



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UMA PROPOSTA METODOLÓGICA DE APRENDIZADO DE MÁQUINA PARA
DETECÇÃO DE CERATOCONE

Francis Spiegel Rubin

Orientadores

D.Sc. Adriana Cesário de Faria Alvim

Ph.D. Carlos Eduardo Ribeiro de Mello

RIO DE JANEIRO, RJ - BRASIL
ABRIL DE 2022

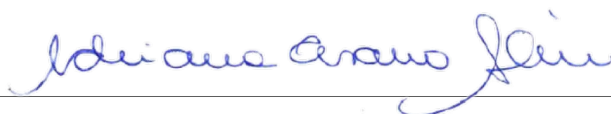
UMA PROPOSTA METODOLÓGICA DE APRENDIZADO DE MÁQUINA PARA
DETECÇÃO DE CERATOCONE

Francis Spiegel Rubin

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO
DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFOR-
MÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNI-
RIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Em conformidade com a Resolução nº 5.257 de 25/03/2020 e a Ordem de Serviço
PROPGPI nº 3 de 2/07/2020, esta ata vai somente por mim assinada, atestando que a
defesa ocorreu com a participação dos componentes abaixo listados.

Aprovada por:



Adriana Cesário de Faria Alvim, D.Sc. — UNIRIO

Carlos Eduardo Ribeiro de Mello, Ph.D. — UNIRIO

Eduardo Fonseca Mendes, Ph.D. — FGV-RIO

Pedro Nuno de Souza Moura, D.Sc. — UNIRIO

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 2022.

Catálogo informatizada pelo(a) autor(a)

- | | |
|------|---|
| S896 | <p>Spiegel Rubin, Francis</p> <p>Uma Proposta Metodológica de Aprendizado de Máquina para Detecção de Ceratocone / Francis Spiegel Rubin. -- Rio de Janeiro, 2022.</p> <p>95 f</p> <p>Orientadora: Adriana Cesário de Faria Alvim.
Coorientador: Carlos Eduardo Ribeiro de Mello.
Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2022.</p> <p>1. Aprendizado de Máquina. 2. Ceratocone. 3. Bioinformática. I. Cesário de Faria Alvim, Adriana, orient. II. Ribeiro de Mello, Carlos Eduardo, coorient. III. Título.</p> |
|------|---|

Agradecimentos

Ao meu pai que já se foi, mas que nunca se foi, à minha mãe, irmãos, esposo e filhos pelo apoio, carinho e amor incondicional.

Aos professores Adriana Cesário de Faria Alvim e Carlos Eduardo Ribeiro de Mello, pela orientação, dedicação, paciência e pela oportunidade concedida. Aos professores Rodrigo Pereira dos Santos e María del Rosario Girardi Gutiérrez pelos ensinamentos e contribuição na minha formação acadêmica.

À Universidade Federal do Estado do Rio de Janeiro e à coordenação do Programa de Pós-Graduação em Informática PPGI-UNIRIO, por ter me acolhido como aluna do curso de mestrado.

Aos amigos e todas as pessoas que acreditaram e me apoiaram, mesmo que indiretamente.

RUBIN, FRANCIS SPIEGEL **Uma Proposta Metodológica de Aprendizado de Máquina para Detecção de Ceratocone**. UNIRIO, 2022. 95 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Ceratocone é um problema que provoca alterações no formato e espessura da córnea, deixando-a mais fina e menos resistente. Pacientes que foram submetidos à cirurgia refrativa LASIK tem uma chance maior de desenvolver a enfermidade pelo afinamento corneano, podendo causar cegueira e exigir o transplante de córnea. Esta pesquisa apresenta uma proposta metodológica para detectar estágios iniciais de ceratocone em pacientes candidatos à cirurgia refrativa, aplicando algoritmos de aprendizado de máquina em dados oftalmológicos da córnea. A contribuição da pesquisa tem como diferencial a geração de modelos robustos de aprendizado de máquina supervisionado, capazes de auxiliar no diagnóstico médico de pacientes candidatos à cirurgia LASIK com risco de desenvolvimento do ceratocone. A metodologia proposta é baseada no cálculo dos intervalos de confiança das medidas de interesse, aplicando as técnicas de *Bootstrap* e *k-fold* para geração das distribuições. Para validação da proposta, é realizado um experimento computacional aplicando os algoritmos de Floresta Aleatória e XGBoost sobre uma amostra de dados disponibilizada por uma clínica de olhos.

Palavras-chave: Aprendizado de Máquina; Ceratocone; Bioinformática

ABSTRACT

Keratoconus is a problem that causes changes in the shape and thickness of the cornea, making it thinner and less resistant. Patients who have undergone refractive surgery LASIK have a greater chance of developing a corneal thinning disease, which can cause blindness and require corneal transplantation. This research presents a methodological proposal to detect the early risk of keratoconus in patients who are candidates for refractive surgery, applying machine learning algorithms to ophthalmic corneal data. The contribution of the research has the differential of generating robust and reliable supervised machine learning models, capable of assisting in the medical diagnosis of patients candidates for LASIK surgery at risk of developing keratoconus. The proposed methodology is based on the calculation of confidence intervals of the measures of interest, applying *Bootstrap* and *k-fold* techniques to generate the distributions. To validate the proposal, a computational experiment is carried out applying the Random Forest and XG-Boost algorithms on a data sample provided by an eye clinic.

Keywords: Machine Learning; Keratoconus; Bioinformatics

Sumário

1	Introdução	1
1.1	Contexto e Motivação	1
1.2	Caracterização do Problema	4
1.3	Objetivos	5
1.4	Justificativa	5
1.5	Organização do Trabalho	5
2	Caracterização do Problema	6
2.1	Anatomia da Córnea	6
2.2	Impacto da Cirurgia Refrativa - LASIK	9
2.3	Cálculo da Espessura Residual do Estroma	12
2.4	O Problema do Ceratocone	14
2.5	Diagnóstico de Ceratocone	15
2.6	O Problema Computacional	16
2.6.1	Conjunto de Dados	16
2.6.2	Análise Exploratória dos Dados	18
2.6.3	Correlação entre Variáveis	20
2.7	Resumo	21

3	Revisão Bibliográfica	22
3.1	Fundamentação Teórica	22
3.1.1	Aprendizado de Máquina	22
3.1.2	Matriz de Confusão	23
3.1.3	Valor de Corte	26
3.1.4	Viés e Variância	27
3.1.5	A Perda Logarítmica	27
3.1.6	Árvore de Decisão	28
3.1.7	Entropia, Ganho de Informação e Índice Gini	29
3.1.8	O Algoritmo de Floresta Aleatória	31
3.1.9	O Algoritmo XGBoost	33
3.1.10	Busca em Grade para a Seleção de Hiperparâmetros	37
3.1.11	Validação Cruzada (em inglês, <i>Cross-Validatiton</i> , CV)	38
3.1.11.1	Validação Cruzada <i>k-fold</i>	39
3.1.11.2	Validação Cruzada Aninhada	40
3.1.12	Interpretação Local dos Dados	40
3.1.13	Cálculo do Intervalo de Confiança	41
3.1.14	<i>Bootstrap</i> para o Cálculo da Distribuição dos Estimadores	42
3.2	Mapeamento de Trabalhos Relacionados	43
3.2.1	Trabalho Comentado - Yoo et al. [71]	45
3.2.2	Trabalho Comentado - Lavric et al. [39]	45
3.2.3	Trabalho Comentado - Leão et al. [40]	46
3.3	Resumo	46
4	Proposta: Uma Nova Metodologia Experimental	48

4.1	Motivação Geral	49
4.2	Hipótese da Proposta	51
4.3	Linhas Gerais da Proposta	52
4.3.1	O Critério Restritivo do Intervalo de Confiança	56
4.3.2	Estratégia para Sistematizar a Geração das Estimativas	57
4.4	Resumo	61
5	O Experimento Computacional	63
5.1	Geração do Ranking de Modelos	63
5.2	Fluxo de Treinamento do Classificador	63
5.2.1	Tratamento e Limpeza de Dados	64
5.2.2	Divisão em Dados de Treino e Teste	65
5.2.3	Modelo Experimental de Floresta Aleatória	65
5.2.4	Modelo Experimental de XGBoost	66
5.2.5	Definição do Valor de Corte (em inglês, <i>threshold</i>)	66
5.2.6	Geração das Medidas de Interesse	67
5.2.7	Avaliação do Modelo	67
5.3	Resumo	67
6	Resultados e Discussão	68
6.1	Modelo por Cobertura - Floresta Aleatória	68
6.2	Modelos por <i>Negative Log Loss</i>	71
6.2.1	Histograma das Distribuições	71
6.2.2	<i>Bootstrap</i> : Matriz de Confusão Consolidada por <i>Negative Log Loss</i>	74
6.2.3	<i>Bootstrap</i> : Estatísticas por <i>Negative Log Loss</i>	79

6.2.4	Diagrama de Caixa das Distribuições	82
6.3	Discussão dos Resultados	84
6.4	Resumo	87
7	Considerações Finais	88
	Referências Bibliográficas	89

Lista de Figuras

1.1	Ilustração do Disco de Plácido	2
1.2	Ilustração de córnea normal	3
1.3	Ilustração de córnea assimétrica	3
1.4	Ilustração da evolução do ceratocone e classificação de Krumeich	4
2.1	Anatomia de um olho normal, figura do lado esquerdo (<i>Normal cornea</i>), e um olho com ceratocone, figura do lado direito (<i>Keratoconus</i>)	8
2.2	Esquema do globo ocular. Onde se lê <i>FlapThickness</i> , leia-se Espessura do retalho (<i>flap</i>); onde se lê <i>AblationDepth</i> , leia-se Profundidade da ablação; onde se lê <i>RSBCentral</i> , leia-se Leito estromal residual central; onde se lê <i>RSBMin</i> , leia-se Leito estromal residual no ponto mais fino; onde se lê <i>Pa- chyApex</i> , leia-se Espessura da córnea no ápice; onde se lê <i>ThinnestPachy</i> , leia-se Espessura da córnea no ponto mais fino; onde se lê <i>PTACentral</i> , leia-se % de ablação no ponto central; e onde se lê <i>PTAMin</i> , leia-se % de ablação no ponto mais fino.	8
2.3	Medidas da córnea - seção transversal	9
2.4	Camadas da córnea. Onde se lê <i>Epithelium</i> , leia-se Epitélio; onde se lê <i>Bowman's layer</i> , leia-se camada de Bowman; onde se lê <i>Stroma</i> , leia-se Estroma; onde se lê <i>Descement's membrane</i> , leia-se membrana de Desce- ment; e onde se lê <i>Endothelium</i> , leia-se Endotélio.	10
2.5	Esquema das camadas na cirurgia LASIK	11
2.6	Medidas consideradas na cirurgia LASIK.	14

2.7	Mapas refrativos capturados pelo Oculus Pentacam	16
3.1	Workflow da geração de um modelo de aprendizado supervisionado para o problema de classificação	23
3.2	Matriz de confusão	24
3.3	Ilustração de uma árvore de decisão	28
3.4	Funcionamento da Floresta Aleatória	32
3.5	Funcionamento da validação cruzada aninhada. À esquerda da linha vertical, é exibido o esquema de CV tradicional: no loop interno, o treinamento dos $(k - 1)$ - <i>folds</i> (fundo branco) e o teste sendo aplicado em cada <i>fold</i> (fundo cinza escuro); no loop externo, a validação é aplicada sobre a amostra (fundo azul) separada do treino. Ao final, a estimativa de erros é calculada com base na média de erros em r repetições, $k - 1$ loops internos e k loops externos.	40
4.1	Simulação para geração de amostras de uma população. O gráfico superior representa um cenário com intervalos de confiança de 95%. O gráfico inferior representa um cenário com intervalos de confiança de 80% . . .	50
4.2	Simulação de 1000 execuções para geração de intervalos de confiança de 95% sobre uma amostra gerada por distribuição normal de tamanho $N=100$ observações, média = 40 e desvio padrão = 2. IC com sucesso de cobertura (em preto). IC com falha de cobertura (em vermelho).	51
4.3	Distribuições hipotéticas da medida de interesse para amostras de treino e teste com distintos intervalos de confiança calculados: IC 60% e IC 95%	53
4.4	Esboço da estratégia metodológica	55
4.5	Distribuições sobrepostas (gráfico à esquerda) e distribuições não sobrepostas (gráfico à direita)	60
4.6	Design da proposta metodológica	62
5.1	Fluxo para o treinamento dos modelos candidatos	64

6.1	Diagramas de caixa (gráficos na metade superior) e histogramas (gráficos na metade inferior) com os valores de cobertura sobre os dados de teste (em azul) e validação (em vermelho) - Floresta Aleatória 100 rodadas - <i>Bootstrap</i> com validação cruzada <i>k-fold</i> igual a 10 - Modelos candidatos retornados na posição 1 (à direita) e na posição 96 (à esquerda) do ranking de modelos ajustados pela medida de cobertura	70
6.2	Modelos candidatos de RF por posição no <i>ranking</i> (eixo x) em duas escalas: à esquerda, % de amostras (em azul) que atenderam ao critério restritivo do IC de 95%; à direita, desempenho da medida de cobertura (em preto). Valores observadas dos 100 modelos candidatos com melhor posição no <i>ranking</i> por cobertura, aplicando-se <i>k-fold</i> sobre a validação e <i>Bootstrap</i> , em 100 execuções sobre o teste por modelo	72
6.3	Histogramas com os valores de <i>negative log loss</i> sobre o teste (barras em azul) e validação (barras em vermelho) - Floresta Aleatória 100 rodadas <i>Bootstrap</i> sobre o teste - Modelo ajustado por <i>negative log loss</i>	75
6.4	Histogramas com os valores de <i>negative log loss</i> sobre os dados de teste (barras em azul) e validação (barras em vermelho) - Floresta Aleatória 100 rodadas <i>Bootstrap</i> sobre o teste - Modelo IC 0,95 por <i>negative log loss</i>	76
6.5	Histogramas com os valores de <i>negative log loss</i> sobre os dados de teste (barras em azul) e validação (barras em vermelho) - XGBoost 100 rodadas <i>Bootstrap</i> sobre o teste - Modelo ajustado por <i>negative log loss</i>	77
6.6	Histogramas com os valores de <i>negative log loss</i> sobre os dados de teste (gráfico superior) e validação (gráfico inferior) - XGBoost 100 rodadas <i>Bootstrap</i> sobre o teste - Modelo IC 0,95 por <i>negative log loss</i>	78
6.7	Matriz de confusão consolidada - Floresta Aleatória 100 rodadas - <i>Bootstrap</i> - Modelo ajustado (à esquerda) vs. Modelo IC 0,95 (à direita)	79
6.8	Matriz de confusão consolidada - XGBoost 100 rodadas - <i>Bootstrap</i> - Modelo ajustado (à esquerda) vs. Modelo IC 0,95 (à direita)	80
6.9	<i>Boxplot</i> com os valores de <i>negative log loss</i> sobre os dados de teste (barras em azul) e validação (barras em vermelho) - <i>Bootstrap</i> de 100 rodadas - Floresta Aleatória (RF) e XGBoost - Modelo IC 0,95 e Modelo ajustado	83

Lista de Tabelas

2.1	Relação de termos técnicos	7
2.2	Medidas da profundidade da ablação (em micrômetros)	11
2.3	Classificação de Krumeich	14
3.1	Exemplo de um conjunto de dados S para treinamento de uma árvore de decisão	30
3.2	Níveis de confiança mais comuns para o cálculo do intervalo de confiança	42
3.3	Amostra de trabalhos para diagnóstico de ceratocone	43
4.1	Simulação para cálculo dos intervalos de confiança	57
6.1	Estatísticas para as distribuições de validação e teste - Modelos por <i>nega-</i> <i>tive log loss</i>	81

Lista de Nomenclaturas

AUC	<i>Area Under the Curve</i>
AUCPR	<i>Area Under the Curve Precision-Recall</i>
bIOP	<i>biomechanically-corrected intraocular pressure</i>
Corvis ST	<i>Corneal Visualization Scheimpflug Technology</i>
CV	<i>Cross Validation</i>
DCR	Resposta da deformação da córnea
FN	Falso Negativo
FP	Falso Positivo
HCTP	Perfil da espessura corneana horizontal
IC	Intervalo de Confiança
IOP	<i>Intraocular Pressure</i>
KDD	<i>Knowledge Discovery on Databases</i>
LASIK	<i>Laser-Assisted in Situ Keratomileusis</i>
LIME	<i>Local Interpretable Model-Agnostic Explanations</i>
LOOCV	<i>Leave-One-Out Cross-Validation</i>
ML	<i>Machine Learning</i>
MSE	<i>Mean Squared Error</i>
NPV	<i>Negative Predicted Value</i>
PCA	<i>Principal Component Analysis</i>
PIO	Pressão Intraocular
PPV	<i>Positive Predicted Value</i>
R^2	<i>R Squared</i>
RF	<i>Random Forest</i>
RMSE	<i>Root Mean Squared Error</i>
ROC	<i>Receiver Operating Characteristic</i>
tSNE	<i>Distributed Stochastic Neighbor Embedding</i>
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

1. Introdução

1.1 Contexto e Motivação

A cirurgia LASIK (*Ceratômileose a laser in situ*) é um procedimento de cirurgia refrativa indicado para a correção da miopia, astigmatismo e hipermetropia de comprovada segurança e eficácia [5]. Segundo dados da empresa Market Scope¹, cerca de 0,8 cirurgias LASIK são realizadas para cada grupo de mil brasileiros. Esta relação é de 2,4 nos EUA; 1,7 na Alemanha; 1,2 na França; 0,7 na China e 0,5 na América Latina. Como qualquer procedimento cirúrgico, pode envolver complicações. Uma complicação rara, mas bastante séria que acomete os pacientes no pós-operatório, é a ectasia pós-LASIK, conhecida como ceratocone (em inglês, *keratoconus*) [52].

Ceratocone (do grego *kerato- horn*, córnea e *konos*, cone) é uma enfermidade na córnea do olho humano, com incidência de, aproximadamente, entre 4 a 600 pessoas a cada 100.000 (0.004% a 0.6%) na população mundial [36]. Caracteriza-se pelo afinamento progressivo e assimétrico da córnea, com formato cônico e irregular. Nos Estados Unidos, cerca de 54,5 em 100.000 pessoas desenvolvem a doença, segundo estudo realizado por Kennedy et al. [31]. De acordo com o CBO², o ceratocone é a principal causa de transplantes de córnea feitos no Brasil, sendo esta uma das características mais importantes a se considerar na avaliação pré-operatória de pacientes candidatos à cirurgia refrativa [5].

A ectasia ocular é o nome dado para a presença de alterações na curvatura da córnea. Na oftalmologia, exames como ceratometria, paquimetria, tomografia e topografia permitem avaliar a córnea, bem como diagnosticar ectasias, como o ceratocone [15]. O exame de paquimetria ocular é o processo que nos permite medir a espessura da córnea. Esta é medida em micrômetros (microns, micra ou μm), em que um micron equivale a 1 milionésimo de metro (1×10^{-6} m). Entre os principais métodos conhecidos na literatura

¹<https://www.market-scope.com/pages/reports/lasik>

²Conselho Brasileiro de Oftalmologia

para tratar o problema de detecção de ceratocone, é possível citar: análise de índices/indicadores topográficos, tomográficos e índices biomecânicos [52,59].

O exame de topografia da córnea tem grande utilidade clínica para o diagnóstico do ceratocone e definição do estágio evolutivo da doença, bem como para a indicação de tratamento do ceratocone. Pode-se citar, como exemplo, o exame de topografia de córnea com Disco de Plácido, que avalia a curvatura da face anterior da córnea, pelo reflexo da imagem nos discos circulares do exame [4]. Já a tomografia é uma técnica que permite a reconstrução da estrutura analisada em uma imagem tridimensional [5] e pode ser realizada por meio dos aparelhos Galilei G4 ³, Orbscan II TM (Baush + Lomb) ⁴, Pentacam ⁵ e Sirius ⁶. Neste contexto, o Galilei é um aparelho de diagnóstico que combina ambas as funções de tomografia e topografia por meio de câmeras de Scheimpflug duplas associadas ao Disco de Plácido [5], fornecendo dados como a espessura e curvatura da córnea, tamanho da pupila, profundidade da câmara anterior ⁷, além de gerar mapas tridimensionais da córnea.

No exame de topografia, a curvatura de uma córnea normal é regular e simétrica, sem distorções observadas. Já a córnea com ceratocone apresenta distorções no reflexo dos anéis de Plácido. Este exame é, portanto, fundamental para a avaliação de córneas elegíveis à cirurgia refrativa e diagnóstico de ceratocone. A Figura 1.1 ilustra o instrumento Disco de Plácido para realizar o exame de topografia.

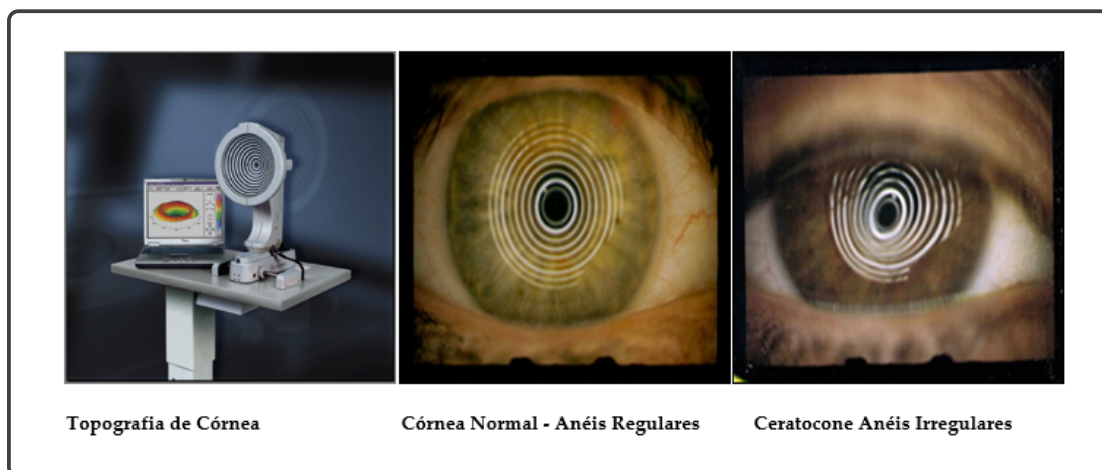


Figura 1.1: Ilustração do Disco de Plácido

Fonte: <https://www.draandrea.com.br/wp-content/uploads/2017/09/topoc1.png> (2021)

O resultado do exame da topografia corneana é conhecido de imediato por meio das

³<https://www.ziemergroup.com/en/products/galilei/>

⁴<https://www.drholzman.com/technologies/orbscan-ii/>

⁵<https://www.pentacam.com/int/technology/measurement-principle-licences-network.html>

⁶<https://www.csoitalia.it/en/prodotto/elenco#filter=.13>

⁷A câmara anterior (frontal) é a região do globo ocular que se estende da córnea até a íris.

imagens construídas pelos meios computacionais associados ao exame. A córnea normal apresenta características simétricas na sua curvatura. Já as córneas levemente assimétricas apresentam um aumento discreto da curvatura que pode causar alterações na sua biomecânica ⁸, com susceptibilidade ao desenvolvimento de ceratocone. Olhos com córneas levemente assimétricas, geralmente, apresentam boa acuidade visual e podem ou não ter graus oculares de miopia ou hipermetropia e astigmatismo, sendo classificados como “córneas suspeitas”. As Figuras 1.2 e 1.3 ilustram, respectivamente, exemplos de córnea normal e córnea com discreta presença de assimetria na sua curvatura, respectivamente.

