1.00 ± 0.01	0.99 ± 0.02	0.99 ± 0.01	0.99 ± 0.03	1.0 ± 0.0	0.99 ± 0.03	0.51 ± 0.06
negativos	XGBoost	AnE	0.91 ± 0.04	0.78 ± 0.06	0.83 ± 0.04	0.6 ± 0.1	0.6 ± 0.2	0.90 ± 0.09	0.44 ± 0.05
negativos	XGBoost	Aps	1.00 ± 0.01	0.96 ± 0.03	0.97 ± 0.02	0.92 ± 0.05	0.90 ± 0.06	1.0 ± 0.0	0.43 ± 0.03
negativos	XGBoost	Asp	0.93 ± 0.03	0.82 ± 0.06	0.86 ± 0.04	0.6 ± 0.1	0.7 ± 0.2	0.91 ± 0.07	0.47 ± 0.08
negativos	XGBoost	dPolE	1.00 ± 0.01	0.96 ± 0.03	0.96 ± 0.03	0.92 ± 0.06	0.97 ± 0.05	0.95 ± 0.05	0.58 ± 0.06
negativos	XGBoost	Psi	0.91 ± 0.04	0.80 ± 0.06	0.83 ± 0.05	0.6 ± 0.1	0.8 ± 0.2	0.8 ± 0.1	0.50 ± 0.06
negativos	XGBoost	mm12	0.99 ± 0.01	0.92 ± 0.07	0.94 ± 0.05	0.8 ± 0.1	0.8 ± 0.2	0.99 ± 0.02	0.50 ± 0.07
negativos	XGBoost	mm13	0.99 ± 0.01	0.90 ± 0.03	0.92 ± 0.02	0.80 ± 0.06	0.76 ± 0.07	1.0 ± 0.0	0.51 ± 0.04
negativos	XGBoost	mm14	0.98 ± 0.01	0.91 ± 0.05	0.93 ± 0.04	0.8 ± 0.1	0.9 ± 0.1	0.94 ± 0.04	0.52 ± 0.04
negativos	XGBoost	mm21	1.00 ± 0.01	0.94 ± 0.05	0.96 ± 0.04	0.89 ± 0.09	0.9 ± 0.1	0.99 ± 0.03	0.52 ± 0.06
negativos	XGBoost	mm22	1.00 ± 0.01	0.91 ± 0.02	0.93 ± 0.01	0.82 ± 0.04	0.77 ± 0.05	1.0 ± 0.0	0.18 ± 0.03
negativos	XGBoost	mm23	0.99 ± 0.01	0.92 ± 0.04	0.93 ± 0.03	0.84 ± 0.08	0.8 ± 0.1	0.97 ± 0.05	0.50 ± 0.06
negativos	XGBoost	mm24	0.97 ± 0.03	0.92 ± 0.05	0.93 ± 0.04	0.84 ± 0.09	0.9 ± 0.1	0.96 ± 0.05	0.58 ± 0.06
negativos	XGBoost	mm31	0.97 ± 0.02	0.88 ± 0.03	0.91 ± 0.03	0.77 ± 0.07	0.74 ± 0.09	0.98 ± 0.04	0.41 ± 0.07
negativos	XGBoost	mm32	0.98 ± 0.01	0.90 ± 0.02	0.92 ± 0.02	0.80 ± 0.05	0.76 ± 0.05	0.99 ± 0.02	0.46 ± 0.04
negativos	XGBoost	mm33	0.99 ± 0.01	0.92 ± 0.04	0.93 ± 0.03	0.84 ± 0.07	0.80 ± 0.09	1.0 ± 0.0	0.40 ± 0.06
negativos	XGBoost	mm34	0.68 ± 0.06	0.67 ± 0.06	0.71 ± 0.06	0.3 ± 0.1	0.6 ± 0.1	0.70 ± 0.09	0.56 ± 0.06
negativos	XGBoost	MM_{11}	1.00 ± 0.01	0.98 ± 0.03	0.98 ± 0.03	0.96 ± 0.06	1.00 ± 0.02	0.97 ± 0.05	0.53 ± 0.06