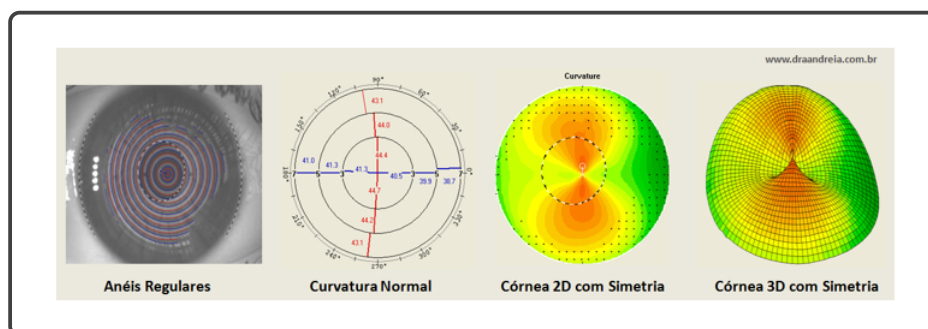


Figura 1.2: Ilustração de córnea normal
Fonte: <http://www.drandreia.com.br> (2021)

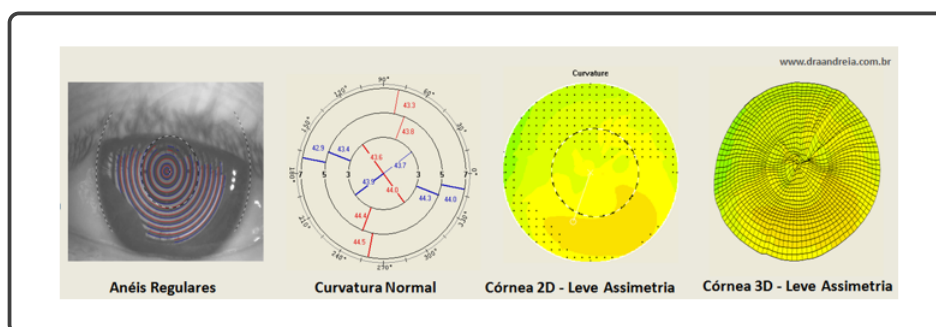


Figura 1.3: Ilustração de córnea assimétrica
Fonte: <http://www.drandreia.com.br> (2021)

As córneas com ceratocone subclínico são córneas propensas a desenvolverem o ceratocone se forem expostas a certas condições propícias para o surgimento de alterações biomecânicas, como: hábito de coçar os olhos, traumas, mudanças hormonais, cirurgia refrativa LASIK, entre outras [21]. Estas córneas já possuem uma discreta assimetria, acima do padrão esperado de normalidade. O diagnóstico de córneas com ceratocone é baseado em critérios topográficos/ou tomográficos das medidas da curvatura da córnea. A severidade do diagnóstico pode ser classificada em Grau I, II, III ou IV, de acordo com a gravidade do estágio evolutivo [37]. Esta classificação utiliza os parâmetros da curvatura

⁸Estudo da estrutura e função do globo ocular a partir das leis da Física.

da córnea, do erro refrativo (grau da miopia/astigmatismo), da paquimetria (espessura central da córnea) e da biomicroscopia (córnea transparente ou com opacidade e/ou perfuração), conforme é ilustrado na Figura 1.4, sendo portanto decisiva para a seleção de pacientes candidatos à cirurgia refrativa.

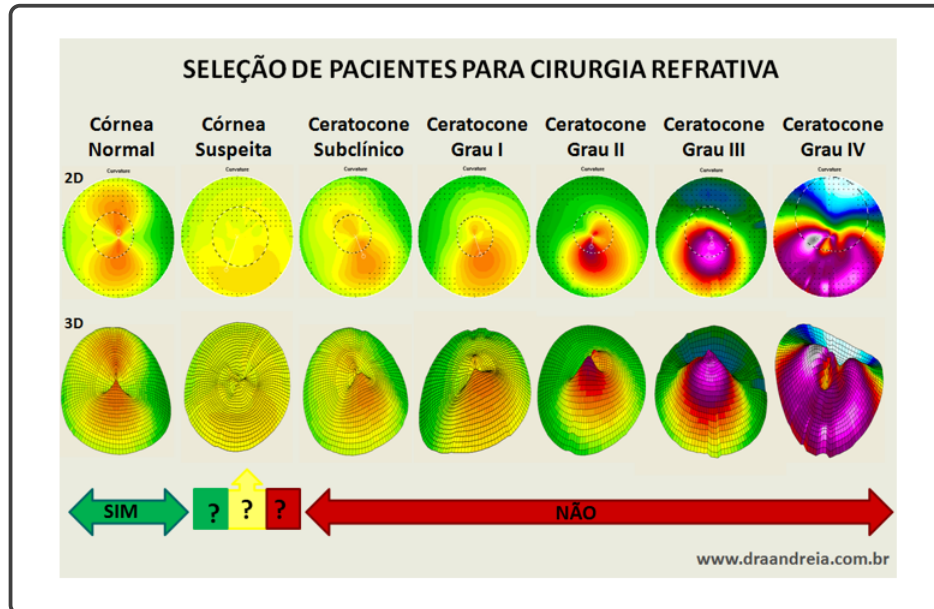


Figura 1.4: Ilustração da evolução do ceratocone e classificação de Krumeich

Fonte: <http://www.drandreia.com.br> (2021)

Estudos [43] indicam que os índices biomecânicos, paquimétricos e ceratométricos da córnea são essenciais para analisar fatores de risco para o desenvolvimento de ceratocone, e podem auxiliar na detecção precoce da doença, antes da evolução para os estágios mais avançados da doença. É importante destacar que os critérios topográficos não devem ser analisados de forma isolada, mas sim aplicados em conjunto com o julgamento clínico, para um correto diagnóstico do paciente.

1.2 Caracterização do Problema

Ceratocone é uma enfermidade que provoca alterações no formato e espessura da córnea, deixando-a mais fina e menos resistente. Pacientes que foram submetidos à cirurgia refrativa a laser estão mais propensos a desenvolver a enfermidade pelo afinamento corneano, podendo causar cegueira e exigir o transplante de córnea [29]. O foco da pesquisa é metodológico e aplicado em Ciência de Dados. O objetivo da proposta é auxiliar no diagnóstico médico de pacientes candidatos à cirurgia refrativa LASIK com risco de desenvolver o ceratocone. As metodologias experimentais tradicionais utilizadas largamente na literatura estão sujeitas a variância nas estimativas em função do esquema de amostra-

gem aplicado. Neste contexto, é apresentada uma proposta de metodologia experimental para a obtenção de classificadores robustos com capacidade de detecção de pacientes com risco de ceratocone. Para a condução desta pesquisa, as medidas oftalmológicas foram disponibilizadas por uma clínica de olhos.

1.3 Objetivos

O objetivo desta pesquisa é propor uma metodologia experimental que permita detectar com robustez e confiabilidade estágios iniciais de ceratocone em pacientes candidatos à cirurgia refrativa, aplicando algoritmos de aprendizado de máquina para classificar córneas normais de portadores de risco de ceratocone. A contribuição esperada inclui a obtenção de um modelo de aprendizado de máquina que apresente boa capacidade preditiva em novas amostras de dados.

1.4 Justificativa

As metodologias tradicionais presentes na literatura para geração de modelos de aprendizado de máquina, em sua maioria, seguem protocolos experimentais para maximizar as medidas de desempenho. Porém, nem sempre há garantia de boa generalização para novas amostras de dados. Portanto, torna-se fundamental o desenvolvimento de uma metodologia experimental que produza resultados robustos para as próximas eventuais amostras.

1.5 Organização do Trabalho

A pesquisa está estruturada da seguinte maneira: No Capítulo 1 são abordados os aspectos gerais e introdutórios da pesquisa, como a contextualização e motivação do problema, objetivos, justificativa e organização. No Capítulo 2 é apresentado a caracterização do problema. No Capítulo 3 é apresentada a fundamentação teórica dos conceitos de aprendizado de máquina, incluindo um mapeamento de trabalhos de ML para o diagnóstico de ceratocone. No Capítulo 4 é apresentada uma proposta de metodologia experimental para geração de classificadores robustos, aplicados ao problema de detecção de risco de ceratocone. No Capítulo 5 é apresentado o experimento computacional para validação da proposta metodológica. No Capítulo 6 são apresentados os resultados e discussão para a metodologia experimental proposta do Capítulo 4. E, por fim, no Capítulo 7 são apresentados os comentários, limitações e considerações finais.

2. Caracterização do Problema

Neste capítulo apresenta-se a caracterização do problema que servirá de guia para a investigação científica da pesquisa. Na Seção 2.1 é apresentada a anatomia da córnea. Na Seção 2.2 é apresentado o impacto da cirurgia refrativa LASIK. Na Seção 2.3 é apresentado o cálculo da espessura residual do estroma, medida pré-determinante para a cirurgia LASIK. Nas Seções 2.4 e 2.5 são apresentados, respectivamente, o problema e diagnóstico do ceratocone, e na Seção 2.6 é apresentada uma análise exploratória preliminar sobre os dados.

Na literatura relacionada ao problema do ceratocone é comum o uso de termos técnicos na língua inglesa. A Tabela 2.1 apresenta a relação dos termos técnicos utilizados nesta dissertação, e a correspondente descrição em inglês e português. No decorrer do texto, os termos em inglês e português poderão ser utilizados de forma alternada.

2.1 Anatomia da Córnea

A córnea é a parte saliente e anterior do globo ocular [65]. É totalmente transparente, contribui com cerca de $2/3$ do poder dióptrico ocular, apresenta espessura central aproximada de 0,6 mm e periférica de 1,3 mm, e diâmetro médio de 12 mm, variando entre 11 mm e 12,5 mm. A curvatura da córnea é discretamente mais acentuada (em média +44,00 D) na vertical do que na horizontal, e se aplanar à medida que se afasta da zona óptica central [10, 14]. A Figura 2.1 apresenta a anatomia de uma córnea normal e outra com a condição de ceratocone. Pode-se notar que a córnea com ceratocone possui um formato de “cone” ou “sino” com grande protuberância para fora do globo ocular.

A córnea, por ser responsável por mais de 50% da função de foco do olho, precisa ser convexa e finamente transparente para formar as imagens na retina [10, 23]. A Figura 2.2 exibe, de forma simplificada, o esquema do globo ocular. Pode-se observar a represen-

Tabela 2.1: Relação de termos técnicos

Termo	Descrição em inglês	Descrição em português
μm	<i>Micron</i>	Micrômetro
D	<i>Diopter</i>	Diopia
CCT,CT	<i>Central Corneal Thickness</i>	Espessura Central da Córnea
PTA	<i>Percentage of Tissue Altered</i>	Percentual de Tecido Alterado
AD	<i>Ablation Depth</i>	Profundidade da Ablação
Femto	<i>Femtosecond</i>	Femtosegundo
Flap	<i>Flap</i>	Retalho
FT	<i>Flap Thickness</i>	Espessura do Retalho
RSB	<i>Residual Stromal Bed</i>	Leito Estromal Residual
RSBMin	<i>Minimum Residual Stromal Bed</i>	RSB no ponto mínimo
RSBCentral	<i>Central Residual Stromal Bed</i>	RSB no ponto central
PachyApex	<i>Pachymetric Value At the Corneal Apex</i>	Espessura da córnea no ápice
PTACentral	<i>PTA At the Central Corneal Thickness</i>	PTA no ponto central
PTAMin	<i>PTA At the Minimum Corneal Thickness</i>	PTA no ponto mínimo
ThinnestPachy	<i>Pachymetry Thinnest Location</i>	Espessura da córnea no ponto mais fino
DifPachyCentralMin	<i>Pachymetry Difference (PachyApex - ThinnestPachy)</i>	Diferença da espessura (PachyApex - ThinnestPachy)
SphEq	<i>Spherical Equivalent Error</i>	Erro Esférico Equivalente
PreopSph	<i>Preoperative Sphere</i>	Esfera no pré-operatório
PreopCyl	<i>Preoperative Cylinder</i>	Cilindro no pré-operatório
PreopAxis	<i>Preoperative Axis</i>	Meridiano da lente no pré-operatório

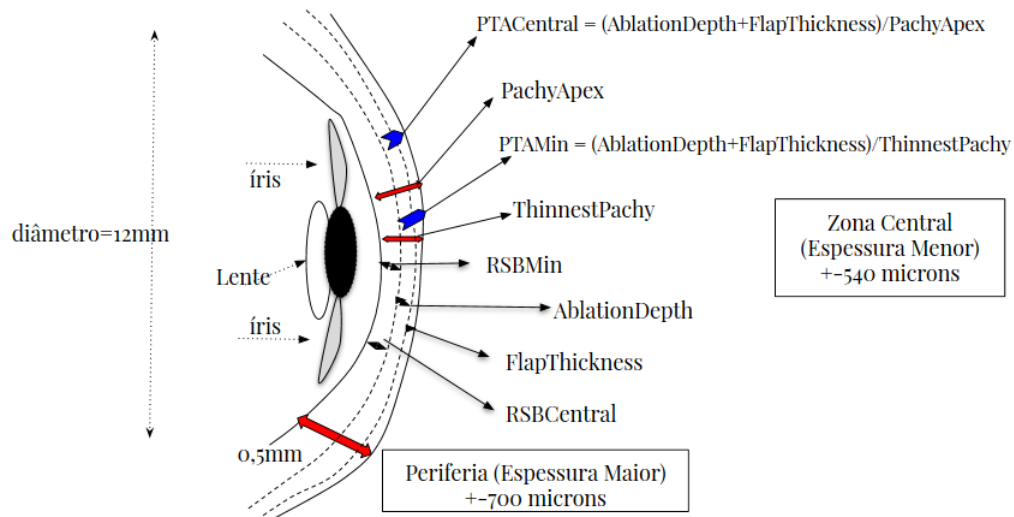


Figura 2.3: Medidas da córnea - seção transversal

A espessura da córnea contém três camadas distintas: (i) O epitélio, que é a camada mais externa, bem parecida com a superfície da pele e cujas células são reconstituídas toda semana. (ii) O estroma, que é a camada intermediária, e a mais larga da córnea. Esta camada concentra as fibras de colágeno, fundamentais para a estrutura e força da córnea, sendo mais rígida na camada externa, e mais frouxa e flexível na camada interna. Contém cerca de 500 μm de espessura, o que representa aproximadamente 90% da espessura total da córnea [14]. (iii) O endotélio, que é a camada mais interna da córnea, tem a função de hidratação das camadas, bombeando fluido (íons de sódio, potássio e água) para a córnea, deixando-a seca e transparente. Esta camada não tem capacidade regenerativa e, à medida que envelhecemos, o número de células diminui bem como sua espessura. Entre a camada de epitélio e o estroma encontra-se a membrana de Bowman. Em adultos, esta camada varia entre 8-12 μm de espessura [27]. Entre a camada do endotélio e o estroma existe uma outra camada denominada membrana de Descemet. Esta membrana não possui células e é muito resistente [14]. A Figura 2.4 exibe o detalhamento de cada camada da córnea.

2.2 Impacto da Cirurgia Refrativa - LASIK

As cirurgias refrativas são cirurgias de correção de grau como nos casos de astigmatismo, miopia e hipermetropia. O erro refrativo ocorre quando a luz não chega com nitidez à retina para a formação de uma imagem, que será enviada ao cérebro. A cirurgia LASIK consiste na remoção de tecido corneano. Ou seja, reformula a superfície da córnea para corrigir o erro de refração. A quantidade de tecido da córnea que é removida do estroma é chamada de profundidade da ablação. Na miopia, o tecido da córnea é removido do centro

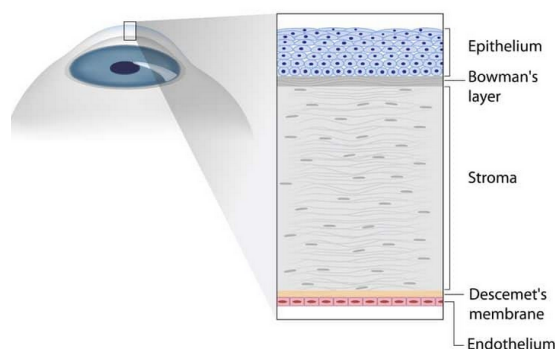


Figura 2.4: Camadas da córnea. Onde se lê *Epithelium*, leia-se Epitélio; onde se lê *Bowman's layer*, leia-se camada de Bowman; onde se lê *Stroma*, leia-se Estroma; onde se lê *Descemet's membrane*, leia-se membrana de Descemet; e onde se lê *Endothelium*, leia-se Endotélio.

Fonte: <https://parkercornea.com/keratoconus-2/bowman-layer-transplant/> (2019)

para achatar a superfície da córnea. Na hipermetropia, o tecido da córnea é removido da periferia para aumentar a superfície da córnea. O restante do tecido corneano subjacente é chamado de leito estromal residual. Existem três camadas a serem consideradas ao realizar a cirurgia LASIK: a espessura do retalho (FT), a profundidade da ablação (AD) e o leito estromal residual (RSB). Todos os três precisam ser levados em consideração antes que a cirurgia LASIK possa ser executada. A profundidade da ablação depende do diâmetro da pupila, do grau de erro de refração que precisa ser corrigido e dos parâmetros da cirurgia LASIK [52]. A Figura 2.5 mostra o esquema das camadas consideradas na cirurgia LASIK e a profundidade de ablação necessária para a correção da visão. A organização americana “United States Food and Drug Administration” (FDA) recomenda um mínimo de 250 μm de espessura de leito estromal residual, ou seja, este é o valor mínimo indicado para se preservar a integridade da córnea e para se evitar complicações no pós-operatório.

Durante a cirurgia LASIK, uma incisão de retalho é criada na camada externa epitelial da córnea. Com a aba aberta, o cirurgião pode remodelar as camadas estromais da córnea para obter a forma precisa sobre a qual a luz pode refletir diretamente na retina e produzir uma visão clara. Como requisito pré-operatório, o paciente deve ter uma córnea suficientemente espessa. Se a espessura corneana não for suficiente para suportar o retalho, maior é o risco de complicações na cirurgia LASIK, como a ectasia corneana, um abaulamento da córnea que pode realmente piorar a saúde e visão do paciente. Como a segurança do paciente é a prioridade, a cirurgia LASIK não deve ser realizada em córneas finas ou fracas [36]. A Figura 2.5 exhibe as camadas a considerar na cirurgia LASIK e o quadro da Tabela 2.2 exhibe a espessura necessária (em micrômetros) para a profundidade da abla-

ção, variando com o erro de refração a ser corrigido e o diâmetro da pupila, baseada na Fórmula de Munnerlyn [47]. Segundo o autor, a profundidade da ablação = diâmetro da ablação ao quadrado x grau de correção (em dioptrias).

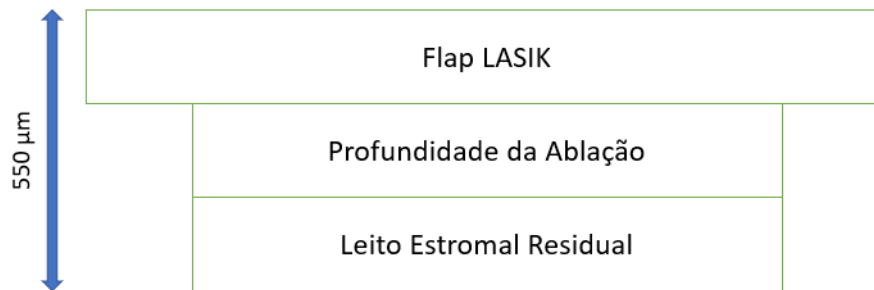


Figura 2.5: Esquema das camadas na cirurgia LASIK
Fonte: Camadas adaptadas de <http://www.lasikguider.com>

Tabela 2.2: Medidas da profundidade da ablação (em micrômetros)

Grau de correção do paciente (em <i>diopters</i>)	Ablação (µm) (pupila de 5 mm)	Ablação (µm) (pupila de 6 mm)	Ablação (µm) (pupila de 7 mm)
-1,00	-8,3	-12,0	-16,3
-2,00	-16,7	-24,0	-32,7
-3,00	-25,0	-36,0	-49,0
-4,00	-33,3	-48,0	-65,3
-5,00	-41,7	-60,0	-81,7
-6,00	-50,0	-72,0	-98,0
-7,00	-58,3	-84,0	-114,3
-8,00	-66,7	-96,0	-130,7
-9,00	-75,0	-108,0	-147,0
-10,00	-83,3	-120,0	-163,3
-11,00	-91,7	-132,0	-179,7
-12,00	-100,0	-144,0	-196,0

Fonte: Dados adaptados de https://www.lasik1.com/LASIK_Tables.html (2021)

Em experimento realizado por Randleman et al. [52] foram avaliados 171 casos de ectasia. A ectasia ocorreu em 164 casos (95,9%) com LASIK e em sete casos (4,1%) com PRK ¹. Os autores compararam os pacientes com olhos normais e pacientes que desenvolveram a ectasia corneana após a cirurgia LASIK e observaram que este último grupo apresentava topografias preoperativas anormais (35,7% versus 0%); eram mais jovens (34,4 versus 40,0 anos); mais míopes (-8,53 versus -5,09 dioptrias); possuíam córneas mais finas antes da cirurgia (521,0 versus 546,5 µm); e apresentaram menor espessura do

¹Photorefractive Keratectomy

leito estromal residual (256,3 versus 317,3 μm). No experimento realizado, os autores aplicaram regressão logística e verificaram que a topografia anormal era o fator mais significativo para discriminação dos casos, seguido da espessura RSB, idade, e espessura da córnea no pré-operatório. A partir desta análise, Randleman et al. [54] propuseram um sistema de ranqueamento para candidatos à cirurgia LASIK denominado ERSS (*Ectasia Risk Score System*), que permitiu a análise de imagens de Plácido, sendo fundamental na evolução da topografia corneana. Aplicando o sistema de ranqueamento ERSS no Orbscan II, os autores propuseram cinco fatores de risco em ordem de importância, baseada em uma pontuação entre 0 a 4: 1-Topografia pré-operatória anormal; 2- Espessura do leito estromal residual baixo ($\text{RSB} < 250 \mu\text{m}$); 3- Idade jovem (< 30 anos); 4- Baixa espessura da córnea no pré-operatório ($\text{CT} < 510 \mu\text{m}$); 5- Miopia alta. Embora seja hoje considerado um sistema de pontuação controverso [45], o ERSS ainda é relevante para candidatos à cirurgia refrativa, sendo reconhecido por combinar distintos fatores, em contraste aos critérios individuais adotados para a análise, como $\text{CT} < 500 \mu\text{m}$ ou $\text{RSB} < 250 \mu\text{m}$.

2.3 Cálculo da Espessura Residual do Estroma

A profundidade da ablação (tecido removido do estroma) é restrita pela medida de paquimetria, ou seja, pela espessura da córnea. A espessura da córnea é medida para determinar a adequação ao procedimento de LASIK. Segundo trabalhos da literatura [52, 59], o mínimo absoluto de espessura residual é de 250 μm . Se a espessura residual da córnea for menor que 250 μm , o risco de alterações na forma da córnea aumenta. A maioria dos cirurgiões prefere deixar cerca de 270-275 μm de córnea residual, a fim de fornecer uma margem de segurança ao paciente.

A cirurgia envolve cerca de 160 μm de tecido para o retalho corneano e a remoção de 12 a 14 μm para cada grau de diopia da visão a ser corrigida. Por exemplo, para pacientes com -3,00 graus de miopia, 36 (3x12) a 42 (3x14) μm de tecido da córnea são removidos durante a cirurgia LASIK. Adicione isso aos 160 μm do retalho da córnea, além dos 250 μm que devem permanecer após a cirurgia, o paciente com prescrição de -3D precisaria de pelo menos 452 μm de tecido da córnea no pré-operatório para ser considerado um bom candidato à cirurgia. Para a maioria dos pacientes, a espessura da córnea varia entre 540 a 550 μm , e é medida por um instrumento chamado de paquímetro. Pacientes com córnea fina e com alto grau de diopia podem não ter espessura corneana suficiente para se submeter à cirurgia LASIK [59]. A seguir, dois exemplos de cálculo da espessura corneana. Para fins de ilustração, considera-se a espessura típica do retalho LASIK de 160 μm , e uma pupila de 6 milímetros de diâmetro. Uma diopia de prescrição requer

a remoção de 12 µm de tecido da córnea. No exemplo a seguir, o paciente apresenta espessura corneana suficiente para se submeter à cirurgia LASIK:

- Erro refrativo do paciente: -8,00 dioptrias;
- Valor a eliminar: $8D \times 12 \mu m = 96 \mu m$;
- Espessura da córnea do paciente no ponto mais fino: 580 µm;
- Espessura do retalho: 160 µm;
- Espessura residual mínima do leito estromal: 250 µm.

De acordo com a fórmula da espessura da córnea no ponto mais fino, temos que: $\text{ThinnestPachy} = \text{AblationDepth} + \text{FlapThickness} + \text{RSBMin}$. Portanto, deriva-se que: $\text{AblationDepth} = \text{ThinnestPachy} - \text{FlapThickness} - \text{RSBMin} = 580 - 160 - 250 = 170 \mu m$. Neste exemplo, o paciente precisaria eliminar 96 µm de 170 µm de tecido estromal da córnea. Isso significa que a córnea é espessa o suficiente para a cirurgia LASIK prosseguir. O leito estromal residual, neste caso, será de 324 µm, o que é muito seguro e apresenta baixo risco de complicações.

Considere o próximo exemplo, em que o paciente possui espessura corneana fina para a realização da cirurgia LASIK:

- Erro refrativo do paciente: -14,00 dioptrias;
- Valor a eliminar: $14D \times 12 \text{ micrômetros} = 168 \mu m$;
- Espessura da córnea do paciente: 510 µm;
- Espessura do retalho LASIK: 160 µm;
- Espessura residual mínima do leito estromal: 250 µm;
- Tecido disponível para ablação: $510 - 160 - 250 = 100 \mu m$.

Neste exemplo, o paciente precisaria eliminar 168 µm de 100 µm de tecido estromal da córnea. Portanto, a córnea é fina demais para ser submetida à cirurgia LASIK. Se a cirurgia LASIK fosse realizada, haveria um risco aumentado de ectasia da córnea. A Figura 2.6 ilustra um esquema de medidas consideradas na cirurgia LASIK.

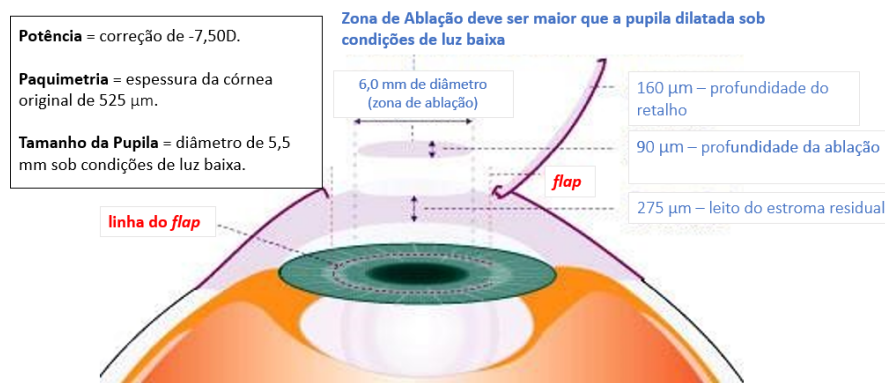


Figura 2.6: Medidas consideradas na cirurgia LASIK.

Fonte: Adaptado de https://www.lasik1.com/LASIK_Envelope.html (2021)

2.4 O Problema do Ceratocone

O desenvolvimento do ceratocone pode acontecer não apenas em córneas com suscetibilidade elevada ao ceratocone no pré-operatório, mas também em córneas normais que sofreram grande impacto biomecânico, decorrentes do ato de esfregar os olhos, e/ou pré-disposição genética, entre outros [24, 36]. Desta forma, é fundamental avaliar dados da fotoablação (ablação por laser) e do estroma residual, para identificação dos riscos à cirurgia refrativa. Neste contexto, pode-se citar o indicador PTA, desenvolvido por Santiago et al. [59], para detectar pacientes com risco de alterações na córnea. Este indicador leva em consideração a relação entre a espessura da córnea no pré-operatório, a profundidade de ablação (quantidade de tecido a ser removido para correção) e a espessura do retalho.

Em outro trabalho, os autores classificaram os estágios do ceratocone segundo sua severidade. Por ser uma doença progressiva, a classificação de Krumeich [37] é amplamente empregada e utiliza parâmetros da curvatura da córnea (curvatura média nos 3 mm da córnea), do erro refrativo (grau da miopia/astigmatismo), da paquimetria (espessura central da córnea) e da biomicroscopia (córnea transparente ou com opacidades e/ou perfuração) para definir a extensão da doença, com graus de I a IV. O grau IV representa a fase mais acentuada da doença. A Tabela 2.3 apresenta a tabela de classificação de Krumeich.

Tabela 2.3: Classificação de Krumeich

Graus	Ceratometria Média (D)	Espessura Corneal (μm)	Equivalente Esférico (D)
I	< 48	> 500	< -5
II	48 – 53	400 – 500	-5 a -8
III	54 – 55	200 – 400	> -8
IV	> 55	< 200	Não mensurável

Fonte: Krumeich, 1998 [37]

Mesmo antes do ceratocone ser identificado, mudanças na estrutura da córnea são notadas. Atualmente, a explicação mais plausível que dispara a formação do ceratocone está associada a um conjunto de fatores relacionados a mudanças na estrutura da córnea [24]. Em estudo realizado por Blackburn et al. [7], os autores apresentaram um resumo das estruturas mais impactadas da córnea com o desenvolvimento do ceratocone, dentre as quais é possível citar alterações na estrutura da córnea, no filme lacrimal, no epitélio, no colágeno da membrana epitelial, no colágeno da camada de Browman, e no colágeno da camada do estroma.

2.5 Diagnóstico de Ceratocone

A determinação da paquimetria ocular é fundamental no diagnóstico de ceratocone. Este exame tem por finalidade medir a espessura da córnea, e é indicado principalmente a pacientes candidatos à cirurgia refrativa, de catarata ou de glaucoma [22]. Com este exame, pode-se obter um mapa tridimensional da espessura em cada ponto da córnea, sendo fundamental na avaliação pré-operatória da cirurgia refrativa. A Figura 2.7 ilustra um exemplo de mapa tridimensional refrativo. Cores vermelhas indicam a existência de uma elevação. Cores claras, tais como verde ou amarelo, mostram uma distribuição uniforme da elevação do parâmetro. Portanto, os mapas refrativos permitem verificar alterações na curvatura da face dorsal (posterior) ² da córnea e no mapa paquimétrico da córnea, que podem ocorrer antes de alterações na curvatura da face frontal (anterior) ³. O equipamento Pentacam também permite a geração de análises automatizadas, comparando os mapas com referências numéricas de medidas normais e alteradas.

Mais recentemente, novas tecnologias vêm sendo introduzidas como o aparelho Oculus Corvis ST ⁴, que permite a avaliação da pressão intra-ocular e das características biomecânicas da córnea. Este aparelho permite medir a Pressão Intraocular (PIO), injetando um jato de ar e capturando mais de 4.300 imagens por segundo, possibilitando registrar com uma câmera de alta precisão a PIO, a espessura corneana, e as métricas biomecânicas da córnea, com base nas imagens de Scheimpflug (Pentacam HR ⁵).

²Região do globo ocular delimitada pela íris e o cristalino

³Região do globo ocular delimitada pela córnea e a íris

⁴*Corneal Visualization Scheimpflug Technology*

⁵<https://www.pentacam.com/int/opticianoptometrist-with-pentacamr/models/pentacamr-hr/core-functions.html>

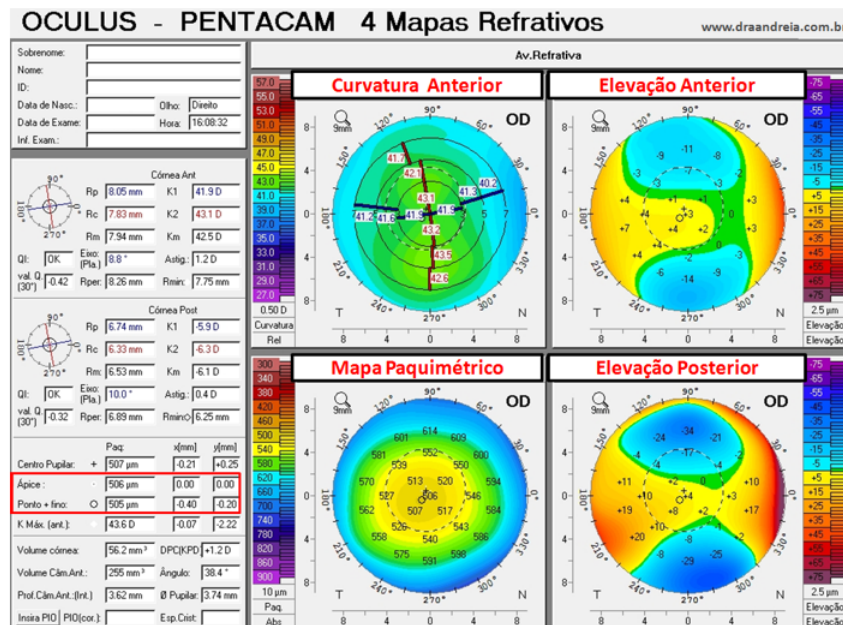


Figura 2.7: Mapas refrativos capturados pelo Oculus Pentacam

Fonte: <https://www.draandrea.com.br/wp-content/uploads/2017/09/tomogr2.png> (2021)

2.6 O Problema Computacional

Para a condução desta pesquisa, uma amostra com medidas oftalmológicas foi disponibilizada por uma clínica de olhos. A Seção 2.6.1 apresenta as variáveis da amostra, a Seção 2.6.2 apresenta uma análise preliminar exploratória dos dados, e a Seção 2.6.3 apresenta um estudo de correlações entre as variáveis.

2.6.1 Conjunto de Dados

O conjunto de dados disponibilizado contém 3.383 registros de pacientes com informações oculares distribuídas em 18 variáveis. A variável categórica e binária IDE (“Yes” ou “No”) indica o risco do ceratocone, mais precisamente, designa a suscetibilidade a desenvolver ceratocone após a cirurgia LASIK. O conjunto de dados contém 3.278 registros de olhos saudáveis (IDE = “No”) e 105 registros de olhos com risco de ceratocone (IDE = “Yes”). As seguintes variáveis foram informadas na amostra:

1. IDE: Indica o risco de ceratocone (“Yes”) ou a ausência de risco de ceratocone (“No”).
2. AGE: Indica a idade do paciente. Idade mais baixa (menor que 25 anos) pode ser um fator de risco significativo para o desenvolvimento da ectasia em pacientes sem outros fatores de risco.

3. FlapTechnique: Nome da técnica de cirurgia a laser (Femto, Intralase, Microkeratome).
4. FlapThickness: Indica a espessura do retalho. Com o desenvolvimento da técnica de Femto, a maioria dos cirurgiões utiliza uma espessura do retalho próxima a 100 µm. Com a técnica de microcerátomo (Microkeratome), as espessuras mais usadas são de 120 µm. Se a espessura do retalho for maior que o planejado, menor é o leito estromal para ablação. Nestes casos, a cirurgia LASIK deve ser evitada.
5. AblationDepth: Indica a profundidade da ablação. A fórmula de Munnerlyn [47] auxilia a estimar a profundidade da ablação nas correções de miopia. A fórmula é dada por: $t = S^2D/3$, onde: t = profundidade da ablação em µm; S = diâmetro da zona óptica em milímetros; D = correção em dioptrias (SphEq).
6. FlapAge: Tempo do retalho em dias.
7. ExamEye: Olho Direito (em inglês, *Right*), Esquerdo (em inglês, *Left*).
8. PreopSph: Indica a potência da lente, medida em dioptrias, para corrigir o astigmatismo ou miopia. Um sinal de menos (-) indica míope; um sinal de mais (+) indica astigmatismo. O termo “esfera” significa que a correção para a miopia ou astigmatismo é “esférica”, ou igual em todos os meridianos do olho.
9. PreopCyl: Indica o poder da lente para o astigmatismo. O termo “cilindro” significa que não é esférico. O número do cilindro pode ser precedido por um sinal de “-” (para correção de miopia com astigmatismo) ou um sinal de “+” (para hipermetropia com astigmatismo). O poder do cilindro sempre segue o poder da esfera em uma prescrição de óculos.
10. PreopAxis: Descreve o meridiano da lente para corrigir o astigmatismo. O eixo (em inglês, *axis*) é um número entre 1 e 180. O número 90 corresponde ao meridiano vertical do olho, e o número 180 corresponde ao meridiano horizontal. Se uma prescrição de óculos inclui a potência do cilindro (PreopCyl), também deve incluir um valor do eixo.
11. SphEq: Indica o equivalente esférico, que é uma estimativa do erro refrativo. É calculado pela fórmula: $SphEq = PreopSph + 1/2 * PreopCyl$, sendo PreopSph e PreopCyl os componentes do erro refrativo.
12. ThinnestPachy: Indica a espessura no ponto mais fino da córnea. Valores menores que 500 µm são considerados córneas finas. Valores menores que 470 µm são considerados córneas altamente suspeitas de ceratocone.

13. **PachyApex:** Paquimetria da espessura no ápice da córnea. Pacientes com córneas muito finas estão no grupo de riscos para desenvolver o ceratocone no pós-LASIK.
14. **DifPachyCentralMin:** Uma diferença da paquimetria entre o ápice e o ponto mais fino superior a 5 μm já é um indicativo para o desenvolvimento do ceratocone. Em outras palavras, as córneas com ceratocone afinam mais rapidamente da periferia para o ponto mais fino do que as córneas normais.
15. **RSBCentral:** Indica a espessura do leito estromal residual no ponto central da córnea.
16. **RSBMin:** Indica a espessura do leito estromal residual no ponto mais fino. A espessura da RSB é especialmente importante após a LASIK pois a cirurgia reduz a integridade estrutural da córnea [34, 53]. O risco de ceratocone aumenta para RSB menor que 300 μm .
17. **PTACentral:** Indica o percentual de tecido alterado no ponto central.
18. **PTAMin:** Indica o percentual de tecido alterado no ponto mínimo. Um PTA acima de 40% aumenta o risco de ceratocone em olhos com topografia pré-operatória normal. A relação destas medidas pode acometer em alterações na córnea, dada pela fórmula $\text{PTA} = (\text{FT} + \text{AD})/\text{CCT}$, onde:
 - PTA = Percentual de tecido alterado;
 - FT = Espessura do retalho;
 - AD = Profundidade da ablação;
 - CCT = Espessura da córnea no pré-operatório.

A despeito de todos os métodos e estratégias descritos na literatura, não há uma concordância universal sobre os critérios mínimos para o diagnóstico do ceratocone.

2.6.2 Análise Exploratória dos Dados

A fim de se identificar as causas de maior relevância para o diagnóstico do ceratocone, é fundamental a análise exploratória sobre a amostra de pacientes que apresentaram risco de ceratocone. A possibilidade de detecção de padrões de correlação entre variáveis, a compreensão de comportamentos anômalos e a validação de hipóteses são alguns dos objetivos da análise exploratória sobre os dados observados. As seguintes conjecturas foram levantadas sobre os dados da amostra:

1. A cirurgia LASIK deixa a córnea mais fina: a espessura no ponto mais fino (ThinnestPachy) e espessura no ápice (PachyApex) estão diretamente relacionadas ao afinamento da córnea. A cirurgia LASIK também deixa a córnea mais plana, caracterizada pela diferença (DifPachyCentralMin) da espessura entre o ápice (PachyApex) e ponto mais fino (ThinnestPachy).
2. Córneas com leito estromal mais fino (RSB menor que 250 μm) têm mais risco de desenvolver ceratocone: 100% dos pacientes com RSB menor que 250 μm foram classificados com risco de ceratocone, referente a 3,6% dos pacientes com risco de ceratocone.
3. Pacientes mais jovens têm mais risco de desenvolver ceratocone: 55% dos pacientes com risco de ceratocone apresentaram idade menor que 30 anos.
4. Pacientes mais míopes (PreopSph e SphEq maior que -5D) possuem maior risco de desenvolvimento do ceratocone pois necessitam de um valor a eliminar maior para a correção da visão: 28,1% dos pacientes com risco de ceratocone apresentaram grau de correção acima de -5D.
5. Valor a eliminar maior que a profundidade da ablação (AblationDepth) aumenta o risco de ceratocone: 21,8% dos pacientes com risco de ceratocone apresentaram valores a eliminar maior que a profundidade da ablação.
6. Espessura do retalho maior que 120 μm associada à profundidade da ablação maior que 70 μm aumenta o risco de ceratocone: 20,9% dos pacientes com risco de ceratocone apresentaram FT maior que 120 μm e AD maior que 70 μm .
7. PTACentral/PTAMin acima de 40% aumenta o risco de ceratocone: 32,7% dos pacientes com risco de ceratocone possuíam PTACentral acima de 40%.
8. A FlapTechnique caracterizada por “Microkeratome” apresenta uma taxa média de descamação em 158,9 μm , ou seja, 3% acima das técnicas por Femto (119,9 μm) e Intralase (120 μm).
9. Míopes possuem, em média, o leito estromal residual mais fino que hipermétropes: RSBMin = 361 μm (míope) versus RSBMin = 387 μm (hipermétrope).
10. Astigmatismo (PreopCyl maior que zero) é um fator que leva ao leito estromal mais fino: RSBMin médio de 359 μm em míopes com astigmatismo versus RSBMin médio de 371 μm em míopes sem astigmatismo.
11. Hipermétropes com astigmatismo tem RSBMin médio de 344 μm e hipermétropes sem astigmatismo tem RSBMin médio de 387 μm .

12. PTA alto, acima de 40%, é o fator individual mais efetivo para prever o risco de ceratocone: 34% dos pacientes com risco de ceratocone apresentaram PTA alto.

2.6.3 Correlação entre Variáveis

O estudo das correlações analisa o grau de correlação entre as variáveis independentes. As seguintes equações são inferidas a partir das regras aplicadas às métricas da córnea:

1. Espessura no ponto mais fino da córnea é dada pela equação: $\text{ThinnestPachy} = \text{AblationDepth} + \text{FlapThickness} + \text{RSBMin}$.
2. Espessura no ápice da córnea é dada pela equação: $\text{PachyApex} = \text{AblationDepth} + \text{FlapThickness} + \text{RSBCentral}$.
3. Diferença da espessura entre o ponto central e o de menor espessura é dada pela equação: $\text{DifPachyCentralMin} = \text{PachyApex} - \text{ThinnestPachy}$.
4. Percentual de tecido alterado no ponto central é proporcional à relação entre a espessura do retalho, a profundidade da ablação e a espessura da córnea no ápice, dado pela equação: $\text{PTACentral} = (\text{FlapThickness} + \text{AblationDepth}) / \text{PachyApex}$.
5. Percentual de tecido alterado no ponto mínimo é proporcional à relação entre a espessura do retalho, a profundidade da ablação e a espessura do ponto mínimo da córnea no pré-operatório, dado pela equação: $\text{PTAMin} = (\text{FlapThickness} + \text{AblationDepth}) / \text{ThinnestApex}$.
6. Equivalência Esférica é dada pela equação: $\text{SphEq} = \text{PreopSph} + 1 / 2 * \text{PreopCyl}$.
7. Presença de astigmatismo ⁶: se $\text{PreopCyl} > 0$, PreopAxis varia entre 0 e 180 graus.
8. Valor a eliminar ⁷ = $\text{SphEq} * 14 \mu\text{m}$. Valores a eliminar maior que a profundidade da ablação (AblationDepth) aumentam o risco de ceratocone (1D equivale à 12-14 μm).
9. Tipo de correção ⁸ = $\text{PreopSph} > 0$ (hipermetropia); $\text{PreopSph} < 0$ (miopia).

⁶Característica derivada da variável PreopCyl

⁷Característica derivada do equivalente esférico (SphEq) e profundidade da ablação (AblationDepth)

⁸Característica derivada da variável PreopSph

2.7 Resumo

Em resumo, este capítulo apresentou os principais conceitos da anatomia da córnea e o impacto da cirurgia LASIK no desenvolvimento de possíveis complicações no pós-operatório. O capítulo também apresentou o racional de cálculo envolvido para avaliar a segurança de uma possível cirurgia refrativa com base nas medidas oculares e espessura corneana dos pacientes. Por fim, o capítulo contextualizou o problema do ceratocone e introduziu o tema do ponto de vista computacional, a partir do conjunto de dados disponibilizado para a condução desta pesquisa. O capítulo finaliza com uma breve análise exploratória sobre os dados. No próximo capítulo é apresentada a revisão bibliográfica para a condução desta pesquisa.

3. Revisão Bibliográfica

Este capítulo está dividido em duas seções. A primeira seção apresenta a fundamentação teórica dos conceitos de aprendizado de máquina para condução desta pesquisa. A segunda seção apresenta um mapeamento de trabalhos da literatura que aplicaram técnicas de aprendizado de máquina para o diagnóstico de ceratocone.

3.1 Fundamentação Teórica

Na Seção 3.1.1 é apresentado o conceito de aprendizado de máquina. Nas próximas quatro seções são apresentados definições dos termos básicos e métricas da matriz de confusão. Na Seção 3.1.6 é apresentado o conceito de árvore de decisão. Na Seção 3.1.7 são apresentados os conceitos de entropia, ganho de informação e índice Gini. Nas Seções 3.1.8 e 3.1.9 são apresentados os algoritmos de Floresta Aleatória e XGBoost. Na Seção 3.1.10 é apresentado o conceito de busca em grade. Na Seção 3.1.11 são apresentados protocolos experimentais de seleção de parâmetros, tais como validação cruzada por *k-fold* e *k-fold* aninhada. As três últimas seções apresentam os conceitos de interpretação local, intervalo de confiança, e o *Bootstrap* para geração dos estimadores.

3.1.1 Aprendizado de Máquina

Aprendizado de Máquina (em inglês, *Machine Learning* - ML) é um campo de estudo da Inteligência Artificial e da Ciência da Computação. Dentre as inúmeras áreas de atuação, pode-se citar o estudo dedicado à construção de algoritmos. Tais algoritmos podem ser aplicados para aprender com um dado contexto, de forma automatizada, e realizar previsões em novos contextos. Em 1959, o cientista Arthur Samuel [58] definiu o aprendizado de máquina como o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados. Modelos gerados por técnicas de

aprendizado de máquina possibilitam a tomada de decisões com base na descoberta de padrões, *insights*, relações e tendências históricas sobre dados e têm sido aplicados com sucesso em diversos campos de pesquisa: reconhecimento de padrões, visão por computador, engenharia de naves espaciais, finanças, entretenimento, biologia e medicina.

Neste contexto, os modelos preditivos são funções matemáticas que mapeiam um conjunto de variáveis de entrada (preditivas) em um valor alvo de saída, e são gerados a partir de um conjunto de dados de treinamento que contém amostras com características conhecidas (variáveis preditivas) e uma classe alvo que se deseja prever. Este processo é denominado aprendizado supervisionado: o algoritmo aprende corrigindo os erros. A tarefa de prever uma classe alvo é denominada de classificação, quando a classe alvo é uma variável categórica, ou regressão, quando a classe alvo é contínua. No contexto desta pesquisa, a detecção de risco de ceratocone é um problema de classificação binária pois há somente duas saídas para a classe alvo: “sim” para a detecção de risco de ceratocone ou “não” para a ausência de risco de ceratocone, e esta detecção é realizada por meio de modelos de classificação de aprendizado de máquina supervisionado. Estes modelos são compostos pelo algoritmo “treinado” com os dados de entrada, e têm como objetivo generalizar um comportamento para uma nova entrada de dados, de forma automatizada, robusta e assertiva. O fluxo (em inglês, *workflow*) do processo de aprendizado de máquina, contemplando a geração de modelos de ML e da predição em um problema de classificação, é ilustrado na Figura 3.1.