Tabela 12 – Resultados dos modelos com todas as amostras

Grupo amostral	Algoritmo	Propriedade	AUC	Acr	F1	MCC	Espc.	Sens.	Threshold
todas as amostras	KNN	AnE	0.76 ± 0.04	0.66 ± 0.06	0.67 ± 0.07	0.4 ± 0.1	0.81 ± 0.06	0.56 ± 0.08	0.5 ± 0.0
todas as amostras	KNN	Aps	0.65 ± 0.05	0.62 ± 0.06	0.68 ± 0.06	0.2 ± 0.1	0.56 ± 0.08	0.65 ± 0.07	0.34 ± 0.01
todas as amostras	KNN	Asp	0.66 ± 0.05	0.55 ± 0.04	0.53 ± 0.05	0.21 ± 0.08	0.80 ± 0.05	0.40 ± 0.06	0.33 ± 0.01
todas as amostras	KNN	dPolE	0.66 ± 0.04	0.63 ± 0.04	0.72 ± 0.03	0.2 ± 0.1	0.4 ± 0.1	0.78 ± 0.06	0.45 ± 0.01
todas as amostras	KNN	Psi	0.76 ± 0.03	0.70 ± 0.03	0.73 ± 0.03	0.39 ± 0.06	0.72 ± 0.06	0.68 ± 0.05	0.5 ± 0.0
todas as amostras	KNN	mm12	0.77 ± 0.05	0.72 ± 0.03	0.79 ± 0.02	0.39 ± 0.08	0.52 ± 0.09	0.85 ± 0.04	0.44 ± 0.01
todas as amostras	KNN	mm13	0.63 ± 0.04	0.63 ± 0.03	0.71 ± 0.03	0.20 ± 0.07	0.47 ± 0.07	0.72 ± 0.05	0.45 ± 0.01
todas as amostras	KNN	mm14	0.66 ± 0.05	0.47 ± 0.04	0.33 ± 0.08	0.2 ± 0.1	0.91 ± 0.06	0.21 ± 0.06	0.5 ± 0.0
todas as amostras	KNN	mm21	0.70 ± 0.04	0.67 ± 0.03	0.76 ± 0.03	0.26 ± 0.07	0.40 ± 0.08	0.83 ± 0.06	0.43 ± 0.02
todas as amostras	KNN	mm22	0.72 ± 0.04	0.63 ± 0.03	0.76 ± 0.01	0.1 ± 0.1	0.12 ± 0.06	0.95 ± 0.02	0.31 ± 0.01
todas as amostras	KNN	mm23	0.67 ± 0.05	0.56 ± 0.04	0.57 ± 0.05	0.17 ± 0.08	0.70 ± 0.05	0.47 ± 0.06	0.5 ± 0.0
todas as amostras	KNN	mm24	0.67 ± 0.05	0.41 ± 0.03	0.16 ± 0.07	0.0 ± 0.1	0.92 ± 0.04	0.10 ± 0.05	0.33 ± 0.01
todas as amostras	KNN	mm31	0.68 ± 0.05	0.56 ± 0.05	0.50 ± 0.09	0.26 ± 0.09	0.87 ± 0.07	0.4 ± 0.1	0.65 ± 0.03
todas as amostras	KNN	mm32	0.65 ± 0.05	0.61 ± 0.06	0.7 ± 0.1	0.2 ± 0.1	0.5 ± 0.2	0.7 ± 0.2	0.42 ± 0.06
todas as amostras	KNN	mm33	0.66 ± 0.05	0.54 ± 0.06	0.5 ± 0.1	0.22 ± 0.09	0.85 ± 0.06	0.4 ± 0.1	0.64 ± 0.03
todas as amostras	KNN	mm34	0.72 ± 0.05	0.64 ± 0.06	0.7 ± 0.1	0.34 ± 0.08	0.7 ± 0.2	0.6 ± 0.2	0.50 ± 0.07