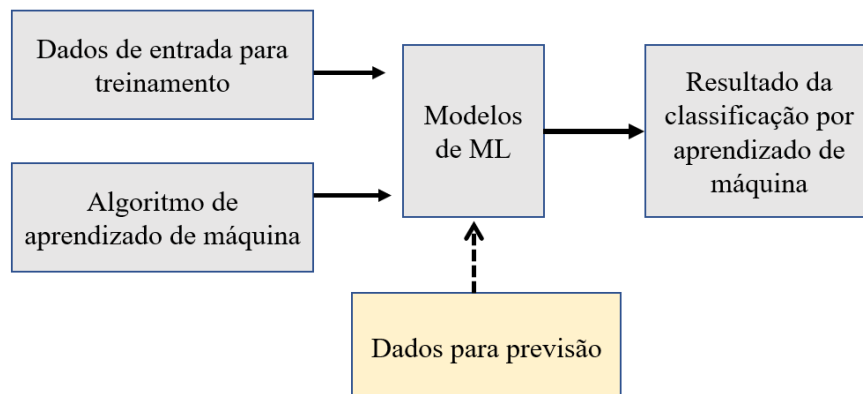


Figura 3.1: Workflow da geração de um modelo de aprendizado supervisionado para o problema de classificação

3.1.2 Matriz de Confusão

Esta seção apresenta a matriz de confusão aplicada ao problema da pesquisa. A matriz de confusão permite visualizar o desempenho do algoritmo selecionado no aprendizado de máquina, comparando os resultados gerados com a classe real. Quanto mais o mo-

delo preditivo responder corretamente, maior será a sua assertividade. Por exemplo, em problemas de classificação binária, os resultados podem ser divididos em quatro interpretações possíveis para o resultado da predição sobre os dados de teste: duas corretas e duas incorretas. Para o problema desta pesquisa, o teste está correto quando é positivo na presença de risco de ceratocone (verdadeiros positivos), ou negativo na ausência de risco de ceratocone (verdadeiros negativos). Por outro lado, o teste está incorreto quando é positivo na ausência de risco de ceratocone (falso positivo), ou negativo quando o risco de ceratocone é real (falso negativo). A matriz de confusão de um classificador binário é visualizada na Figura 3.2, e mostra a relação entre os resultados de um teste e o diagnóstico verdadeiro. Cada linha representa as instâncias previstas para a classe alvo, denominada nesta pesquisa como IDE, após a execução do modelo de ML. Cada coluna representa as instâncias reais da classe IDE. A classe IDE = “Yes” indica a presença de risco de ceratocone, e IDE = “No” representa a ausência de risco de ceratocone. Em síntese, as previsões ocorrem nos possíveis cenários:

1. Previsões corretas:

- I Verdadeiro positivo (VP): o modelo de ML previu como 1 (IDE = “Yes”) e o valor verdadeiro é 1 (IDE = “Yes”).
- II Verdadeiro negativo (VN): o modelo de ML previu como 0 (IDE = “No”) e o valor verdadeiro é 0 (IDE = “No”).

2. Previsões incorretas:

- I Falso positivo (FP): o modelo ML previu como 1 (IDE = “Yes”), mas o valor verdadeiro é 0 (IDE = “No”).
- II Falso negativo (FN): o modelo ML previu como 0 (IDE = “No”), mas o valor verdadeiro é 1 (IDE = “Yes”).

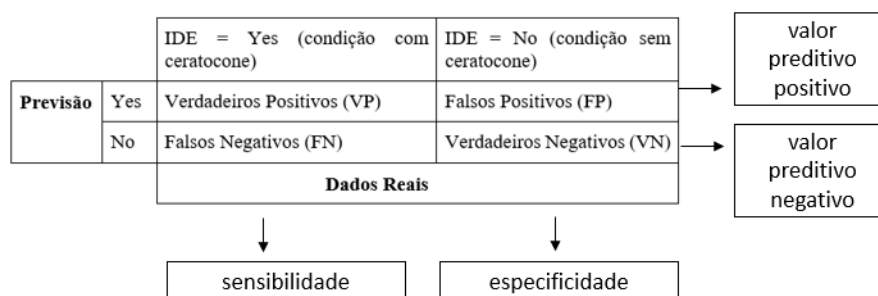


Figura 3.2: Matriz de confusão

As métricas da matriz de confusão podem mascarar, quando usadas de forma isolada, a presença de deficiências no classificador aparentemente não perceptíveis mas que podem esconder algumas fragilidades do modelo. As métricas não devem ser consideradas como inquestionáveis pois dependem da maleabilidade do teste aplicado e da predominância da classe alvo nas amostras capturadas. Neste contexto, é importante questionar o rigor empregado nos modelos de classificação levantados da literatura bem como a aderência dos dados de validação e teste em relação aos dados de treino. Portanto, faz-se necessário interpretar todas as métricas a fim de garantir a assertividade na previsão dos dados. Os termos básicos e métricas da matriz de confusão aplicados ao problema da pesquisa são apresentados a seguir:

- Verdadeiros Positivos (VP): São os casos em que o modelo previu um risco de ceratocone e os pacientes de fato possuem risco de ceratocone.
- Verdadeiros Negativos (VN): São os casos em que o modelo não previu risco de ceratocone e os pacientes de fato não são do grupo de risco de ceratocone.
- Falsos Positivos (FP): São os casos em que o modelo previu que os pacientes são do risco de ceratocone mas, na realidade, são pacientes que não desenvolveram ceratocone.
- Falsos Negativos (FN): São os casos em que o modelo não previu risco de ceratocone mas, na realidade, são pacientes que desenvolveram ceratocone.
- Precisão (PPV): A precisão é a probabilidade de que as pessoas com um resultado positivo na previsão tenham, de fato, a condição de risco de ceratocone. A fórmula é dada pela Equação 3.1 e representa a proporção de verdadeiros positivos em relação a todas as predições positivas.

$$PPV = \frac{VP}{VP + FP}, \quad (3.1)$$

onde VP é a quantidade de casos verdadeiros positivos e FP é a quantidade de casos falsos positivos.

- Sensibilidade ou Cobertura (em inglês, *Recall*): No diagnóstico médico, é comum associar a sensibilidade para indicar a capacidade do modelo em classificar corretamente os olhos na classe positiva, com risco de ceratocone. Dado que a classe positiva IDE = “Yes” representa os olhos com risco de ceratocone, a sensibilidade quantifica a probabilidade dos registros com risco de ceratocone serem efetivamente

classificados com o risco de ceratocone, e pode ser calculada conforme a Equação 3.2:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}, \quad (3.2)$$

onde VP é a quantidade de casos verdadeiros positivos e FN é a quantidade de casos falsos negativos.

- Especificidade (em inglês, *Specificity*): Mede a capacidade do modelo em detectar eventos na classe negativa (VN). Dado que IDE = “No” representa os olhos sem o risco de ceratocone, a especificidade quantifica a probabilidade dos registros de olhos normais serem efetivamente classificados sem o risco de ceratocone, e é calculada conforme a Equação 3.3.

$$\text{Especificidade} = \frac{VN}{FP + VN}, \quad (3.3)$$

onde VN é a quantidade de casos verdadeiros negativos e FP é a quantidade de casos falsos positivos.

- Acurácia (em inglês, *Accuracy*): Descreve o quanto o classificador acerta, ou seja, é o número de previsões corretas com base em todas as previsões feitas, e pode ser calculada conforme a Equação 3.4:

$$\text{Acurácia} = \frac{CC}{CO} = \frac{VP + VN}{VP + VN + FP + FN}, \quad (3.4)$$

onde CC é a quantidade de casos corretamente classificados e CO é a quantidade de casos observados.

3.1.3 Valor de Corte

Em situações de dados desbalanceados, os algoritmos de aprendizado de máquina normalmente classificam corretamente a classe majoritária (de maior proporção) porém falham na classificação da classe minoritária (de menor proporção). Este cenário é especialmente crítico quando um erro de classificação da classe minoritária tem maior impacto do que os da classe majoritária. Neste contexto, o valor de corte (em inglês, *threshold*) define a probabilidade de uma observação da amostra pertencer a uma classe. Como a estratégia de definição do valor de corte impacta diretamente nas métricas da matriz de confusão, sua definição depende do peso dado para as medidas de precisão e sensibilidade. Se o problema de classificação for mais sensível para os falsos negativos, o valor de

corte pode ser definido para um valor menor que 50%, tornando-se mais sensível e menos preciso. De forma análoga, se o valor de corte definido for maior que 50%, o modelo torna-se mais preciso e menos sensível.

3.1.4 Viés e Variância

O viés representa o erro (resíduos) ao aproximar valores observados de uma medida de interesse por uma função polinomial, e a variância indica a flexibilidade do modelo de aprendizado de máquina. Quanto mais simples o modelo, maior o viés e menor a variabilidade. Modelos com baixa variabilidade indicam um comportamento previsível dos valores observados em amostras distintas. Quanto mais ajustado o modelo, menor o viés, porém maior a variabilidade. Modelos com alta variabilidade não apresentam resultados congruentes nos valores observados para amostras distintas entre treino e teste. Este comportamento é conhecido como o *trade-off* viés-variância.

3.1.5 A Perda Logarítmica

A perda logarítmica (em inglês, *log loss*) é uma função de perda bastante utilizada para medir o desempenho de modelos de ML em problemas de classificação cujos parâmetros de entrada indicam probabilidades de previsões entre 0 e 1. Portanto, quanto menor o *log loss* (mais próximo de zero), melhor o desempenho do classificador. A perda logarítmica aumenta à medida que a probabilidade prevista diverge da classificação real. Em problemas de classificação binária, a função de perda *log loss* para uma amostra é dada por:

$$\text{Logloss}(y, p) = -(y \log(p) + (1 - y) \log(1 - p)), \quad (3.5)$$

onde y indica os valores possíveis reais da classe alvo, para $y \in \{0, 1\}$ e p indica a estimativa da probabilidade no intervalo entre 0 e 1.

A função de perda *log loss* é comumente denominada de perda logística ou perda de entropia cruzada. Para que esta métrica esteja consistente com as outras métricas de desempenho do classificador em que, geralmente, quanto maior o valor, melhor a métrica, multiplica-se o *log loss* por -1 para manter o mesmo padrão. O *log loss* negativado (em inglês, *negative log loss*) é definido por: $\text{negative log loss} = -\text{log loss}$.

3.1.6 Árvore de Decisão

A árvore de decisão é uma estrutura de dados de fácil entendimento que permite representar modelos de classificação e previsão de dados, sendo bastante aplicada ao aprendizado de máquina supervisionado [44]. É composta por um nó inicial (conhecido também como raiz), nós internos, ramos e folhas. A cada conjunto de nós internos uma camada de nível é adicionada à árvore. Os nós internos possuem tanto ramos que chegam a eles quanto ramos que partem deles. As folhas possuem apenas ramificações que chegam a elas e nenhum ramo partindo delas. A Figura 3.3 ilustra um exemplo de árvore de decisão simplificado para o naufrágio do Titanic com $x_1, \dots, x_n$ atributos: 'Sexo', 'Criança' e 'Primeira classe'. Cada percurso da árvore representa uma regra de negócio, e em cada nó de decisão há um teste para classificar um atributo (ou característica). Os ramos representam os possíveis valores de cada atributo e as folhas representam os resultados da classe alvo, indicando se o passageiro sobreviveu ou não ao naufrágio.

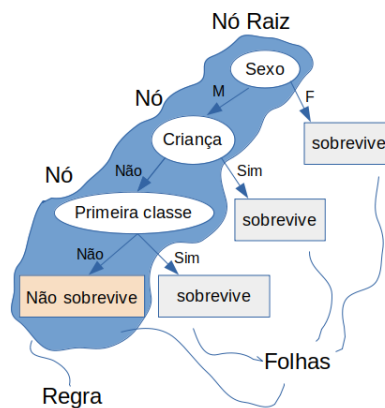


Figura 3.3: Ilustração de uma árvore de decisão

O algoritmo começa pelo topo, ou raiz, e vai se ramificando progressivamente até a obtenção de uma classificação final (classe alvo) nas folhas, denominada de rótulo. Portanto, uma árvore de decisão pode ser definida como um modelo preditor, $h : X \rightarrow Y$, que prediz o rótulo associado a uma instância $x \in X$ percorrendo o caminho de um nó raiz (nó de maior nível hierárquico) à folha. Para problemas de classificação binária, Y representa a classe alvo e pode ser definida como: $Y = \{0, 1\}$. Uma folha é um nó que não possui sucessores e contém um rótulo específico, caracterizada por uma classe ou uma distribuição de probabilidades. Em cada percurso da árvore, no caminho da raiz para a folha, o nó sucessor é escolhido com base na partição das características (atributos) de x ou por um conjunto predefinido de regras [60]. Para problemas de classificação binária, os critérios de partição mais conhecidos são baseados na entropia e índice Gini.

Um problema bastante comum no treinamento de árvores de decisão é denominado

sobreajuste (em inglês, *overfitting*). Este problema ocorre quando o modelo se ajusta muito bem aos dados de treino, mas se mostra ineficaz para prever novos resultados. Para prevenir o sobreajuste em problemas de classificação, pode-se aplicar a estratégia de poda [8]. Esta estratégia consiste na remoção de partes da árvore, diminuindo sua complexidade para promover uma melhor generalização. Em geral, os nós selecionados para remoção contribuem pouco para a classificação das instâncias da árvore [25].

3.1.7 Entropia, Ganho de Informação e Índice Gini

No contexto das árvores de decisão, a entropia é utilizada para medir o quão diferentes/iguais são os elementos de um conjunto de dados de treinamento, isto é, estimar a aleatoriedade da classe alvo em uma amostra. Esta medida consiste em um número entre 0 e 1, onde o valor igual a 0 corresponde à completa igualdade, ou seja, todos os dados da amostra pertencem à mesma classe (nó puro), e o valor igual a 1 corresponde à completa desigualdade entre os dados da amostra (nó impuro). Neste caso, o conjunto de dados contém um número igual de elementos de ambas as classes [44].

Dado um conjunto de entrada S de uma amostra contendo c classes distintas, a entropia de S é dada pela Equação 3.6, onde p_i é a proporção de dados em S que pertencem à classe i :

$$\text{Entropia}(S) = \sum_{i=1}^c -p_i \log_2 p_i. \quad (3.6)$$

Para um problema de classificação onde a classe alvo é binária, a entropia de S pode ser calculada usando a fórmula da Equação 3.7, onde p e q indicam as probabilidades de ocorrência ou não de um evento: por exemplo, passageiro sobreviveu ao Titanic (classe positiva) ou passageiro não sobreviveu ao Titanic (classe negativa).

$$\text{Entropia}(S) = -p \log_2 p - q \log_2 q. \quad (3.7)$$

A entropia atinge o ponto máximo em 0,5 quando $p = q$. Exemplificando, a Tabela 3.1 ilustra um exemplo de conjunto de dados S , representando os passageiros do Titanic. Com base no treinamento dos dados em S , o classificador irá determinar se o passageiro sobreviveu ou não ao naufrágio. p representa a proporção de sobreviventes e q a proporção de não sobreviventes. Os seguintes valores foram definidos: $|S| = 14$ (total de passageiros), $p = 9/14$ e $q = 5/14$. Neste exemplo, a entropia é calculada pela Equação 3.8:

$$\text{Entropia}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0,939. \quad (3.8)$$

sexo	criança	primeira classe	sobreviveu
M	Não	Não	Não
M	Não	Não	Não
M	Não	Não	Não
M	Não	Não	Não
M	Não	Não	Não
F	Não	Não	Sim
F	Não	Não	Sim
M	Sim	Não	Sim
M	Sim	Não	Sim
M	Sim	Não	Sim
M	Não	Sim	Sim
M	Não	Sim	Sim
F	Sim	Não	Sim
F	Sim	Sim	Sim

Tabela 3.1: Exemplo de um conjunto de dados S para treinamento de uma árvore de decisão

Uma abordagem comum para geração de uma árvore de decisão é utilizar a entropia da classe alvo e o ganho de informação das variáveis preditivas para obter a efetividade em classificar os dados de treinamento [44]. Quanto menor a entropia, mais uniforme e homogênea está a amostra de treinamento, e maior é o ganho de informação. Dentre as possíveis aplicações, é comum sua utilização como critério para a seleção do nó raiz e a cada passo da construção da árvore, de forma a selecionar a variável preditiva para o nó corrente, denominada de característica, com maior influência na classificação da classe alvo. Portanto, pode-se definir o $\text{Ganho}(S, C)$ como a medida do ganho de informação de uma característica C selecionada para a partição de um conjunto de dados S , conforme a Equação 3.9. $\text{Valores}(C)$ representa os possíveis valores de C , e S_v representa o subconjunto de S onde a característica C possui valor v . $\text{Entropia}(S)$ é a entropia do conjunto de dados S , e o segundo termo da equação é a soma das entropias de cada valor presente na característica C . O valor $\frac{|S_v|}{|S|}$ é a proporção da frequência de v em relação ao conjunto todo, e $\text{Entropia}(S_v)$ é a entropia do subconjunto que contém esses valores para cada característica.

$$\text{Ganho}(S, C) = \text{Entropia}(S) - \sum_{v \in \text{Valores}(C)} \frac{|S_v|}{|S|} * \text{Entropia}(S_v). \quad (3.9)$$

Outra medida bastante conhecida para medir a desigualdade de um conjunto de dados é o índice Gini [20]. Este índice é bastante similar ao ganho de informação e mede o grau de heterogeneidade dos dados, critério importante para a seleção das características que compõem cada nó da árvore de decisão de forma a produzir maior redução de impureza. A fórmula do índice Gini, aplicada em cada nó, é dada pela Equação 3.10, onde p_i é a proporção de elementos de uma amostra de cada classe, em cada nó, e c é o número de classes. O nó é considerado puro quando o índice Gini é igual a 0. De forma análoga, é considerado impuro quando este valor aproxima-se de 1.

$$\text{Indice_Gini} = 1 - \sum_{i=1}^c (p_i)^2. \quad (3.10)$$

3.1.8 O Algoritmo de Floresta Aleatória

Floresta Aleatória (em inglês, *Random Forest* - RF) é um algoritmo de aprendizado supervisionado introduzido por Breiman [8], que combina a geração de múltiplas árvores de decisão para obter um resultado de previsão com maior acurácia, podendo ser utilizado tanto em problemas de regressão como de classificação. Ao contrário da árvore de decisão que busca a melhor característica na partição dos nós, apenas um subconjunto aleatório de características é considerado. Como árvores mais enxutas são geradas a partir de tais subconjuntos, há uma maior diversidade de subárvores combinadas no treinamento do modelo.

No primeiro passo do algoritmo, selecionam-se aleatoriamente amostras do conjunto de treino de tamanho N . Por meio do *Bootstrap*, que é um método de reamostragem em que as amostras selecionadas podem estar repetidas a cada seleção, a primeira árvore de decisão é gerada. Para definir o primeiro nó da árvore de decisão (nó raiz), escolhe-se a característica que leva a uma partição com a menor entropia/índice Gini em relação às outras partições possíveis. Portanto, a variável com o menor índice de impureza é escolhida para compor a primeira divisão nos dados, levando ao melhor ganho de informação. À medida que aprofunda-se na árvore, o nível de incerteza diminui, levando a uma melhor classificação da classe alvo.

No algoritmo de RF, a seleção da variável em cada nó ou partição (em inglês, *split*) não acontece considerando-se todas as variáveis disponíveis. Ao invés de procurar pela melhor característica para cada partição, busca-se a melhor característica dentre uma subamostra de características. Portanto, dado um conjunto de dados com M variáveis, um número $m \ll M$ de variáveis são selecionadas aleatoriamente em cada nó. Para o primeiro

nó, o algoritmo sorteia duas ou mais variáveis. Para a escolha da variável do próximo nó, novamente são escolhidas duas ou mais variáveis, excluindo-se as já selecionadas anteriormente, e o processo de escolha se repete. A árvore é gerada sucessivamente até o último nó. A quantidade de variáveis selecionadas em cada partição é definida previamente na criação do modelo, mantendo-se constante ao longo do processo de geração das árvores. Esta diversidade na geração das árvores é uma das principais características que contribuem no bom desempenho dos modelos de RF.

A Figura 3.4 ilustra o funcionamento da RF para um problema de classificação binária. Pode-se observar um conjunto de árvores de decisão aleatórias representadas pelas cores marrom, azul, amarela e vermelha. Provavelmente, essas árvores são distintas pois a seleção das amostras e das características de cada árvore ocorreu de maneira aleatória. O modelo de RF é treinado e os resultados das árvores são combinados para encontrar o resultado final da predição, que é o resultado da seleção que obteve a maioria dos votos. No exemplo da Figura 3.4, a maioria das árvores de decisão resultou na previsão igual a “SIM” para a classe alvo, sendo 3 votos para “SIM” x 1 voto para “NÃO”.

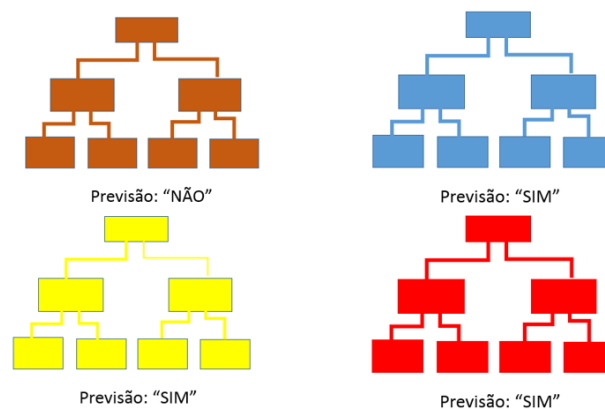


Figura 3.4: Funcionamento da Floresta Aleatória

A técnica de RF é bastante poderosa pois, como são geradas muitas árvores, uma seleção ruim para o nó raiz não compromete tanto o resultado final do classificador. Entretanto, é importante enfatizar que 100% de acerto sobre os dados de treino não garante um baixo erro de classificação na previsão sobre os dados de teste. Outra vantagem dos modelos de RF é que, em geral, não sofrem de *overfitting*. Como um subconjunto de características é selecionado de forma aleatória, e árvores menores são geradas, as chances de *overfitting* diminuem no resultado final da predição.

Em geral, quanto mais árvores são geradas, maior é a diversidade e melhor é o resultado alcançado do classificador. As árvores são geradas até o momento em que a geração de uma nova árvore deixa de agregar melhora significativa no desempenho do modelo.

Assim, o resultado final é obtido a partir da contribuição individual de cada árvore. Para problemas de regressão, calcula-se a média dos valores previstos, e para problemas de classificação, o resultado que mais vezes se repete será o escolhido. É importante considerar que, quanto mais árvores são geradas, maior é o tempo de criação do modelo, elevando o custo computacional. Porém, um custo computacional bem controlado também pode implicar em resultados melhores e mais robustos.

Para a condução desta pesquisa, o algoritmo de RF foi inicialmente selecionado por apresentar uma série de características aderentes ao problema de modelagem preditiva de classificação binária. Podemos citar algumas vantagens do algoritmo de RF:

- Capacidade de modelar padrões não lineares nos dados.
- Escalabilidade (em geral, bom desempenho para grande conjunto de dados).
- Modelos, em geral, obtidos com ótimo desempenho. A aleatoriedade no momento da partição dos nós otimiza a diversidade de subconjuntos. Assim, o resultado final representa a seleção da melhor característica em um subconjunto aleatório de características.
- Facilidade de se medir a importância de cada característica, por meio da análise da impureza geral da floresta. As variáveis que reduzem esta impureza são consideradas as mais importantes.
- Não requer uma configuração precisa dos hiperparâmetros. Em geral, os hiperparâmetros com valores padrões já permitem um bom resultado para o classificador.

3.1.9 O Algoritmo XGBoost

O algoritmo XGBoost (em inglês, *eXtreme Gradient Boosting*) foi desenvolvido como um projeto de pesquisa na Universidade de Washington, e seu trabalho foi apresentado por seus criadores Tianqi Chen e Carlos Guestrin [12] na Conferência SIGKDD¹ em 2016. Desde sua introdução, além de ter sido creditado por ter ganho numerosas competições na plataforma Kaggle², este algoritmo também vem sendo aplicado em vários segmentos da indústria: diagnóstico médico [9], crédito imobiliário [42], mercado financeiro [49], prevenção à fraude e compliance [51], entre outros. Como resultado, há atualmente uma forte comunidade de cientistas de dados contribuindo para a contínua evolução do código. Dentre as principais características do XGBoost, pode-se destacar:

¹<https://www.kdd.org/kdd2016/>

²<https://www.kaggle.com/>

- A capacidade de resolver problemas de aprendizado supervisionado como classificação e regressão.
- O bom desempenho computacional combinando técnicas de otimização de software e hardware (processamento paralelo) para produzir resultados superiores com menos recursos de computação e em menos tempo de execução em relação a outros algoritmos.
- A minimização de *overfitting* nos dados de treino.
- A facilidade de uso em diferentes linguagens de programação, como C++, Python, R, Java, Scala e Julia.

O algoritmo XGBoost é baseado na técnica de Gradient Boosting. Esta técnica treina novos modelos diretamente nos erros gerados em cada iteração, também denominados de resíduos. Em suma, os novos modelos têm como objetivo prever e corrigir o erro (resíduo) encontrado nos modelos anteriores. Por consequência, a predição final é obtida somando-se a predição de cada classificador base (comitê), por meio do empilhamento das árvores geradas nas iterações anteriores. A cada passo, os valores encontrados para os resíduos diminuem até que não seja mais possível diminuir a função de perda. O resultado final é a combinação das predições dos modelos individuais de cada iteração. Este método de aprendizado de máquina que agrega o resultado de vários modelos para se chegar a um resultado final também é conhecido como método *ensemble*.

O algoritmo Gradient Boosting funciona da seguinte forma: o primeiro modelo faz uma aproximação bem simples da predição para obter os erros residuais (valor observado - valor previsto); após, treinam-se mais modelos nos erros residuais para prever o erro encontrado no primeiro modelo. Dessa forma, quando as predições de cada modelo são somadas para geração da predição final, uma versão mais assertiva da primeira predição é obtida. Repete-se esse processo por várias iterações, gerando-se novas árvores e atualizando cada amostra com a nova previsão, baseada na previsão anterior mais o valor do resíduo encontrado na árvore de decisão recém gerada. Assim, obtêm-se erros residuais cada vez menores a cada iteração. O critério de parada é definido quando não for mais possível minimizar o somatório dos ruídos que definem a função de perda. Desta forma, a função de perda para o algoritmo de Gradient Boosting avalia o quanto o modelo consegue acertar na previsão. A função de perda mais comum para a regressão é apresentada na Equação 3.11. O cálculo se dá pela soma dos n desvios (resíduos) ao quadrado, onde y_i é o valor observado real, e $\hat{y}_i$ é o valor da previsão estimado pelo classificador que corresponde a cada uma das n observações $(y_1, \dots, y_n)$, para $i = 1, \dots, n$:

$$L(y_i, \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (3.11)$$

A função de perda mais comum para problemas de classificação binária é dada pela Equação 3.12:

$$L(y_i, \hat{y}_i) = -[y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)]. \quad (3.12)$$

Para diminuir a sensibilidade da previsão nas observações, pode-se adicionar um termo de regularização nos dados de treino como estratégia para minimizar a variância nos dados de teste. Portanto, este parâmetro tem como objetivo gerar viés no modelo, ou seja, produzir um ruído para minimizar a variância nos dados de teste, e consequentemente as chances de *overfitting*.

Os resíduos representam os desvios de cada folha entre valor observado e valor estimado. A partir dos resíduos, pode-se auferir a qualidade da folha por meio do cálculo do grau de similaridade. Seja λ um parâmetro que representa um valor escalar entre 0 e 1. Em problemas de regressão, o grau de similaridade em cada nó da árvore é calculado pela Equação 3.13, sendo num.resíduos_i o número de resíduos em cada folha i :

$$\text{GrauSimilaridade} = \frac{\sum (\text{resíduos}_i)^2}{(\text{num.resíduos}_i + \lambda)}. \quad (3.13)$$

Em problemas de classificação, o grau de similaridade é dado pela Equação 3.14, onde prev_i é a probabilidade estimada no passo anterior de cada observação. A cada passo, são geradas novas árvores a partir dos resíduos calculados. Quando $\lambda > 0$, o grau de similaridade diminui. O algoritmo é interrompido quando resíduos mínimos são encontrados ou quando um número máximo de árvores é alcançado.

$$\text{GrauSimilaridade} = \frac{\sum (\text{resíduos}_i)^2}{(\sum [\text{prev}_i * (1 - \text{prev}_i) + \lambda])}. \quad (3.14)$$

Com base nas funções de perda definidas nas Equações 3.11 e 3.12, o algoritmo XGBoost utiliza o resultado da função de perda de regressão ou classificação com o objetivo de minimizar o somatório dos erros encontrados a cada geração de árvores. A função de perda do algoritmo XGBoost pode ser definida conforme a Equação 3.15, onde y_i cor-

responde aos valores observados, $\hat{y}_i$ corresponde aos valores previstos, n corresponde ao número de observações, e λ corresponde ao valor escalar introduzido na equação para minimizar o valor de saída “O” calculado para cada folha. O objetivo da função de perda é encontrar o valor de saída “O” para cada folha da árvore de forma a minimizar o resultado final da equação.

$$F_{\text{perda}} = \left[\sum_{n=1}^n L(y_i, \hat{y}_i) \right] + 1/2\lambda * O_{\text{value}}^2. \quad (3.15)$$

A parte da equação definida por $1/2\lambda * O_{\text{value}}^2$ também é denominada termo de regularização. Na primeira iteração ($i = 1$) do cálculo da função de perda, o valor $\hat{y}_i$ é definido como $y_0 + O_{\text{value}}$. Assim, a equação da função de perda para cada árvore pode ser aproximada pela Equação 3.16:

$$L(y_i, \hat{y}_i) \approx L(y_i, \hat{y}_i + O_{\text{value}}) \approx L(y_i, \hat{y}_i) + gO_{\text{value}} + 1/2 * hO_{\text{value}}^2. \quad (3.16)$$

A variável g corresponde ao cálculo da primeira derivada da função de perda e h corresponde a segunda derivada da função de perda (*Hessian*). Se o valor de h é baixo, significa que a folha já está “pura” o suficiente e não deve ser mais particionada.

Assim, a cada geração de uma nova árvore, computa-se o valor de saída O_{value} de cada folha por meio do cálculo da primeira derivada (gradiente) g_i , e calcula-se o grau de similaridade de cada folha, que é definido pelo cálculo da segunda derivada (*Hessian*) h_i da função de perda mais o termo de regularização. Como a função de erro para uma folha é quadrática (função polinomial de segundo grau), o mínimo é definido como o ponto de inflexão da parábola, em que a primeira derivada é igual a zero. Portanto, como o objetivo é minimizar o somatório das funções de perda geradas para cada árvore, infere-se o resultado das derivadas igualando a Equação 3.17 a zero:

$$\frac{d}{dO_{\text{value}}} (g_1 + \dots + g_n)O_{\text{value}} + \frac{1}{2}(h_1 + \dots + h_n + \lambda)O_{\text{value}}^2 = 0. \quad (3.17)$$

Desta forma, calcula-se a Equação 3.17 para encontrar o valor de saída O_{value} , onde $g_1 + \dots + g_n$ corresponde à soma dos resíduos e $h_1 + \dots + h_n$ corresponde ao total de resíduos em cada folha. Assim, o valor de saída O_{value} de cada folha é dado por:

$$O_{\text{value}} = \frac{-(g_1 + \dots + g_n)}{(h_1 + \dots + h_n + \lambda)}. \quad (3.18)$$

Simplificando a Equação 3.18, O_{value} pode ser definido pela Equação 3.19:

$$O_{\text{value}} = \frac{\text{Soma dos resíduos}}{(\text{Total de Resíduos} + \lambda)}. \quad (3.19)$$

Ou seja, para $\lambda = 0$, o valor de saída O_{value} é a média dos resíduos.

A função de perda encontrada é utilizada para avaliar o ganho de informação a cada divisão de cada nova árvore, assim como a entropia ou o índice Gini são utilizados tradicionalmente na geração das árvores de decisão. Portanto, o $\text{Ganho}_{\text{folha}}$ define como os resíduos devem ser agrupados e divididos da melhor forma possível. A partir da similaridade da esquerda, similaridade da direita e similaridade da raiz, o cálculo do ganho de cada folha é definido pela Equação 3.20:

$$\text{Ganho}_{\text{folha}} = \text{Similaridade}_{\text{esq}} + \text{Similaridade}_{\text{dir}} - \text{Similaridade}_{\text{raiz}}. \quad (3.20)$$

O $\text{Ganho}_{\text{folha}}$ de maior valor é escolhido. Quanto menor são os valores de similaridade encontrados, mais facilmente realiza-se a poda das árvores e, consequentemente, maior são as chances de evitar o *overfitting*. Considerando o elevado custo computacional para calcular os resíduos em cada folha, o artigo original [12] também propõe algumas estratégias de aproximação para maximizar o $\text{Ganho}_{\text{folha}}$.

3.1.10 Busca em Grade para a Seleção de Hiperparâmetros

Para que o modelo de ML tenha capacidade de generalização, um conjunto de valores de hiperparâmetros deve ser escolhido de forma a produzir um modelo que minimize uma função de perda predefinida, de maneira reprodutível e otimizada [13]. Desta forma, a estratégia de seleção de hiperparâmetros denominada busca em grade (em inglês, *grid search*) tem como objetivo encontrar um conjunto ótimo de hiperparâmetros que resulte no melhor desempenho para a(s) medida(s) de interesse. Um bom desempenho pode significar conceitos diferentes. Por exemplo, pode retornar o modelo mais preciso para problemas de classificação, enquanto que para um problema de regressão pode retornar o RMSE (raiz quadrada média do erro) mais baixo. A busca em grade pode ser aplicada,

por exemplo, por duas técnicas:

- Busca cartesiana (em inglês, *cartesian grid search*): Pesquisa exaustivamente todas as combinações possíveis para os hiperparâmetros informados. Este método é o mais indicado quando o número de possíveis combinações for relativamente baixo para o custo computacional disponível.
- Busca aleatória (em inglês, *random grid search*): Este método processa uma combinação aleatória de hiperparâmetros amostrada uniformemente a partir do conjunto de todas as combinações possíveis de hiperparâmetros. Ao contrário da técnica de busca cartesiana, onde é necessário testar exaustivamente todas as combinações possíveis por força bruta, somente um subconjunto de combinações aleatórias é testado. Este método também permite definir um critério de parada para interromper a pesquisa aleatória dos hiperparâmetros. A busca aleatória intercala uma combinação aleatória dos valores e interrompe o processamento quando um determinado nível de melhoria tenha sido alcançado, ou um determinado tempo tenha sido excedido, ou uma determinada quantidade de modelos tenha sido executada (ou uma combinação destas alternativas tenha sido satisfeita). Embora esta técnica provavelmente não encontre o melhor modelo, normalmente traz bons resultados. Se o número de possíveis combinações for relevante, considerando o custo computacional disponível, este é o método mais indicado.

Assim como o *trade-off* entre viés e variância, há um *trade-off* clássico entre a precisão dos resultados e o custo computacional envolvido. Ambos devem ser considerados na definição da técnica para seleção dos hiperparâmetros. Embora a busca cartesiana seja o método que forneça maior garantia para encontrar a melhor combinação de hiperparâmetros, muitas vezes envolve um custo computacional proibitivo. Por outro lado, a busca aleatória tem como vantagem diminuir o tempo computacional, mas nem sempre fornece os melhores resultados.

3.1.11 Validação Cruzada (em inglês, *Cross-Validatiton*, CV)

A validação cruzada é uma técnica bastante conhecida para estimativa do erro preditivo de um modelo [19] e amplamente utilizada em problemas de regressão e classificação na escolha dos hiperparâmetros, de forma a minimizar a estimativa de erro. Apesar de ser uma técnica bastante popular e consistente para a seleção de modelos, a real capacidade preditiva sobre os dados de teste vem sendo discutida na literatura [69], e trabalhos vêm sendo conduzidos com o propósito de ajustar o viés do estimador sem comprometer a

variabilidade do modelo [18]. A validação cruzada tende a ser otimista quanto a verdadeira estimativa do erro preditivo e este baixo viés é frequentemente correlacionado a uma elevada variabilidade [18, 63, 69].

Em geral, a validação cruzada é aplicada quando o conjunto de teste não é suficientemente grande para estimar a taxa de erro do classificador. Algumas estratégias conhecidas podem ser empregadas para estimar este desempenho, utilizando os dados do conjunto de treino [28, 33]. As próximas seções detalham algumas particularidades destas técnicas. A Seção 3.1.11.1 apresenta a CV por *k-fold* e a Seção 3.1.11.2 apresenta a CV aninhada.

3.1.11.1 Validação Cruzada *k-fold*

Na técnica *k-fold*, um número aleatório de amostras é alocado em k grupos (*folds*) de igual tamanho. Cada *fold* é tratado como um conjunto de teste e o restante dos $k - 1$ *folds* são destinados ao treino. Este processo é repetido k vezes, resultando em k conjuntos de teste distintos, e portanto k taxas de erro [28]. Portanto, a taxa de erro estimada na técnica de validação cruzada *k-fold* é computada pela média dos k erros. Seja a taxa de erro resultante de cada *fold* i , definida como MSE_i , para $i = 1, \dots, k$, a estimativa de erro da validação cruzada *k-fold* é dada pela Equação 3.21:

$$CV_{(k)} = 1/k \sum_{i=1}^k MSE_i. \quad (3.21)$$

Uma quantidade menor de amostras sendo treinadas em cada execução implica em um maior viés e uma menor variância sobre o conjunto de teste [18, 28]. Esta menor variância na técnica por *k-fold* é decorrente da menor sobreposição entre conjuntos de treinamento. Como as estimativas de erro de teste são menos correlacionadas, o valor médio de erro de teste não apresentará tanta variância. Este comportamento é decorrente do *trade-off* entre viés e variância.

É importante destacar que esta técnica tende a ser otimista quanto ao viés calculado para os erros de predição, aparentando ser menor do que seu verdadeiro valor. Tibshirani et al. [63] publicaram uma forma de corrigir a estimativa deste viés a partir da técnica de validação cruzada com 10-*folds* e simularam sua aplicação em diferentes classificadores e algoritmos de aprendizado de máquina. Segundo os autores, a estratégia apresentada permitiu corrigir o viés sobre as taxas de erro, aplicando a correção diretamente sobre cada *fold*. A estratégia se mostrou promissora e com desempenho computacional superior em comparação a outras estratégias, como a CV aninhada, técnica a ser apresentada na próxima seção.

3.1.11.2 Validação Cruzada Aninhada

Em trabalho realizado por Bates et al. [6], os autores apresentaram uma proposta de estimador, denominada validação cruzada aninhada (em inglês, *nested cross validation*), para a predição do erro médio entre conjuntos de dados de treino, sugerindo que o mesmo pode ser útil para a seleção de hiperparâmetros. Segundo os autores, uma das principais contribuições deste estimador é a capacidade de geração de intervalos de confiança de maior cobertura, mais precisos e menos enviesados sobre os erros estimados na previsão. A Figura 3.5 apresenta o esquema adaptado do CV aninhado, que é baseado em dois *loops*: um loop interno para treinamento dos $(k-1)$ -*folds*, e um loop externo para validação do modelo no k -ésimo *fold*, extraído em separado dos dados treinados. Ao final de r repetições, os autores demonstraram que o CV não estima o erro de um modelo específico no conjunto de treinamento observado, mas sim o erro médio sobre distintos conjuntos de treinamento, e isso também é válido para outras estimativas de erro de previsão, como por divisão de dados e *Bootstrap* [6]. É importante destacar que esta técnica requer um processamento computacional custoso, o que pode inviabilizar sua aplicação.

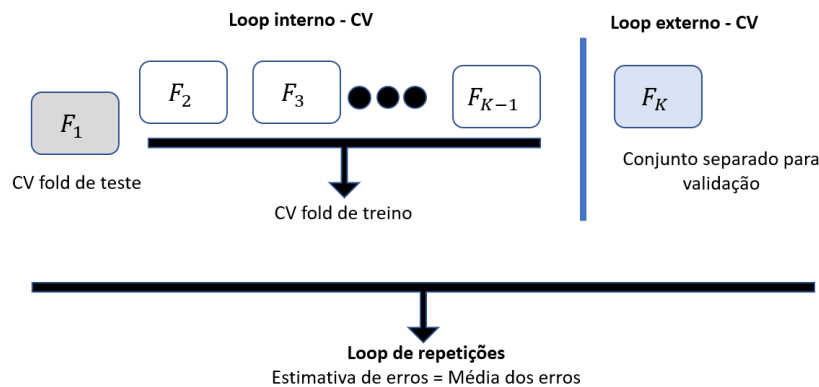


Figura 3.5: Funcionamento da validação cruzada aninhada. À esquerda da linha vertical, é exibido o esquema de CV tradicional: no loop interno, o treinamento dos $(k-1)$ -*folds* (fundo branco) e o teste sendo aplicado em cada *fold* (fundo cinza escuro); no loop externo, a validação é aplicada sobre a amostra (fundo azul) separada do treino. Ao final, a estimativa de erros é calculada com base na média de erros em r repetições, $k-1$ loops internos e k loops externos.

Fonte: Adaptada de Bates et al. [6]

3.1.12 Interpretação Local dos Dados

Os modelos de ML são frequentemente considerados “caixas pretas”. Modelos de ML, como *Random Forests* (RF), Redes Neurais Artificiais (ANN), entre outros, tipicamente apresentam boa precisão para prever fenômenos observados. Infelizmente, uma maior precisão vem, muitas vezes, à custa da interpretabilidade. A necessidade da inter-

pretabilidade significa que, para determinados problemas, não é suficiente obter a previsão (o quê). O modelo também deve explicar como chegou à previsão (o porquê) [16]. As seguintes razões impulsionam a demanda por interpretabilidade e explicações: entendimento do problema, documentação, aceitação e confiança nos resultados encontrados [16].

Portanto, compreender como as variáveis influenciam no resultado é essencial para a confiança nos resultados, principalmente em aplicações de ML na medicina diagnóstica [64]. Os falsos positivos e falsos negativos nesses cenários costumam ter custo muito alto e não devem ser ignorados. As explicações para os casos em que as previsões erraram para um grupo específico da amostra amplia o entendimento sobre o problema e os fatores que mais pesaram em cada classificação. Em geral, as métricas não são capazes de, isoladamente, informar o motivo de cada previsão. Em geral, no mapeamento de artigos da literatura, aplicando ML para detecção de ceratocone (a serem apresentados na Seção 3.2), não foi identificado com clareza como os autores chegaram nos resultados do ponto de vista das características apresentadas. Apesar do bom desempenho das métricas nestes trabalhos, não se sabe com clareza quais são as características mais determinantes em uma amostra específica. A ausência de explicações mais detalhadas sobre os dados e a supervalorização da acurácia corroboram para a suspeita da confiabilidade, principalmente no meio médico. Nestes casos, a métrica de alta acurácia é, na realidade, pior quando analisada em conjunto com as interpretações geradas. Pode-se destacar duas técnicas relevantes para a interpretação local dos dados: A técnica LIME [55] e a técnica de Shapley [61].

3.1.13 Cálculo do Intervalo de Confiança

Na Seção 3.1.2 foi apresentada uma seleção de métricas que podem ser estimadas a partir de uma amostragem. Entretanto, é importante entender quão boa é a estimativa obtida. Em estatística, uma estimativa de intervalo para uma métrica ou parâmetro de uma população caracteriza o intervalo de confiança [48]. Suponha que $X_1, \dots, X_n$ seja uma amostra proveniente de uma distribuição normal. Uma estimativa para o intervalo de confiança é um intervalo em que os extremos (limites inferior e superior) são calculados a partir dos dados da amostra. Quando o desvio padrão da população é conhecido, o cálculo do intervalo de confiança (IC) para uma média populacional é dado pela Equação 3.22:

$$IC = \bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}. \quad (3.22)$$

Portanto, os limites inferior (LI) e superior (LS) são calculados pelas Equações 3.23 e 3.24:

$$LI = \bar{x} - Z^* \frac{\sigma}{\sqrt{n}} \quad (3.23)$$

$$LS = \bar{x} + Z^* \frac{\sigma}{\sqrt{n}}, \quad (3.24)$$

onde $\bar{x}$ é a média da amostra; σ é o desvio padrão da população; n é o tamanho da amostra; Z^* representa o valor padrão da distribuição normal para cada nível de confiança.

A Tabela 3.2 ilustra os níveis de confiança mais comuns para o cálculo do intervalo de confiança:

Nível de Confiança	Valor de Z^*
80%	1,280
90%	1,645
95%	1,960
98%	2,330
99%	2,580

Tabela 3.2: Níveis de confiança mais comuns para o cálculo do intervalo de confiança

3.1.14 *Bootstrap* para o Cálculo da Distribuição dos Estimadores

Em geral, o desvio-padrão e a variância são utilizados para o cálculo dos intervalos de confiança, fixando-se um coeficiente de confiança. Uma alternativa sem validação cruzada para calcular a distribuição da estatística é aplicando a técnica de *Bootstrap*. Esta técnica é baseada no método de reamostragem para aprimorar a estimativa da incerteza sobre uma amostra. Foi introduzida por Efron e Tibshirani [17], e está fundamentada na reamostragem para o cálculo das medidas de incerteza dos estimadores, como, por exemplo, os intervalos de confiança. A ideia básica é tratar a amostra como a própria população que deu origem aos dados, e usar a reamostragem com reposição para gerar amostras *Bootstrap*. A amostragem com reposição obtém uma observação a partir da amostra e, em seguida, retorna a observação de volta na amostra para ser utilizada novamente em uma nova amostragem. O *Bootstrap* tem a vantagem de repetir a experiência a partir de repetidas amostragens dentro de uma única amostra. Quando o *Bootstrap* é aplicado a um problema de previsão, este fornece, em geral, uma versão suavizada da validação cruzada, com variabilidade reduzida, porém com um viés ascendente [18]. Em trabalho realizado

por Efron et al. [18], os autores propuseram um estimador *Bootstrap* com o objetivo de reduzir o viés sem comprometer tanto a variabilidade. Como contribuição da técnica, os autores também citaram a geração de estimativas mais precisas para os erros de previsão e a possibilidade de avaliação da variabilidade dos parâmetros estimados bem como dos erros preditivos.

3.2 Mapeamento de Trabalhos Relacionados

Esta seção apresenta trabalhos da literatura que serviram de base científica para o desenvolvimento desta pesquisa. Em seguida, três destes trabalhos são comentados.