todas as amostras	KNN	MM_{11}	0.81 ± 0.03	0.68 ± 0.06	0.69 ± 0.08	0.43 ± 0.08	0.84 ± 0.09	0.6 ± 0.1	0.56 ± 0.05
todas as amostras	LR	AnE	0.82 ± 0.03	0.73 ± 0.04	0.79 ± 0.04	0.43 ± 0.08	0.6 ± 0.1	0.81 ± 0.08	0.53 ± 0.04
todas as amostras	LR	Aps	0.74 ± 0.04	0.63 ± 0.01	0.77 ± 0.01	0.04 ± 0.08	0.02 ± 0.03	1.00 ± 0.01	0.11 ± 0.02
todas as amostras	LR	Asp	0.65 ± 0.06	0.63 ± 0.04	0.72 ± 0.04	0.2 ± 0.1	0.4 ± 0.2	0.8 ± 0.1	0.54 ± 0.05
todas as amostras	LR	dPolE	0.78 ± 0.03	0.74 ± 0.03	0.79 ± 0.03	0.44 ± 0.07	0.61 ± 0.09	0.81 ± 0.06	0.59 ± 0.03
todas as amostras	LR	Psi	0.87 ± 0.02	0.78 ± 0.04	0.81 ± 0.04	0.57 ± 0.07	0.84 ± 0.06	0.74 ± 0.07	0.59 ± 0.03
todas as amostras	LR	mm12	0.81 ± 0.03	0.76 ± 0.03	0.81 ± 0.02	0.49 ± 0.07	0.6 ± 0.1	0.84 ± 0.04	0.51 ± 0.04
todas as amostras	LR	mm13	0.73 ± 0.04	0.69 ± 0.04	0.78 ± 0.03	0.3 ± 0.1	0.4 ± 0.1	0.88 ± 0.05	0.31 ± 0.02
todas as amostras	LR	mm14	0.68 ± 0.06	0.66 ± 0.04	0.77 ± 0.03	0.2 ± 0.1	0.2 ± 0.1	0.92 ± 0.06	0.31 ± 0.03
todas as amostras	LR	mm21	0.76 ± 0.03	0.71 ± 0.02	0.75 ± 0.02	0.42 ± 0.05	0.72 ± 0.06	0.71 ± 0.04	0.39 ± 0.01
todas as amostras	LR	mm22	0.80 ± 0.03	0.73 ± 0.04	0.79 ± 0.03	0.43 ± 0.08	0.60 ± 0.08	0.82 ± 0.06	0.40 ± 0.02
todas as amostras	LR	mm23	0.76 ± 0.04	0.69 ± 0.03	0.79 ± 0.02	0.3 ± 0.1	0.28 ± 0.09	0.94 ± 0.04	0.30 ± 0.04
todas as amostras	LR	mm24	0.69 ± 0.05	0.64 ± 0.06	0.68 ± 0.08	0.3 ± 0.1	0.66 ± 0.08	0.6 ± 0.1	0.61 ± 0.04
todas as amostras	LR	mm31	0.65 ± 0.04	0.59 ± 0.04	0.70 ± 0.03	0.07 ± 0.09	0.28 ± 0.08	0.79 ± 0.06	0.46 ± 0.02
todas as amostras	LR	mm32	0.71 ± 0.04	0.66 ± 0.03	0.78 ± 0.02	0.20 ± 0.09	0.15 ± 0.06	0.96 ± 0.02	0.19 ± 0.03
todas as amostras	LR	mm33	0.75 ± 0.03	0.68 ± 0.03	0.76 ± 0.02	0.30 ± 0.09	0.5 ± 0.1	0.82 ± 0.04	0.36 ± 0.03
todas as amostras	LR	mm34	0.72 ± 0.02	0.66 ± 0.03	0.68 ± 0.04	0.35 ± 0.05	0.76 ± 0.08	0.59 ± 0.07	0.47 ± 0.02
todas as amostras	LR	MM_{11}	0.85 ± 0.03	0.78 ± 0.03	0.82 ± 0.02	0.52 ± 0.07	0.67 ± 0.09	0.84 ± 0.05	0.41 ± 0.03