Distintos algoritmos vêm sendo amplamente aplicados na medicina e bioinformática para previsão de diagnósticos [56]. Segundo Oh et al. [50], o aprendizado de máquina tornou-se popular na medicina, dada a sua capacidade de lidar com Big Data e de auxiliar no diagnóstico médico com alta precisão. Segundo Jimenez et al. [30], a detecção de ceratocone é chave para identificação de candidatos à cirurgia refrativa. A Tabela 3.3 apresenta uma amostra de trabalhos da literatura para diagnóstico de ceratocone, aplicando técnicas de ML. Para cada artigo, a primeira linha “Referência” apresenta os autores do artigo e o ano de publicação; a linha “Técnica” apresenta a técnica de aprendizado de máquina selecionada; a linha “Dados de entrada” indica os marcadores selecionados para geração do modelo classificador; e, por fim, a linha “Desempenho” indica o resultado alcançado e a métrica selecionada para a avaliação do resultado.

Tabela 3.3: Amostra de trabalhos para diagnóstico de ceratocone

Referência: Lavric et al. [39] (2019)	
Técnica:	CNN
Dados de entrada:	Topografia das imagens - análise da curvatura por meio das cores da topografia da córnea
Desempenho:	98,3% (AUC)
Referência: Yoo et al. [71] (2019)	
Técnica:	Ensemble [38]: SVM + ANN + RF + Lasso + AdaBoost
Dados de entrada:	Dados clínicos e oftalmo-métricos pré-operatórios
Desempenho:	98,3% (validação interna); 97,2% (validação externa)
Referência: Yousefi et al. [72] (2018)	
Técnica:	PCA (redução dos 420 p/ 8 componentes principais) + tSNE (8 componentes principais p/ 2 <i>eigen</i> -parâmetros) + algoritmo não supervisionado p/ a identificação de <i>clusters</i>

Desempenho:	94,1% (especificidade); 97,7% (sensibilidade)
Dados de entrada:	420 parâmetros de topografia, elevação e paquimetria foram selecionados por meio de imagens adquiridas do sistema SS-1000 CASIA OCT ³
Referência: Ambrósio et al. [2] (2017)	
Técnica:	RF + LOOCV
Desempenho:	99,6% (AUC) ; 98,8% (especificidade); 96,2% (sensibilidade)
Dados de entrada:	Dados tomográficos baseados nas imagens tomográficas de Scheimpflug (Pentacam HR) + biomecânicos (Corvis ST)
Referência: Leão et al. [40] (2019)	
Técnica:	Função discriminante linear (Abordagem-1) e Diagonal quadrática (Abordagem-2, com dados de perfil da espessura horizontal)
Desempenho:	Abordagem-1: 95,4% (AUC); 88,2% (sensibilidade); 97,4% (especificidade). Abordagem-2: 96% (AUC); 89,7% (sensibilidade); 96,4% (especificidade)
Dados de entrada:	Amplitude de deflexão (DA) compensado por IOP e bIOP. Corvis ST (imagens e dados biomecânicos)
Referência: Chastang et al. [11] (2000)	
Técnica:	Árvore de Decisão Binária
Desempenho:	95% (AUC); 88,5% (sensibilidade); 94,9% (especificidade)
Dados de entrada:	Índices topográficos derivados do sistema de videoceratografia EyeSys System 2000 ⁴ e da queratometria ⁵
Referência: Twa et al. [67] (2005)	
Técnica:	Árvore de Decisão + validação cruzada 10-fold
Desempenho:	97% (AUC); 92% (acurácia)
Dados de entrada:	4/36 coeficientes de Zernike ⁶ (<i>inferior elevation, greater sagittal depth, oblique toricity, e trefoil</i>) calculados a partir dos dados de elevação, oriundos de instrumentos de videoceratografia clínicos
Referência: Hidalgo et al. [26] (2016)	
Técnica:	SVM
Desempenho:	95,2% (AUC)
Dados de entrada:	22 parâmetros obtidos dos dados topográficos - Pentacam
Referência: Toutounchian et al. [66] (2012)	
Técnica:	NN com MLP, RBFNN, SVM e árvore de decisão

³<https://www.beye.com/product/casia-cornealanterior-segment-oct-ss-1000>

⁴<http://eyesys.com/products/vista.html>

⁵Instrumento diagnóstico para medir a curvatura da superfície anterior da córnea.

⁶Córneas com ceratocone apresentam porcentagens aumentadas nos coeficientes de Zernike

Desempenho:	92,21% (acurácia)
Dados de entrada:	125 mapas topográficos da córnea (imagens) do Pentacam
Referência: Arbelaez et al. [3] (2012)	
Técnica:	SVM
Desempenho:	Sensibilidade: 96,0% (olhos anormais); 95,0% (olhos com ceratocone); 92,0% (ceratocone subclínico); 97,2% (normal)
Dados de entrada:	Parâmetros da superfície posterior da córnea e espessura da córnea obtidos da câmera Scheimpflug combinadas com a topografia de Plácido

Legenda: NN (*Neural Networks*); SVM (*Support Vector Machine*); MLP (*Multilayer Perceptron*); CNN (*Convolution Neural Networks*); RF (*Random Forest*); LOOCV (*Leave-one-out cross-validation*); PCA (*Principal Component Analysis*); tSNE (*Distributed Stochastic Neighbor Embedding*); RBFNN (*Radial basis function network*); OCT (*Optical coherence tomography*)

3.2.1 Trabalho Comentado - Yoo et al. [71]

Pode-se destacar o trabalho de Yoo et al. [71]. Os autores aplicaram técnicas distintas de algoritmos com validação cruzada (*10-fold*) a partir de dados pré-operativos coletados por instrumentos oftalmológicos de pacientes candidatos à cirurgia refrativa. Os dados foram divididos em um conjunto de treino (80% = 10.561 registros) e validação interna (20% = 2.640 registros). A validação externa foi realizada em 5.279 novos pacientes (40% do total da amostra). O modelo desenvolvido superou estatisticamente outras metodologias, como o PTA e a tabela de Randleman [52], reclassificando corretamente pacientes que desenvolveram ectasia no pós-operatório. Do total de 18.428 pacientes, 8,8% apresentaram contraindicação à cirurgia refrativa. Destes pacientes, 9,5% são do conjunto de treino, 9,5% pacientes são da validação interna e 7,1% pacientes são da validação externa. Para cada algoritmo (SVM, ANN, RF, LASSO e AdaBoost), os modelos foram comparados entre si com base no número de características e nos métodos de seleção. Algumas fragilidades encontradas pelos autores: (i) Os dados do pós-operatório não foram analisados; (ii) O estudo foi aplicado em um único país da Ásia, a Coreia do Sul. Este fato pode influenciar as predisposições genéticas centradas em um único grupo, não sendo possível confirmar a generalização em outros grupos de pacientes; (iii) Os resultados não foram comparados com outros índices oftalmológicos.

3.2.2 Trabalho Comentado - Lavric et al. [39]

Em trabalho realizado por Lavric et al. [39], os autores utilizaram imagens (pixels) da topografia da córnea para desenvolver uma rede neural convolucional com capacidade

de classificar a saída em duas categorias: com ou sem padrões de ceratocone. O algoritmo foi baseado na decomposição das imagens em uma matriz de pixels que formaram a entrada da rede neural. As imagens foram pré-processadas e divididas em três grupos: conjunto de treino, validação e teste. A rede neural foi treinada sobre o treino e a acurácia foi auferida sobre o teste. Dada a dificuldade para treinar um conjunto de dados suficientemente grande, o modelo estocástico SyntEys KTC [57] foi aplicado para geração de dados sintéticos da topografia da córnea. O algoritmo proposto foi aplicado em 1.500 olhos saudáveis e 1.500 olhos com ceratocone. A validação foi realizada sobre o teste, representado por 400 topografias da córnea, sendo 200 topografias de olhos com ceratocone e 200 topografias de olhos normais. Segundo os autores, a principal contribuição do trabalho foi o desenvolvimento de um modelo de alto desempenho para auxiliar no diagnóstico médico de ceratocone.

3.2.3 Trabalho Comentado - Leão et al. [40]

Em trabalho realizado por Leão et al. [40], os autores investigaram a DA (Amplitude de Deflexão) com ou sem a compensação pela IOP⁷ e bIOP⁸, utilizando um método que conjuga funções discriminantes e seleção de atributos. Olhos saudáveis e com ceratocone foram considerados para a pesquisa. Os dados foram capturados a partir das imagens do Corvis ST. O critério de seleção foi a maximização da AUC. O primeiro classificador considerou somente parâmetros biomecânicos relacionados à amplitude de deflação, excluindo-se parâmetros de espessura. O segundo e o terceiro classificador foram baseados na análise dos intervalos da DA compensada por IOP e por bIOP, respectivamente, e o quarto classificador acrescentou informações de espessura da córnea. Segundo os autores, foi possível gerar um classificador baseado apenas em informações biomecânicas com resultado semelhante ao CBI⁹ [68]. Como ponto de fragilidade, os autores citaram a necessidade de complementar os dados com informações sobre a espessura da córnea e com dados clínicos dos pacientes.

3.3 Resumo

Em resumo, este capítulo apresentou os principais conceitos e termos utilizados de aprendizado de máquina aplicados nesta pesquisa. O funcionamento e principais características dos algoritmos de Floresta Aleatória e XGBoost foram detalhados por apresenta-

⁷pressão intra-ocular

⁸*biomechanically-corrected* IOP

⁹Índice Biomecânico Corvis

rem aderência para o problema de classificação binária proposto da pesquisa. A estratégia de busca em grade também foi comentada bem como as técnicas de validação cruzada *k-fold* e *k-fold* aninhada. Em seguida, foram apresentados os conceitos de interpretação local, o cálculo do intervalo de confiança e a técnica de *Bootstrap* para aprimorar a estimativa de incerteza das medidas de interesse. O capítulo finaliza com uma amostra de trabalhos da literatura que aplicaram técnicas de aprendizado de máquina para detecção de ceratocone. Foi apresentado um resumo dos critérios adotados em cada trabalho. Em seguida, as técnicas, fragilidades e resultados de três trabalhos foram comentados.

No Capítulo 4 é apresentada uma proposta de metodologia experimental para detecção de risco de ceratocone utilizando modelos de aprendizado de máquina.

4. Proposta: Uma Nova Metodologia Experimental

Neste capítulo, discute-se uma nova proposta metodológica para seleção de modelos de classificação robustos, considerando o contexto específico do problema de detecção de risco de ceratocone e, conseqüentemente, indicação à cirurgia LASIK.

Diversos modelos de classificação podem ser aplicados a uma mesma amostra de dados e alguns deles podem desempenhar melhor do que outros, porém, não havendo garantias de generalizar os resultados em novas amostras de dados [70]. Desta forma, protocolos de avaliação experimental são necessários para realizar previsões com sucesso a respeito de novos casos. A definição de protocolos foi investigada na literatura, conforme amostra de trabalhos apresentada na Tabela 3.3, a fim de selecionar modelos com maior capacidade de generalização. Estes trabalhos, em sua maioria, adotaram metodologias padrões de seleção de modelos, i.e., busca por melhores hiperparâmetros com base em protocolos experimentais de validação padrão.

Neste trabalho, levanta-se a questão de que as métricas de desempenho geradas a partir das metodologias tradicionais podem apresentar incongruências relativas às distribuições dos resultados, impactando na robustez do modelo em novas amostras de dados. Esta questão pode ser bastante sensível, considerando-se o risco da cirurgia LASIK e do potencial impacto de falsos negativos na saúde dos pacientes. O presente trabalho de dissertação de mestrado propõe-se em apresentar uma metodologia experimental de seleção de modelos que garanta maior segurança nas decisões tomadas pelos médicos com base nos resultados gerados pelos modelos de classificação. Esta proposta apresenta um critério para sistematizar a seleção de hiperparâmetros a partir da introdução do intervalo de confiança dos estimadores, com o objetivo de gerar modelos de menor variância, mais previsíveis, e de comportamento similar em amostras distintas.

4.1 Motivação Geral

A motivação da metodologia proposta visa a questionar as metodologias experimentais tradicionais com o intuito de aumentar a robustez dos modelos de aprendizado supervisionado no contexto do problema de ceratocone. Seu objetivo é assegurar que os resultados encontrados sejam robustos, sendo de fato confiáveis na assertividade das previsões, e generalizáveis para novas amostras de dados.

A distribuição amostral da medida de interesse pode apresentar comportamento distinto entre validação e teste. Como a classe alvo é minoritária para os casos positivos de detecção de risco de ceratocone, o modelo acaba “viciado” nos dados da classe majoritária (olhos saudáveis) em que foi treinado para aprender. Este excesso de ajuste nos dados de validação não é um comportamento adequado e implica na geração de modelos com baixo viés, porém, com alta variância. Em outras palavras, quando o modelo é aplicado sobre uma nova amostra, qualquer variação de comportamento nos dados de entrada que não siga o mesmo padrão em que foram treinados resulta em uma previsão falha. Desta forma, busca-se construir um modelo que seja flexível o suficiente para se adaptar a novas amostras de forma que a distribuição das estimativas seja suficientemente previsível (baixa variância). Como há uma expectativa do intervalo de confiança de 95% abranger o valor médio populacional, pressupõe-se que essa medida apresente o mesmo comportamento em amostras extraídas de populações iguais. Ou seja, espera-se que a média populacional esteja enquadrada em 95% dos experimentos processados para as amostras de validação e teste. Caso este cenário não se concretize, levantam-se questionamentos acerca da robustez do classificador.

A Figura 4.1 exibe uma simulação para a geração de 20 amostras hipotéticas de uma população em dois cenários: IC 95% e IC 80%. Devido à natureza aleatória, os intervalos de confiança não são idênticos. A linha horizontal representa o valor médio populacional desconhecido μ . Para o cenário IC 95%, pode-se afirmar que 95% (19 de 20) das amostras produziram intervalos de confiança que incluem μ , e 5% (1 de 20) das amostras produziram um intervalo de confiança (em vermelho) abaixo da linha horizontal, não incluindo μ . De forma análoga, para o cenário IC 80%, 16 amostras produziram intervalos contendo μ e 4 amostras (em vermelho) não produziram intervalos contendo μ . No cenário IC 80%, pode-se observar que os valores observados em cada amostra variaram consideravelmente. Quanto maior a variabilidade dos dados, menor a precisão e robustez da estimativa e, portanto, maior o desvio padrão e a margem de erro. Já no cenário IC 95%, a variabilidade dos dados é menor e a precisão sobre a estimativa μ é maior, trazendo maior previsibilidade ao classificador.

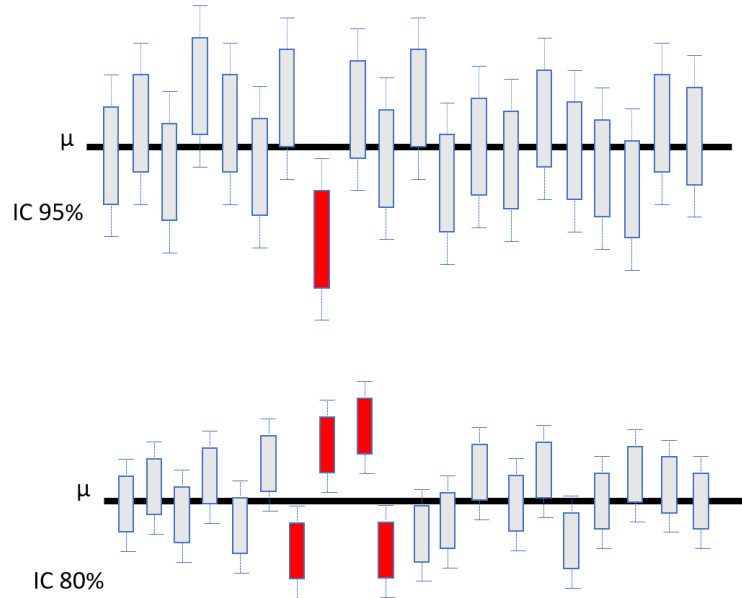


Figura 4.1: Simulação para geração de amostras de uma população. O gráfico superior representa um cenário com intervalos de confiança de 95%. O gráfico inferior representa um cenário com intervalos de confiança de 80%

No exemplo da Figura 4.2, simula-se um processamento sequencial para geração de intervalos de confiança sobre uma amostra com distribuição normal de tamanho $N = 100$, média = 40 e desvio padrão = 2. Repete-se o processamento em 1000 execuções sequenciais e, para cada amostra processada, gera-se o intervalo de confiança de 95% correspondente. Para cada amostra, o valor real se enquadra ou não se enquadra ao IC calculado. 95% dos intervalos de confiança observados englobaram o valor real da média (em preto). Isto significa que aproximadamente 5% dos intervalos de confiança (em vermelho) falharam na cobertura do valor real.

Portanto, o critério proposto não visa somente a otimizar as medidas de interesse mas também garantir que a distribuição amostral seja similar em amostras distintas. Como estratégia para controlar a variância e assegurar medidas de interesse compreendidas em intervalos de confiança de 95%, este critério implica na introdução de um possível viés e sacrifício da medida de interesse. Porém, esta estratégia apresenta a vantagem de garantir resultados compreendidos em um intervalo de confiança de 95% e com a mesma margem de erro se for aplicada em uma nova amostra de dados, o que nos motiva a introduzir o cálculo dos intervalos de confiança como critério para determinar o grau de robustez do modelo.

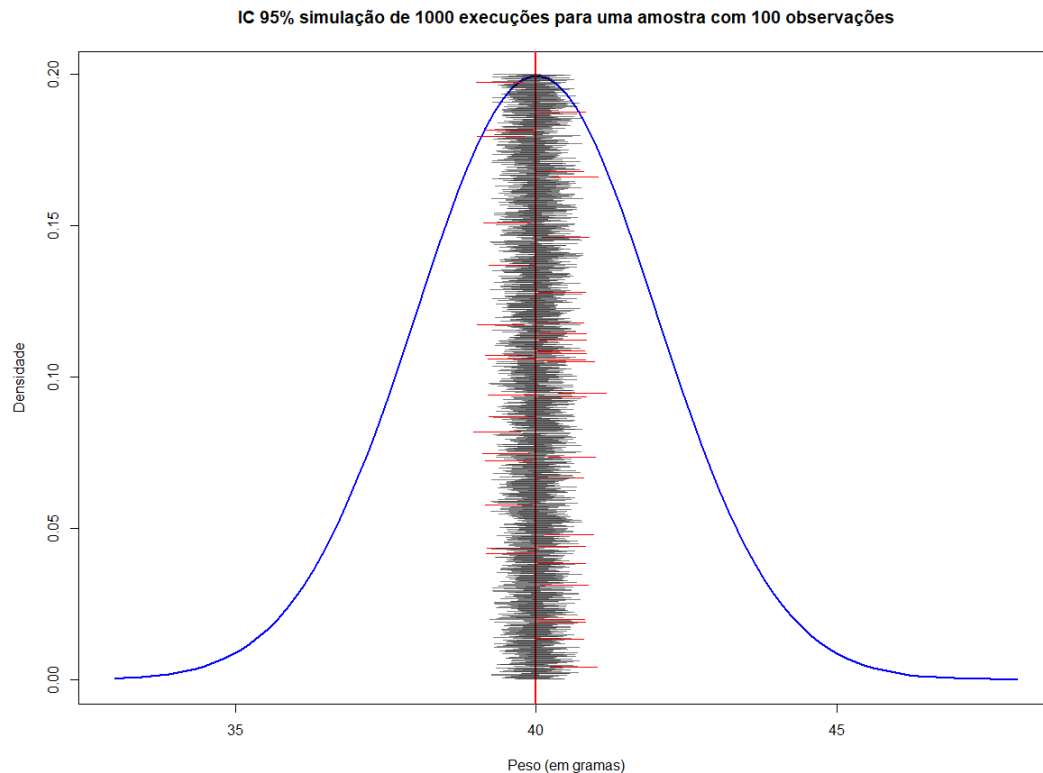


Figura 4.2: Simulação de 1000 execuções para geração de intervalos de confiança de 95% sobre uma amostra gerada por distribuição normal de tamanho $N=100$ observações, média = 40 e desvio padrão = 2. IC com sucesso de cobertura (em preto). IC com falha de cobertura (em vermelho).

4.2 Hipótese da Proposta

Para que modelos de aprendizado supervisionado possam auxiliar no diagnóstico médico, especialmente em uma enfermidade sensível como o ceratocone, torna-se crucial a garantia da robustez deste modelos para sua efetiva aplicação e contribuição na medicina diagnóstica. Porém, conforme mapeamento de trabalhos da literatura da Tabela 3.3, em geral, a metodologia aplicada na etapa de treinamento não apresentou, de forma clara, uma garantia de robustez nos resultados dos experimentos.

A hipótese por trás da proposta desta pesquisa consiste em que modelos preditivos robustos geram necessariamente medidas de interesse estatisticamente iguais se obtidas a partir de amostras de uma mesma população. Ou seja, medidas de erro de classificação computadas a partir de diferentes amostras aleatórias de mesmo tamanho devem seguir uma distribuição de probabilidade. Uma forma de avaliar este comportamento estatístico é a partir do $(1 - \alpha)$ nível de confiança associado a um intervalo de valores da estatística de interesse, onde α é o nível de significância.

É esperado que as distribuições sobre os dados de teste apresentem comportamento similar aos dados de validação e compreendidas dentro do IC de 95% calculado sobre a validação. Quanto maior a amplitude do intervalo de confiança, maior é a confiança (probabilidade) em estimar corretamente a verdadeira classificação do resultado e menor é a variabilidade sobre a estimativa. De forma análoga, quanto mais estreito o intervalo de confiança, maior a variabilidade e a margem de erro sobre a estimativa.

A correta classificação da classe minoritária é cerne para o sucesso do classificador. Nesta linha, a hipótese da proposta é sensível para o problema de detecção de risco de ceratocone, considerando os possíveis impactos do desbalanceamento da classe alvo (3% de olhos com risco de ceratocone) para a geração de modelos robustos. Um modelo que ignore esta desproporcionalidade nos dados não será capaz de diferenciar olhos saudáveis de olhos com risco de ceratocone, mesmo que aparentemente apresente alta acurácia nos resultados¹ e/ou melhores medidas de interesse. Um modelo que apresente um desempenho ligeiramente inferior e mais tolerante para a medida de interesse pode ser muito mais robusto para o problema em questão do que um modelo com desempenho otimizado para a medida de interesse, porém indiferente à classe minoritária.

A Figura 4.3 exibe dois cenários hipotéticos de distribuição dos valores observados da medida de interesse *negative log loss*: IC 95% e IC 60%. No cenário de maior grau de confiança (IC 95%), as distribuições apresentam comportamento similar entre as amostras de treino e teste. As métricas de teste estão contidas dentro do intervalo de confiança de 95% calculado para treino. Neste cenário, há evidência de robustez do modelo. Já no cenário com menor grau de confiança (IC 60%), a distribuição de teste extrapola os limites do intervalo de confiança calculado para treino. Embora o desempenho da medida de interesse para esta distribuição esteja, em média, ligeiramente maior em relação ao cenário IC 95%, não há evidência de robustez para este modelo.

4.3 Linhas Gerais da Proposta

A proposta metodológica deste trabalho foi inspirada na aplicação da técnica *Bootstrap* para aprimorar a incerteza dos estimadores (Seção 3.1.14), no cálculo do intervalo de confiança (Seção 3.1.13), e na técnica de validação cruzada *k-fold* aninhada, proposta por Bates et al. [6] e apresentada na Seção 3.1.11.2. Tem como premissa a seleção de hiperparâmetros que garanta a mesma distribuição dos valores observados da medida de interesse em amostras distintas. Pode ser resumida nos seguintes passos:

¹Paradoxo da acurácia [1]

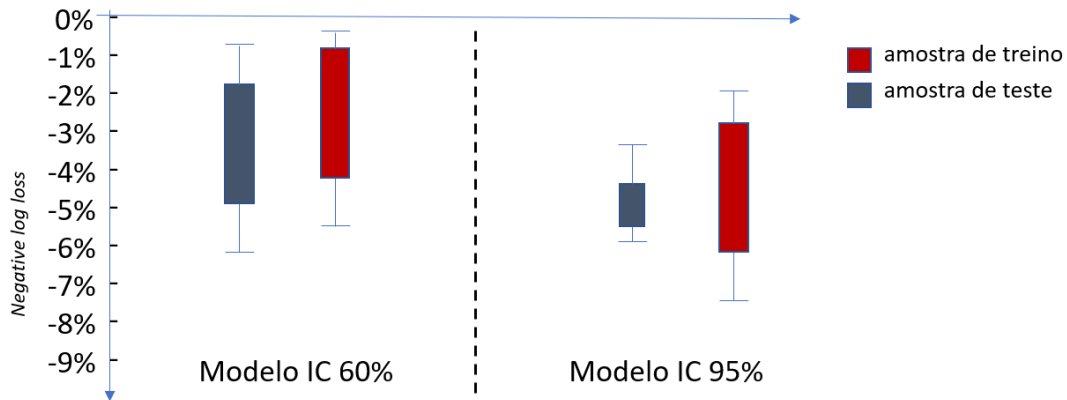


Figura 4.3: Distribuições hipotéticas da medida de interesse para amostras de treino e teste com distintos intervalos de confiança calculados: IC 60% e IC 95%

- Aplica-se a divisão estratificada entre validação e teste sobre o dado original.
- Gera-se um *ranking* de modelos candidatos de ML otimizados pela medida de interesse, treinados por algoritmos de aprendizado supervisionado de classificação binária (Floresta Aleatória/XGBoost) sobre o conjunto de validação, para servir de insumo para o artefato metodológico.
- Ordenam-se os modelos por ordem decrescente de desempenho e seleciona-se o modelo com melhor desempenho pela medida de interesse.
- Aplica-se o modelo candidato de melhor desempenho sobre os dados de validação.
- Geram-se as medidas de interesse sobre os dados de validação, aplicando a técnica de validação cruzada *k-fold* (Seção 3.1.11.1) com R repetições.
- Observa-se a distribuição das medidas de interesse sobre validação e calcula-se o intervalo de confiança de 95% destas amostras (Seção 3.1.13), a partir da média e desvio padrão sobre os *k-folds*.
- Aplica-se a técnica de *Bootstrap* no modelo candidato sobre os dados de teste (Seção 3.1.14), em um total de N execuções.
- Geram-se as N estatísticas de interesse sobre o teste. A medida de interesse é comparada com os limites do IC, e o percentual de amostras de teste que atenderam ao critério restritivo do IC é calculado. Espera-se uma margem de erro menor que 5% nas N execuções.
- Gera-se a matriz de confusão consolidada com a classificação das N execuções *Bootstrap* sobre a amostra de teste.

- Geram-se as distribuições sobre teste e validação. Distribuições não sobrepostas indicam divergência e ausência de robustez, um indicativo de viés nos hiperparâmetros do modelo. De forma análoga, distribuições sobrepostas, com comportamento simétrico entre teste e validação, garantem robustez e previsibilidade do modelo.
- Se o modelo não atender ao critério de restrição, outros modelos do *ranking* são selecionados, com a expectativa de obter um maior percentual de amostras atendendo ao critério restritivo do IC de 95%.
- Modelos dentro do domínio restritivo que satisfaçam ao critério do IC de 95% são considerados válidos. Caso contrário, são descartados e considerados inválidos e sem capacidade de generalização.
- Dentre os modelos válidos, escolhe-se o modelo que obteve o melhor desempenho pela medida de interesse sobre os dados de validação.

A restrição associada ao intervalo de confiança contribui para a redução da variância do modelo. Contudo, isto pode gerar medidas de interesse com desempenho inferior, uma consequência da introdução de viés no modelo. Assim, há uma expectativa dos valores observados sobre o teste estarem estatisticamente próximos daqueles obtidos sobre a validação, e portanto dentro do IC de 95%, em 95% das amostras observadas. Neste contexto, a hipótese é que este protocolo mais estrito de seleção de modelos tende a evitar candidatos enviesados para o conjunto de validação utilizado.

Espera-se que as medidas geradas para validação e teste estejam distribuídas de forma sobreposta. Se este cenário não se concretizar, assume-se a hipótese de que o viés e/ou a variância precisam ser calibrados. Portanto, a escolha de hiperparâmetros que prioriza um menor viés sobre os dados de teste nem sempre garante resultados robustos. Nestes casos, estes modelos ficam excessivamente ajustados aos dados de treinamento, perdendo a capacidade de generalização e de serem efetivamente utilizados além dos dados de treino.

A Figura 4.4 apresenta um esquema das etapas realizadas na proposta metodológica:

1. Observação do fenômeno e captura de amostra da população.
2. Divisão dos dados de forma estratificada entre validação (80%) e teste (20%).
3. Seleção de modelo do ranking de modelos otimizados pela medida de interesse.
4. Treinamento do modelo candidato, aplicando-se a técnica de validação cruzada *k-fold* com R repetições sobre os dados de validação .

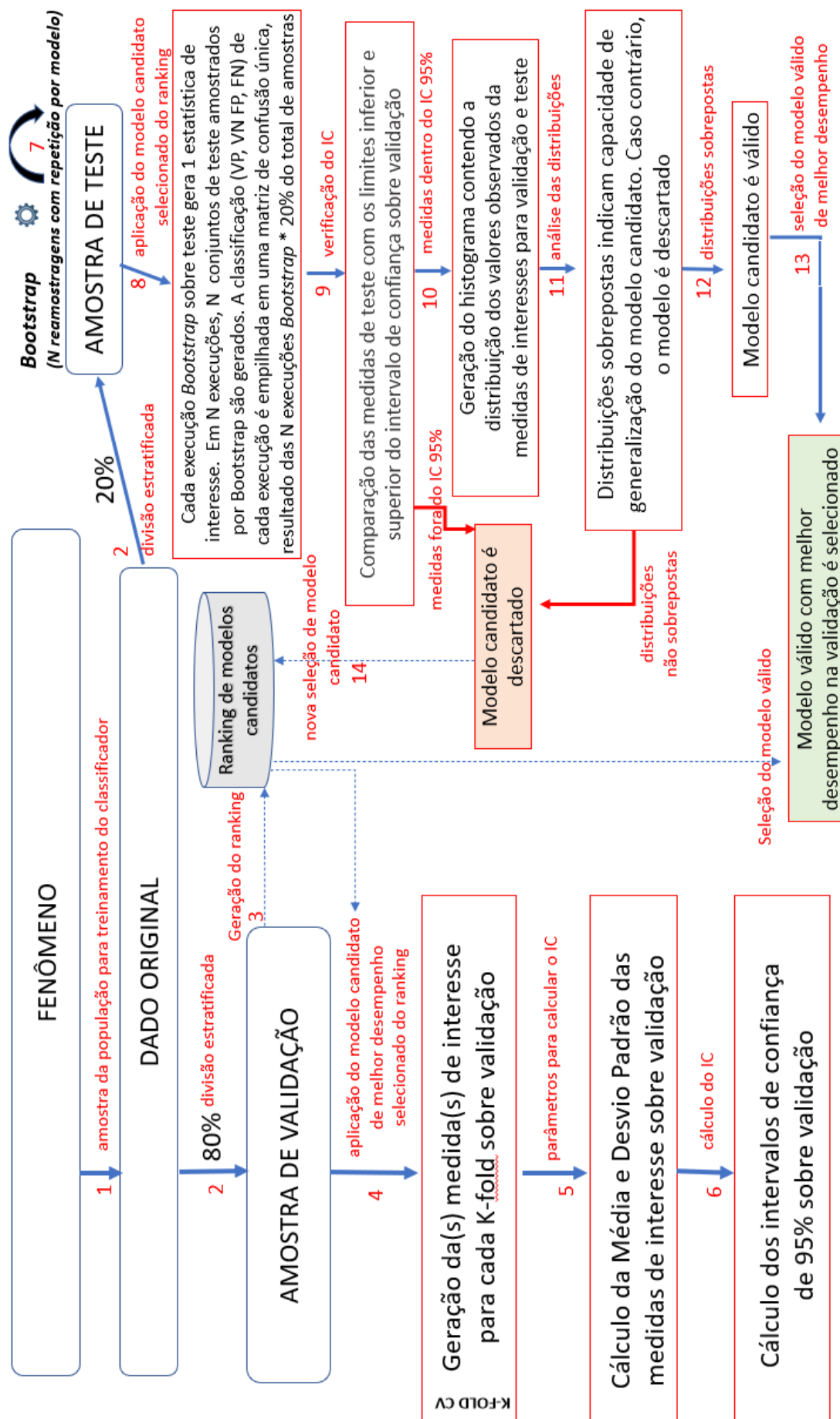


Figura 4.4: Esboço da estratégia metodológica

5. Geração das medidas de interesse em cada *fold* para calcular a média e desvio padrão sobre a amostra de validação.
6. Cálculo do IC de 95% sobre a amostra de validação.
7. Aplicação da técnica de *Bootstrap* sobre a amostra de teste, gerando-se N estatísticas de teste para cada esquema de configuração aplicada na validação cruzada. O *Bootstrap* sobre o teste tem a finalidade de validar o modelo segundo o critério restritivo proposto do IC.
8. Aplicação do modelo candidato sobre a amostra de teste para geração das previsões e captura dos valores observados. A classificação de cada execução *Bootstrap* é armazenada na matriz de confusão consolidada.
9. Valores observados da medida de interesse são comparados com IC de 95%, calculado na etapa 6. Modelo é considerado válido para valores dentro do IC ou inválido para valores fora do IC.
10. Geração dos histogramas sobre os dados de validação e teste.
11. Distribuições não sobrepostas entre validação e teste indicam que o modelo candidato não é válido e deve ser descartado. Neste caso, um novo modelo deve ser selecionado do ranking de modelos sobre a validação.
12. Distribuições sobrepostas indicam que os valores observados sobre os dados de teste estão enquadrados dentro do IC de 95%, em 95% das amostras observadas. Neste caso, o modelo é considerado válido.
13. Nos casos em que houver mais de um modelo válido, seleciona-se o modelo de melhor desempenho na validação.

4.3.1 O Critério Restritivo do Intervalo de Confiança

Esta seção apresenta o critério restritivo introduzido com a aplicação da técnica de validação cruzada *k-fold* para calcular os intervalos de confiança dos valores observados sobre a validação, em conjunto com a técnica de *Bootstrap*, para geração das distribuições sobre o teste. Resultados fora do IC sugerem que os dados treinados estão muito diferentes dos dados testados. Este cenário é contraindicado pois não assegura que os resultados previstos correspondem ao verdadeiro valor da média populacional. Resultados enquadrados dentro do IC de 95%, em 95% das amostras de teste, são considerados robustos.

A Tabela 4.1 apresenta o resultado de uma simulação para o cálculo de intervalos de confiança, conforme descrito na Equação 3.22. Cada linha representa o treinamento do classificador com reamostragem por *Bootstrap*, em um total de N execuções. A coluna “recall” indica as medidas de interesse observadas nos dados de teste. As colunas “recall_mean” e “recall_sd” indicam a média e desvio padrão gerados a partir da técnica de validação cruzada *k-fold*, sendo k igual a 10. As colunas “recall_ic95inf” e “recall_ic95sup” indicam os limites inferior e superior das medidas capturadas sobre os dados de validação, calculados a partir da média e desvio padrão com nível de significância igual a 0,05. Pode-se observar a presença de uma amostra de teste, identificada pela semente (em inglês, *seed*) igual a 4, que não se enquadra dentro do intervalo de confiança definido pelos limites inferior e superior da execução correspondente. Portanto, em um total de 50 (5 execuções x 10 *folds*) amostras de validação, o classificador manteve o “recall” dentro do intervalo de confiança em 80% das amostras.

seed	recall	recall_ic95inf	recall_ic95sup	recall_mean	recall_sd
1	0,76	0,51	0,81	0,66	0,24
2	0,71	0,63	0,79	0,71	0,14
3	0,76	0,62	0,81	0,72	0,15
4	0,81	0,54	0,73	0,64	0,16
5	0,81	0,71	0,91	0,81	0,16

Tabela 4.1: Simulação para cálculo dos intervalos de confiança

4.3.2 Estratégia para Sistematizar a Geração das Estimativas

Esta seção apresenta a estratégia para sistematizar a geração das estimativas do modelo de ML e seleção do modelo robusto de melhor desempenho pela medida de interesse. O Algoritmo 1 de Seleção de Modelo Robusto (SMR) descreve a proposta metodológica. O algoritmo possui os seguintes parâmetros de entrada: conjunto de dados original *A* de tamanho α ; percentual de validação *p_val*; quantidade de *folds* *K*; quantidade de repetições *R*; tamanho do ranking *tam_rank*; total de execuções *Bootstrap* *N*. Como saída, o algoritmo retorna o modelo de ML de melhor desempenho e suas estatísticas correspondentes. Nas duas primeiras linhas é realizada a divisão dos dados de forma estratificada entre validação ($p_val \times \alpha$ dados) e teste ($(1 - p_val) \times \alpha$ dados) gerando, respectivamente, as amostras de validação *X* e de teste *Y*. Na linha 3 inicializa-se o conjunto de modelos de ML Válidos *MLV*, inicialmente vazio. Na linha 4 é executado um procedimento, que será detalhado no próximo capítulo, para criação do ranking de modelos de ML otimizados pela medida de interesse. Esse procedimento retorna um conjunto ordenado que será chamado de Ranking, de tamanho *tam_rank*. Para cada modelo do Ranking (laço das linhas 5-41), serão geradas $R \times K$ medidas de interesse (laços 9-18

e 10-17). Na linha 6 inicializa-se o contador i de medidas de interesse, e, na linha 7, inicializa-se o conjunto de medidas de interesse $MVal_b$ sobre a amostra de validação X . Na linha 8 é definido o modelo de ML ML_b referente ao b -ésimo modelo do ranking. Na linha 11 é definida a amostra de validação excluindo-se o k -fold e, na linha 12 é definida a amostra de validação do k -fold. Em seguida, na linha 13, o modelo é treinado nos $(k - 1)$ -folds e validado no k -ésimo-fold. A i -ésima medida de interesse correspondente ao k -fold na repetição r é armazenada em M_i , na linha 14. Na linha 15 atualiza-se o conjunto de medidas de interesse $MVal_b$ com a i -ésima medida de interesse gerada para o modelo ML_b . O contador i de medidas de interesse é atualizado na linha 16. Ao final do processamento dos K folds em R repetições, calcula-se a média e desvio padrão das medidas de interesse (linhas 19 e 20), os limites inferior (linha 21) e superior (linha 22) do IC de 95%. Na linha 23, inicializa-se a matriz de confusão consolidada $Mat_b = (0, 0, 0, 0)$ como um vetor de quatro posições, em que a primeira posição guarda a quantidade de ocorrências de VP, a segunda posição guarda a quantidade de ocorrências de VN, a terceira guarda a quantidade de ocorrências de FN e a quarta de FP. O conjunto de estatísticas de teste para o b -ésimo modelo é inicializado na linha 24 e o conjunto de medidas de teste válidas para o b -ésimo modelo é inicializado na linha 25. Para cada execução j *Bootstrap* sobre o teste (linha 26), gera-se a amostra *Bootstrap* (linha 27) e aplica-se o b -ésimo modelo de ML (linha 28) sobre Z_j . Para cada conjunto de teste gerado por *Bootstrap*, gera-se a previsão da classe alvo (linhas 29 e 30). Ao final do processamento de cada amostra gerada por *Bootstrap*, calculam-se as medidas VP, VN, FN, FP sobre Z_j e atualiza-se a matriz de confusão consolidada incrementando os valores de cada posição do vetor (linha 32). Em seguida, calcula-se a estatística de teste $Teste_j$ da execução *Bootstrap* corrente (linha 33) e, na linha 34, atualiza-se o conjunto de estatísticas de teste $MTeste_b$ para o b -ésimo modelo, incluindo a estatística de teste $Teste_j$. Se a medida de interesse sobre a amostra *Bootstrap* de teste estiver contida dentro do IC de 95% (linha 35), atualiza-se também o conjunto de medidas de teste válidas $MValidos_b$ para o b -ésimo modelo. Ao final das execuções *Bootstrap*, verifica-se, na linha 39, se 95% do total de execuções encontra-se dentro do IC e, se for o caso, atualiza-se o conjunto de modelos de ML Válidos MLV . Ao final do processamento do ranking, o índice do modelo válido e com melhor desempenho da medida de interesse sobre a validação é armazenado na variável m (linha 43). No final, o algoritmo retorna o modelo robusto de melhor desempenho ML_m e suas estatísticas de interesse (linha 44).

O gráfico à esquerda da Figura 4.5 ilustra um exemplo de simulação para 5000 amostras de validação (em azul) e 500 amostras de teste (em vermelho) com média igual a 8 e desvio padrão igual a 1.5. Os picos representam os valores mais comuns e a dispersão da distribuição indica o quanto os resultados variaram. As distribuições estão sobrepostas

Algorithm 1 Seleção de Modelo Robusto (SMR)

INPUT: Conjunto de dados original A de tamanho $a = |A|$; percentual de validação p_val ; quantidade de *folds* K ; quantidade de repetições R ; tamanho do ranking tam_rank ; total de execuções *Bootstrap* N ;
OUTPUT: Modelo robusto de melhor desempenho ML_m ; matriz de confusão Mat_m ; medidas de teste $MTeste_m$ e validação $MVal_m$ do modelo ML_m .

```
1:  $X \leftarrow$  divisão estratificada de  $A$  em  $a \times p\_val$  dados de amostra de validação
2:  $Y \leftarrow$  divisão estratificada de  $A$  em  $a \times (1 - p\_val)$  dados de amostra de teste
3:  $MLV \leftarrow \emptyset$ 
4: Ranking  $\leftarrow$  conjunto de modelos de ML otimizados pela medida de interesse sobre  $X$ , de tamanho  $tam\_rank$ .
5: for  $b \leftarrow 1, \dots, tam\_rank$  do
6:    $i \leftarrow 1$ 
7:    $MVal_b \leftarrow \emptyset$ 
8:   Seja  $ML_b$  o  $b$ -ésimo modelo de ML sobre  $X$ 
9:   for  $r \leftarrow 1, \dots, R$  do ▷ validação cruzada com repetição
10:    for  $k \leftarrow 1, \dots, K$  do
11:      Seja  $X(k-1)$  a amostra de validação excluindo-se o  $k$ -fold
12:      Seja  $X(k)$  a amostra de validação correspondente ao  $k$ -fold
13:      O modelo é treinado nos  $X(k-1)$  folds e validado no  $X(k)$  fold
14:      Seja  $M_i$  a medida do  $k$ -fold na repetição  $r$ 
15:       $MVal_b \leftarrow MVal_b \cup M_i$ 
16:       $i \leftarrow i + 1$ 
17:    end for
18:  end for
19:  Calcula-se a média  $\overline{M}_b \leftarrow \frac{1}{K \times R} \sum_{i=1}^{K \times R} M_i$ 
20:  Calcula-se o desvio padrão  $DP_b \leftarrow \sqrt{\frac{1}{K \times R} \sum_{i=1}^{K \times R} (M_i - \overline{M}_b)^2}$ 
21:  Calcula-se o limite inferior LI do IC de 95%:  $LI_b \leftarrow \overline{M}_b - 1,96 \times \frac{DP_b}{\sqrt{K \times R}}$ 
22:  Calcula-se o limite superior LS do IC de 95%:  $LS_b \leftarrow \overline{M}_b + 1,96 \times \frac{DP_b}{\sqrt{K \times R}}$ 
23:   $Mat_b \leftarrow (0, 0, 0, 0)$ 
24:   $MTeste_b \leftarrow \emptyset$  ▷ medidas de teste
25:   $MValidos_b \leftarrow \emptyset$  ▷ medidas de teste dentro do IC
26:  for  $j \leftarrow 1, \dots, N$  do
27:    Gera-se a amostra Bootstrap  $Z_j$  sobre  $Y$  de tamanho  $(1 - p\_val) \times a$ 
28:    Aplica-se o modelo  $ML_b$  sobre  $Z_j$ 
29:    for  $i \leftarrow 1, \dots, (1 - p\_val) \times a$  do
30:      Gera-se a previsão  $\hat{z}_i$  sobre  $Z_j$ 
31:    end for
32:    Calculam-se (VP, VN, FN, FP) sobre  $Z_j$  e atualiza-se  $Mat_b$ 
33:    Calcula-se a estatística de teste  $Teste_j$  sobre  $Z_j$ 
34:     $MTeste_b \leftarrow MTeste_b \cup \{Teste_j\}$ 
35:    if  $Teste_j \geq LI_b$  and  $Teste_j \leq LS_b$  then ▷ medida dentro do IC de 95%
36:       $MValidos_b \leftarrow MValidos_b \cup \{Teste_j\}$ 
37:    end if
38:  end for
39:  if  $|MValidos_b| \geq 95\% \times N$  then ▷ 95% das medidas dentro do IC
40:     $MLV \leftarrow MLV \cup \{ML_b\}$ 
41:  end if
42: end for
43:  $m \leftarrow \{i, i = 1, \dots, |MLV| : ML_i \text{ de melhor desempenho para a medida de interesse}\}$ 
44: return  $ML_m, Mat_m, MTeste_m, MVal_m$ 
```

pois os picos estão bem próximos entre si e os valores de teste estão enquadrados nos limites da validação. Este modelo indica boa generalização na previsão das amostras de

teste. Já no gráfico à direita, ilustra-se um exemplo de simulação com distribuições não sobrepostas. Os valores de teste não estão enquadrados nos limites da validação em sua totalidade. Este modelo não conseguiu estimar previsões robustas em 95% dos experimentos.

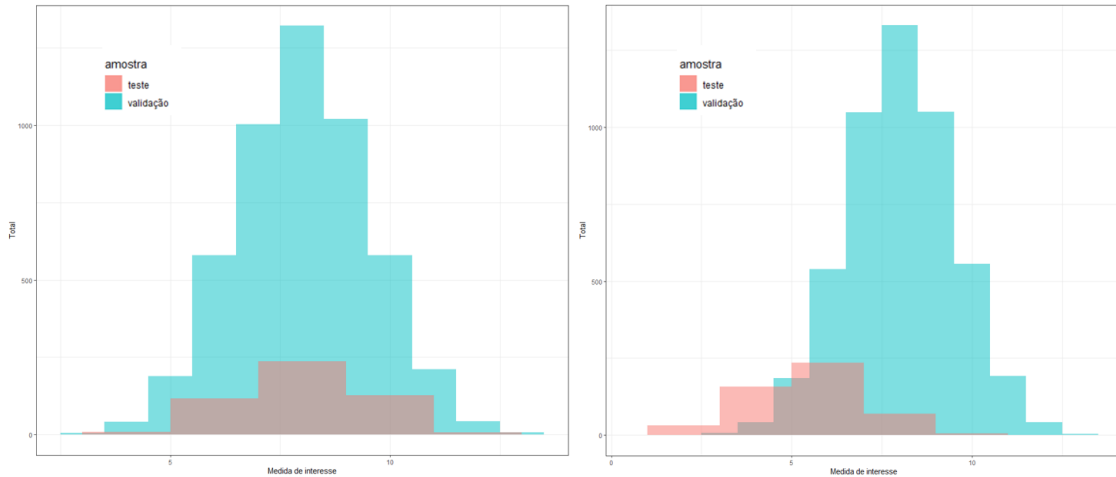


Figura 4.5: Distribuições sobrepostas (gráfico à esquerda) e distribuições não sobrepostas (gráfico à direita)

A Figura 4.6 ilustra o esquema da proposta metodológica. As técnicas de validação cruzada com repetição sobre a amostra de validação e *Bootstrap* sobre a amostra de teste são aplicadas para a seleção do modelo candidato. No esquema, são definidos k -folds, sendo k igual a 5 e um número N de execuções para o procedimento de *Bootstrap*. Para aumentar a robustez das estimativas, também é definido um número R de repetições na técnica de validação cruzada. As medidas $\{M_1, M_2, \dots, M_5\}$ representam os valores observados no k -ésimo *fold* de validação (hachurado em cinza) após o treino nos 4 ($k - 1$) *folds* (fundo branco). “M validation” é a média aritmética dos $k * R$ valores observados, “DP validation” é o desvio padrão dos $k * R$ valores observados, “Hopt” representa os hiperparâmetros do modelo candidato, e “M teste” é o valor observado aplicando o modelo candidato sobre os dados de teste. A partir de “M validation” e “DP validation”, o intervalo de confiança de 95% é calculado como critério para assegurar a robustez dos resultados em “M teste”. O valor de “M teste” é comparado com os limites do IC. Para valores enquadrados dentro dos limites, o modelo candidato é considerado válido. Caso contrário, o modelo não é considerado válido e é descartado. Um novo modelo candidato é selecionado do ranking de modelos sobre a validação. Portanto, cada modelo gera N conjuntos de teste por *Bootstrap*. Ao final de N execuções *Bootstrap* sobre o teste, há N estatísticas de interesse que correspondem aos N conjuntos de teste. Os valores de (VP, VN, FP, FN) da classificação dos N conjuntos de teste são armazenados em uma matriz de confusão única, denominada matriz consolidada. O modelo válido de melhor

desempenho sobre a validação é selecionado.