todas as amostras	MLP	AnE	0.78 ± 0.03	0.69 ± 0.04	0.74 ± 0.04	0.38 ± 0.08	0.7 ± 0.1	0.70 ± 0.07	0.50 ± 0.02
todas as amostras	MLP	Aps	0.72 ± 0.04	0.62 ± 0.01	0.77 ± 0.01	0.02 ± 0.06	0.01 ± 0.02	1.0 ± 0.0	0.11 ± 0.06
todas as amostras	MLP	Asp	0.68 ± 0.05	0.62 ± 0.01	0.77 ± 0.01	0.03 ± 0.08	0.01 ± 0.02	1.00 ± 0.01	0.17 ± 0.05
todas as amostras	MLP	dPolE	0.75 ± 0.04	0.73 ± 0.03	0.80 ± 0.03	0.41 ± 0.08	0.5 ± 0.1	0.90 ± 0.07	0.45 ± 0.04
todas as amostras	MLP	Psi	0.80 ± 0.04	0.72 ± 0.04	0.80 ± 0.03	0.4 ± 0.1	0.4 ± 0.1	0.90 ± 0.06	0.42 ± 0.06
todas as amostras	MLP	mm12	0.81 ± 0.03	0.73 ± 0.04	0.80 ± 0.02	0.4 ± 0.1	0.5 ± 0.1	0.89 ± 0.04	0.46 ± 0.06
todas as amostras	MLP	mm13	0.67 ± 0.04	0.63 ± 0.02	0.77 ± 0.01	0.1 ± 0.1	0.07 ± 0.05	0.97 ± 0.02	0.16 ± 0.06
todas as amostras	MLP	mm14	0.73 ± 0.04	0.67 ± 0.05	0.77 ± 0.03	0.3 ± 0.1	0.4 ± 0.1	0.86 ± 0.04	0.46 ± 0.04
todas as amostras	MLP	mm21	0.71 ± 0.05	0.68 ± 0.05	0.73 ± 0.05	0.33 ± 0.09	0.6 ± 0.1	0.7 ± 0.1	0.52 ± 0.05
todas as amostras	MLP	mm22	0.75 ± 0.04	0.66 ± 0.05	0.77 ± 0.03	0.2 ± 0.2	0.2 ± 0.2	0.92 ± 0.06	0.3 ± 0.1
todas as amostras	MLP	mm23	0.74 ± 0.05	0.63 ± 0.01	0.77 ± 0.01	0.1 ± 0.1	0.04 ± 0.03	0.99 ± 0.02	0.15 ± 0.05
todas as amostras	MLP	mm24	0.70 ± 0.04	0.65 ± 0.03	0.76 ± 0.02	0.21 ± 0.09	0.32 ± 0.08	0.86 ± 0.04	0.41 ± 0.03
todas as amostras	MLP	mm31	0.64 ± 0.05	0.62 ± 0.02	0.76 ± 0.01	0.0 ± 0.1	0.06 ± 0.08	0.96 ± 0.02	0.23 ± 0.08
todas as amostras	MLP	mm32	0.67 ± 0.04	0.63 ± 0.01	0.77 ± 0.01	0.05 ± 0.07	0.01 ± 0.02	1.0 ± 0.0	0.09 ± 0.03
todas as amostras	MLP	mm33	0.75 ± 0.04	0.70 ± 0.04	0.75 ± 0.04	0.37 ± 0.09	0.6 ± 0.1	0.73 ± 0.08	0.52 ± 0.07
todas as amostras	MLP	mm34	0.76 ± 0.03	0.67 ± 0.04	0.77 ± 0.03	0.3 ± 0.1	0.3 ± 0.1	0.88 ± 0.06	0.40 ± 0.04
todas as amostras	MLP	MM_{11}	0.80 ± 0.04	0.73 ± 0.05	0.77 ± 0.04	0.47 ± 0.09	0.77 ± 0.08	0.71 ± 0.07	0.55 ± 0.02
todas as amostras	SVM	AnE	0.77 ± 0.04	0.70 ± 0.03	0.79 ± 0.02	0.34 ± 0.08	0.43 ± 0.09	0.87 ± 0.04	0.48 ± 0.01
todas as amostras	SVM	Aps	0.79 ± 0.03	0.64 ± 0.01	0.77 ± 0.01	0.16 ± 0.09	0.06 ± 0.04	1.00 ± 0.01	0.33 ± 0.02
todas as amostras	SVM	Asp	0.66 ± 0.01	0.63 ± 0.01	0.77 ± 0.01	0.09 ± 0.09	0.02 ± 0.03	1.0 ± 0.0	0.41 ± 0.01
todas as amostras	SVM	dPolE	0.84 ± 0.03	0.74 ± 0.04	0.82 ± 0.02	0.4 ± 0.1	0.42 ± 0.09	0.93 ± 0.03	0.46 ± 0.01
todas as amostras	SVM	Psi	0.77 ± 0.02	0.69 ± 0.03	0.75 ± 0.02	0.35 ± 0.06	0.59 ± 0.08	0.76 ± 0.04	0.50 ± 0.01
todas as amostras	SVM	mm12	0.84 ± 0.03	0.79 ± 0.03	0.84 ± 0.02	0.53 ± 0.08	0.61 ± 0.09	0.89 ± 0.03	0.49 ± 0.02
todas as amostras	SVM	mm13	0.76 ± 0.02	0.66 ± 0.01	0.79 ± 0.01	0.26 ± 0.03	0.11 ± 0.02	1.0 ± 0.0	0.35 ± 0.02
todas as amostras	SVM	mm14	0.77 ± 0.04	0.61 ± 0.01	0.76 ± 0.01	-0.03 ± 0.09	0.01 ± 0.02	0.98 ± 0.01	0.38 ± 0.01
todas as amostras	SVM	mm21	0.81 ± 0.03	0.75 ± 0.04	0.82 ± 0.03	0.46 ± 0.09	0.5 ± 0.1	0.88 ± 0.04	0.49 ± 0.02
todas as amostras	SVM	mm22	0.87 ± 0.03	0.74 ± 0.03	0.82 ± 0.02	0.45 ± 0.08	0.36 ± 0.09	0.97 ± 0.02	0.36 ± 0.03
todas as amostras	SVM	mm23	0.79 ± 0.03	0.64 ± 0.02	0.77 ± 0.01	0.14 ± 0.08	0.07 ± 0.04	0.98 ± 0.01	0.33 ± 0.03
todas as amostras	SVM	mm24	0.72 ± 0.01	0.67 ± 0.01	0.79 ± 0.01	0.30 ± 0.03	0.15 ± 0.04	1.00 ± 0.01	0.43 ± 0.03
todas as amostras	SVM	mm31	0.71 ± 0.03	0.66 ± 0.02	0.78 ± 0.01	0.22 ± 0.08	0.12 ± 0.04	0.98 ± 0.01	0.40 ± 0.01
todas as amostras	SVM	mm32	0.75 ± 0.02	0.64 ± 0.02	0.78 ± 0.01	0.2 ± 0.1	0.07 ± 0.05	1.00 ± 0.01	0.37 ± 0.02
todas as amostras	SVM	mm33	0.78 ± 0.03	0.74 ± 0.04	0.80 ± 0.03	0.43 ± 0.09	0.57 ± 0.09	0.84 ± 0.04	0.44 ± 0.02
todas as amostras	SVM	mm34	0.74 ± 0.03	0.64 ± 0.02	0.77 ± 0.01	0.17 ± 0.07	0.09 ± 0.05	0.98 ± 0.01	0.43 ± 0.01
todas as amostras	SVM	MM_{11}	0.95 ± 0.02	0.89 ± 0.03	0.92 ± 0.02	0.77 ± 0.05	0.70 ± 0.07	1.00 ± 0.01	0.40 ± 0.03
todas as amostras	XGBoost	AnE	0.75 ± 0.03	0.68 ± 0.03	0.79 ± 0.02	0.29 ± 0.08	0.24 ± 0.08	0.95 ± 0.05	0.38 ± 0.04
todas as amostras	XGBoost	Aps	0.70 ± 0.03	0.64 ± 0.02	0.77 ± 0.01	0.1 ± 0.1	0.06 ± 0.04	0.99 ± 0.02	0.17 ± 0.01
todas as amostras	XGBoost	Asp	0.64 ± 0.04	0.62 ± 0.01	0.77 ± 0.01	0.01 ± 0.03	0.00 ± 0.01	1.0 ± 0.0	0.19 ± 0.03
todas as amostras	XGBoost	dPolE	0.82 ± 0.03	0.73 ± 0.03	0.80 ± 0.03	0.43 ± 0.08	0.5 ± 0.1	0.88 ± 0.09	0.49 ± 0.08
todas as amostras	XGBoost	Psi	0.75 ± 0.03	0.66 ± 0.03	0.77 ± 0.02	0.23 ± 0.08	0.24 ± 0.09	0.92 ± 0.05	0.41 ± 0.05
todas as amostras	XGBoost	mm12	0.78 ± 0.05	0.72 ± 0.04	0.79 ± 0.03	0.4 ± 0.1	0.5 ± 0.2	0.86 ± 0.07	0.50 ± 0.06
todas as amostras	XGBoost	mm13	0.72 ± 0.04	0.69 ± 0.03	0.77 ± 0.03	0.30 ± 0.08	0.4 ± 0.1	0.86 ± 0.08	0.44 ± 0.06
todas as amostras	XGBoost	mm14	0.64 ± 0.05	0.6 ± 0.0	0.77 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	1.0 ± 0.0	0.15 ± 0.04
todas as amostras	XGBoost	mm21	0.73 ± 0.04	0.68 ± 0.04	0.78 ± 0.02	0.3 ± 0.1	0.3 ± 0.1	0.90 ± 0.03	0.42 ± 0.05
todas as amostras	XGBoost	mm22	0.76 ± 0.05	0.72 ± 0.04	0.80 ± 0.02	0.4 ± 0.1	0.5 ± 0.1	0.89 ± 0.05	0.47 ± 0.06
todas as amostras	XGBoost	mm23	0.72 ± 0.05	0.64 ± 0.04	0.76 ± 0.02	0.2 ± 0.1	0.19 ± 0.09	0.92 ± 0.03	0.42 ± 0.05
todas as amostras	XGBoost	mm24	0.71 ± 0.05	0.62 ± 0.01	0.77 ± 0.01	0.01 ± 0.04	0.00 ± 0.01	1.0 ± 0.0	0.11 ± 0.07
todas as amostras	XGBoost	mm31	0.62 ± 0.05	0.63 ± 0.03	0.76 ± 0.03	0.1 ± 0.1	0.2 ± 0.1	0.92 ± 0.06	0.33 ± 0.06
todas as amostras	XGBoost	mm32	0.68 ± 0.05	0.65 ± 0.04	0.75 ± 0.04	0.2 ± 0.1	0.3 ± 0.2	0.87 ± 0.09	0.43 ± 0.08
todas as amostras	XGBoost	mm33	0.71 ± 0.04	0.67 ± 0.04	0.75 ± 0.04	0.3 ± 0.1	0.5 ± 0.1	0.80 ± 0.09	0.53 ± 0.06
todas as amostras	XGBoost	mm34	0.72 ± 0.07	0.69 ± 0.05	0.79 ± 0.04	0.3 ± 0.1	0.3 ± 0.1	0.91 ± 0.08	0.49 ± 0.06
todas as amostras	XGBoost	MM_{11}	0.78 ± 0.02	0.71 ± 0.03	0.78 ± 0.02	0.37 ± 0.06	0.49 ± 0.08	0.84 ± 0.05	0.51 ± 0.06