4.4 Resumo

Em resumo, este capítulo apresentou uma proposta metodológica para geração de modelos robustos e com capacidade de generalização no contexto do diagnóstico de risco de ceratocone. Foi introduzido um critério para priorizar a verossimilhança entre as distribuições, em detrimento de um possível sacrifício da medida de interesse, a partir da aplicação das técnicas de *k-fold* e *Bootstrap* para geração das estimativas. Como estratégia para minimizar a variabilidade do modelo, os IC de 95% são calculados para o conjunto de validação e aplicados como critério de enquadramento para as medidas de teste. Distribuições sobrepostas indicam previsibilidade e robustez, pois nestes casos, os valores correspondem ao verdadeiro valor da média populacional, em 95% das amostras.

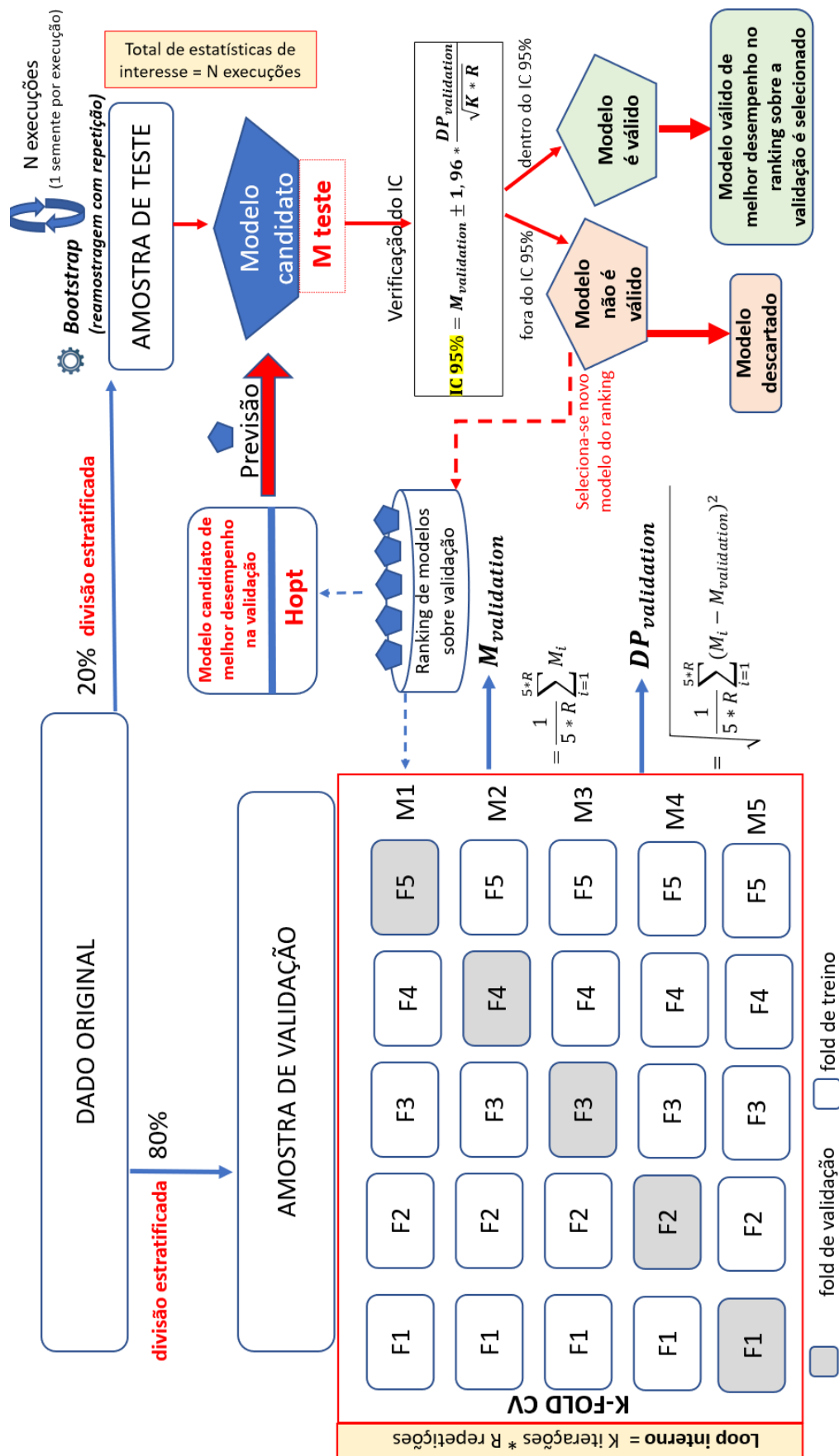


Figura 4.6: Design da proposta metodológica

5. O Experimento Computacional

Neste capítulo é apresentado o desenho do experimento computacional para validação da proposta metodológica. Na Seção 5.1 aborda-se a geração do *ranking* de modelos candidatos sobre a amostra de validação e na Seção 5.2 é apresentado o fluxo das atividades realizadas para treinamento dos modelos candidatos.

5.1 Geração do Ranking de Modelos

A etapa de geração do *ranking* de modelos otimizados pela medida de interesse, apresentada no Algoritmo 1 (SRM) da Seção 4.3, é crucial para o problema da pesquisa. Estes modelos devem apresentar um bom desempenho na medida de interesse sem sacrificar tanto a precisão e ainda atender as restrições introduzidas na proposta metodológica. Para geração destes modelos candidatos sobre a amostra de validação, as seguintes atividades foram realizadas:

- Pré-Processamento: fase de limpeza e seleção dos hiperparâmetros para maximizar a medida de interesse.
- Processamento: aplicação do algoritmo de aprendizado supervisionado sobre a amostra de validação e captura do desempenho.
- Pós-Processamento: geração do *ranking* de modelos ordenados de forma decrescente pela medida de interesse.

5.2 Fluxo de Treinamento do Classificador

O esquema da Figura 5.1 resume o fluxo para o treinamento dos modelos candidatos que serviram de insumo para validação da proposta metodológica (Capítulo 4), culmi-

nando na seleção de um modelo robusto para o classificador. As próximas seções detalham cada etapa do fluxo de treinamento.

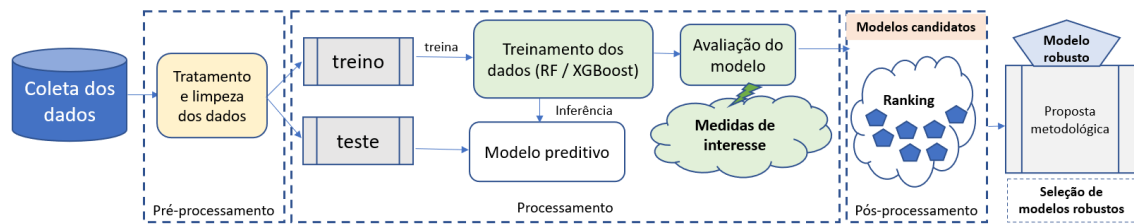


Figura 5.1: Fluxo para o treinamento dos modelos candidatos

5.2.1 Tratamento e Limpeza de Dados

Após a importação dos dados, os registros foram analisados quanto à presença de dados faltantes. Identificou-se a presença de dados faltantes no marcador de astigmatismo (PreopCyl). Como tratamento, um valor padrão foi definido para os dados faltantes. Em sequência, foi realizado o tratamento de valores trocados entre medidas correlacionadas. Foram encontrados valores trocados nos cenários a seguir:

1. Espessura no ápice da córnea (PachyApex) menor que a espessura no ponto mais fino da córnea (ThinnestPachy).
2. Espessura do leito estromal residual no ponto central da córnea (RSBCentral) menor que a a espessura do leito estromal residual no ponto mais fino da córnea (RSBMin).

Para valores trocados no par “espessura no ápice da córnea” (PachyApex) e “espessura no ponto mais fino da córnea” (ThinnestPachy), o marcador “diferença de espessura entre o ponto máximo e mínimo” (DifPachyCentralMin) foi recalculado com o resultado dos marcadores (PachyApex - ThinnestPachy). Ambos os marcadores “percentual de tecido alterado no ponto central” (PTACentral) e “percentual de tecido alterado no ponto mínimo” (PTAMin) também foram recalculados. Para os pares de medidas trocados em “leito estromal residual no ponto central” (RSBCentral) e “leito estromal residual no ponto mínimo” (RSBMin), os marcadores “espessura da córnea no ponto mais fino” (ThinnestPachy) e “espessura da córnea no ápice” (PachyApex) também foram recalculados. Consequentemente, os marcadores “diferença de espessura entre o ponto máximo e mínimo” (DifPachyCentralMin), “percentual de tecido alterado no ponto central” (PTACentral) e “percentual de tecido alterado no ponto mínimo” (PTAMin) também foram recalculados, de acordo com as equações de correlação entre as variáveis (veja Seção 2.6.3).

Em seguida, as variáveis categóricas foram convertidas em fatores. Esta é uma técnica para representar dados categóricos, associando-se um rótulo numérico para cada valor possível. Em relação aos marcadores considerados para treinamento do classificador, somente foram considerados os marcadores que poderiam interferir no risco de ceratocone e inerente às métricas oftalmológicas. Portanto, os marcadores “olho diagnosticado” (ExamEye) e “técnica do retalho” (FlapTechnique) foram removidos do conjunto de dados. A informação de “ExamEye” descreve se o olho diagnosticado é esquerdo ou direito, sendo considerada irrelevante para treinamento do classificador. O marcador “FlapTechnique” foi excluído por não ser diretamente condicionado a uma medida oftalmológica intrínseca do paciente.

5.2.2 Divisão em Dados de Treino e Teste

Os dados foram divididos de forma estratificada, isto é, mantendo a proporção dos rótulos em 80% para o conjunto de treino e 20% para o conjunto de teste. Assim, 2.622 registros sem risco de ceratocone e 84 registros com risco de ceratocone foram coletados. Para o conjunto de teste, 656 registros foram classificados sem risco de ceratocone e 21 registros classificados com risco de ceratocone.

5.2.3 Modelo Experimental de Floresta Aleatória

Com base nas características apresentadas na Seção 3.1.8, o algoritmo de Floresta Aleatória foi selecionado por ser amplamente conhecido e utilizado pela literatura, sendo uma tecnologia madura e bem estabelecida para um problema que exige um modelo com tal grau de maturidade. A busca em grade aleatória dos hiperparâmetros (Seção 3.1.10) foi inicialmente definida de forma a maximizar a métrica de cobertura (Equação 3.2), e que ainda retornasse no mínimo 60% de precisão (Equação 3.1). A cobertura mede a capacidade do modelo em detectar olhos com risco de ceratocone. Considerando que é muito mais crítico indicar para a cirurgia LASIK um olho com risco de ceratocone (falsos negativos) do que contra-indicar um olho sadio para cirurgia LASIK (falsos positivos), esta métrica foi selecionada por ser mais sensível para a presença de falsos negativos. Em um segundo momento, a medida de *negative log loss* também foi selecionada. Os seguintes hiperparâmetros foram definidos para treinamento do classificador: número de árvores para gerar um modelo de classificação (*n_estimators*), profundidade máxima da árvore (*max_depth*), número mínimo de amostras para se dividir um nó interno (*min_samples_split*), número mínimo de amostras em um nó folha (*min_samples_leaf*).

O modelo otimizado por cobertura foi denominado de “Modelo por Cobertura - Flo-

resta Aleatória” e o modelo otimizado por *negative log loss* (Seção 3.1.5) foi denominado de “Modelo Ajustado por *Negative Log Loss* - Floresta Aleatória”.

5.2.4 Modelo Experimental de XGBoost

O XGBoost, apresentado na Seção 3.1.9, foi o segundo algoritmo selecionado no experimento. Este destaca-se como uma técnica *Boosting* avançada de aprendizado gradual em que a geração de um novo modelo base depende do resultado do classificador base anterior (comitê). A cada iteração, o modelo aprende com os erros (ruídos) e corrige os erros. Como resultado, o modelo XGBoost combina as contribuições individuais de cada iteração para retornar a predição final (método *ensemble*). Ao contrário do algoritmo de Floresta Aleatória, o XGBoost permite um maior nível de ajuste aos dados de treino, customização do critério de avaliação do modelo, grande variedade de regularização, entre outras características.

As técnicas de validação cruzada *k-fold* (Seção 3.1.11.1), em conjunto com a técnica de busca em grade da Seção 3.1.10, foram aplicadas de forma a retornar o melhor desempenho para a medida de interesse. Este modelo foi denominado de “Modelo Ajustado - XGBoost”. Os seguintes hiperparâmetros foram definidos: número de árvores (*n_estimators*), taxa de aprendizado (*learning_rate*), termo de regressão (*gamma*), profundidade máxima da árvore (*max_depth*), taxa de amostragem por colunas (*colsample_bytree*), taxa de amostragem (*subsample*), termo de regularização L2 (*lambda*).

5.2.5 Definição do Valor de Corte (em inglês, *threshold*)

Um ponto de atenção é determinar um valor de corte (Seção 3.1.3) que seja aceitável: nem tão baixo a ponto de não ser representativo para o experimento, e nem tão alto mas suficiente para alcançar um desempenho interessante nas medidas de interesse. Para o diagnóstico médico, classificar incorretamente olhos com risco de ceratocone como olhos saudáveis (falsos negativos) é um cenário crítico e que deve ser minimizado, considerando a saúde do paciente candidato à cirurgia LASIK. Portanto, foi definido um valor de corte abaixo de 50% para a classificação das observações, tornando o modelo mais sensível e menos preciso na classificação correta das observações da classe minoritária (de menor proporção), referente aos olhos com o risco de ceratocone (3% do conjunto de dados).

5.2.6 Geração das Medidas de Interesse

A medida de cobertura (*recall*) foi inicialmente selecionada como fator de otimização, por ser sensível em minimizar os falsos negativos. Em um segundo momento, a função de perda *negative log loss* (Seção 3.1.5) também foi selecionada de forma a avaliar o critério de restrição do IC. As medidas da matriz de confusão (cobertura, precisão e especificidade) foram calculadas considerando o valor de corte pré-definido. A matriz de confusão (Seção 3.1.2) relaciona os valores de entrada com os resultados previstos, segregando em quatro quadrantes: os verdadeiros positivos (casos positivos classificados como positivos), verdadeiros negativos (casos negativos classificados como negativos), falsos positivos (casos negativos classificados como positivos) e falsos negativos (casos positivos classificados como negativos). Desta forma, a cobertura (ou sensibilidade) mede a proporção dos registros com risco de ceratocone que foram identificados corretamente. A especificidade mede a proporção de olhos saudáveis que foram identificados corretamente. E a precisão mede a proporção de casos que foram corretamente previstos da classe positiva.

5.2.7 Avaliação do Modelo

O modelo foi avaliado quanto à função de perda *negative log loss* e às métricas da matriz de confusão: cobertura, especificidade e precisão. Os modelos candidatos foram ordenados pela medida de interesse para compor o ranking que serviu de insumo para a proposta metodológica.

5.3 Resumo

Em resumo, este capítulo apresentou as atividades do experimento computacional para geração do ranking de modelos candidatos que serviram de insumo para validação da proposta metodológica. Considerando a criticidade na classificação de pacientes com risco de ceratocone, a busca em grade foi definida de forma a minimizar os falsos negativos, sem sacrificar tanto os falsos positivos. Em um segundo momento, o modelo também foi ajustado por *negative log loss*, com o propósito de avaliar o critério de restrição introduzido no esquema metodológico do Capítulo 4. Os modelos foram treinados com os algoritmos de Floresta Aleatória e XGBoost.

6. Resultados e Discussão

Neste capítulo, os resultados obtidos nos experimentos definidos no Capítulo 4 são apresentados e discutidos, a partir dos modelos candidatos gerados no Capítulo 5.

As técnicas de validação cruzada *k-fold* (Seção 4.3.1) e *Bootstrap* (Seção 3.1.14) foram empregadas para geração das distribuições e cálculo dos intervalos de confiança de 95%. Um IC de 95% significa que em 95% das amostras a média variou entre um valor mínimo e máximo, e em 5% das amostras essa média extrapolou os limites do intervalo. Foram realizados os seguintes experimentos: RF otimizando a cobertura com no mínimo de 60% de precisão; RF otimizando o *negative log loss*; RF que atendeu ao critério do IC 0,95 por *negative log loss*; XGBoost otimizando o *negative log loss*, e XGBoost que atendeu ao critério do IC 0,95 por *negative log loss*. Para avaliação dos resultados, foram gerados os histogramas das distribuições sobre validação e teste, diagramas de caixa, estatísticas de interesse, e as matrizes de confusão correspondentes. Os resultados do modelo por cobertura são apresentados na Seção 6.1, e os resultados dos modelos por *negative log loss* são apresentados na Seção 6.2.

6.1 Modelo por Cobertura - Floresta Aleatória

Para seleção do modelo candidato que retorne um maior percentual de amostras dentro do IC de 95%, foram gerados 100 modelos candidatos otimizados pela medida de cobertura e com no mínimo 60% de precisão. Ao final, gerou-se um ranking decrescente de modelos candidatos de RF (Seção 5.2.3), contendo sua posição no *ranking* e o desempenho da cobertura. Para cada modelo, foi aplicado o esquema metodológico do Capítulo 4, empregando-se as técnicas de *Bootstrap* e validação cruzada *k-fold*, sendo *k* igual a 10. Após 100 execuções *Bootstrap*, cada modelo resultou em 100 estatísticas de teste.

O modelo candidato melhor posicionado no *ranking* foi inicialmente escolhido de

forma a minimizar os falsos negativos, questão essencial para o problema de ceratocone. Este modelo retornou o melhor desempenho sobre a amostra de validação, porém não obedeceu às restrições do IC de 95% quando aplicado sobre os dados de teste, sendo considerado uma solução inválida. Desta forma, buscou-se selecionar um modelo dentre os modelos retornados no *ranking*, àquele que atendesse melhor as restrições do IC sem comprometer tanto as medidas de interesse. Os modelos que não atenderam ao critério do IC de 95% foram considerados inválidos e descartados, apesar de otimizarem a função objetivo.

A Figura 6.1 apresenta os diagramas de caixa (em inglês, *boxplot*) e os histogramas com a distribuição dos valores de cobertura para as amostras de teste (em azul) e validação (em vermelho). São apresentados dois modelos candidatos: à direita, é exibida a distribuição para o modelo que retornou o melhor desempenho sobre os dados de teste, posição 1 no *ranking*; à esquerda, é exibida a distribuição para o modelo na posição 96 do *ranking* sobre os dados de teste. O modelo pior posicionado no *ranking* obteve maior percentual das amostras de teste enquadradas dentro do IC em relação ao modelo posicionado em primeiro lugar, de melhor desempenho. Também pode-se observar que o modelo com melhor desempenho obteve maior percentual de *outliers* sobre os dados de validação, em relação ao modelo de menor desempenho. As médias das distribuições também são exibidas nos histogramas (linhas tracejadas), evidenciando médias mais próximas entre si no modelo de posição 96 do *ranking* em comparação ao modelo de posição 1 do *ranking*. Este comportamento evidencia uma acentuada distorção e assimetria entre as amostras de validação e teste, principalmente nas distribuições geradas pelo modelo de melhor desempenho.

A Figura 6.2 exibe o *ranking* de modelos otimizados em duas escalas: à esquerda, o % de amostras (em azul) que atenderam ao critério do IC de 95%; à direita, a média de cobertura (em preto) observada. As medidas foram coletadas após o processamento de 100 execuções *Bootstrap* por modelo sobre os dados de teste, aplicando o esquema metodológico proposto. Pode-se observar que as soluções candidatas posicionadas no *ranking* entre 1 e 100 não atenderam ao critério do IC, ou seja, não foram consideradas válidas para o critério proposto, apesar de otimizarem a função objetivo. Porém, modelos com posição inferior no *ranking* obtiveram um resultado melhor em percentual de amostras enquadradas dentro do IC de 95%. Estes modelos se aproximaram mais ao critério restritivo proposto e sem impactar em demasiado as medidas de interesse sobre os dados de teste: cobertura = 0,74; precisão = 0,83 (posição 96 com 66% das amostras dentro do IC) vs. cobertura = 0,76; precisão = 0,85 (posição 1 com 4% das amostras dentro do IC). O comportamento acima permitiu observar que modelos mais bem posicionados no *ran-*

king manifestaram acentuada assimetria entre as distribuições de validação e teste, com valores sobre os dados de teste extrapolando os limites calculados do IC de 95% sobre a validação. O modelo com melhor desempenho no *ranking* na verdade ludibria o seu valor real pois, se realmente fosse extraído por amostras distintas da mesma população, o valor observado corresponderia ao valor médio populacional. Este comportamento sugere que as amostras foram geradas por populações diferentes, consequência do *overfitting* no conjunto de validação. Portanto, nesses modelos, não foi possível garantir a previsibilidade e generalização em novas amostras, e o desempenho aparente não corresponde a real capacidade dos modelos gerados. Desta forma, a seleção de um modelo candidato válido é decisiva para se evitar um possível falso diagnóstico da classe positiva de ceratocone. Os modelos candidatos somente são considerados válidos se atenderem ao critério restritivo proposto e ainda alcançar um percentual de cobertura mínimo aceitável para o diagnóstico de risco de ceratocone sem comprometer tanto os falsos positivos.

A medida de cobertura (*recall*) foi inicialmente selecionada pois está diretamente relacionada ao problema da pesquisa. Pode-se verificar que esta medida, mesmo sendo aderente ao desbalanceamento de classes, apresentou acentuada assimetria e distorção entre teste e validação, mostrando-se custosa para atender ao critério do IC. Portanto, pode-se observar que quando a métrica escolhida é aderente ao domínio do problema mas não é o critério natural da função objetivo, produz-se uma elevada assimetria, uma distorção nas distribuições que pode dificultar o atendimento ao critério de restrição do IC. Em um segundo momento, escolheu-se a medida de *negative log loss* (Seção 3.1.5) que é uma função de perda indicada para problemas de classificação binária. As próximas seções apresentam os resultados destes experimentos.

6.2 Modelos por *Negative Log Loss*

Esta seção apresenta os resultados da proposta metodológica do Capítulo 4, aplicando-se as técnicas de validação cruzada *k-fold* sobre a validação, com *k* igual a 10, e *Bootstrap* em 100 execuções sobre o teste. O critério de restrição do IC foi baseado na medida de *negative log loss*.

6.2.1 Histograma das Distribuições

Os histogramas foram gerados para identificar o comportamento das distribuições sobre as amostras de validação e teste em cada experimento realizado. O eixo *y* contém a frequência do total de amostras para cada valor de *negative log loss* (eixo *x*) sobre a

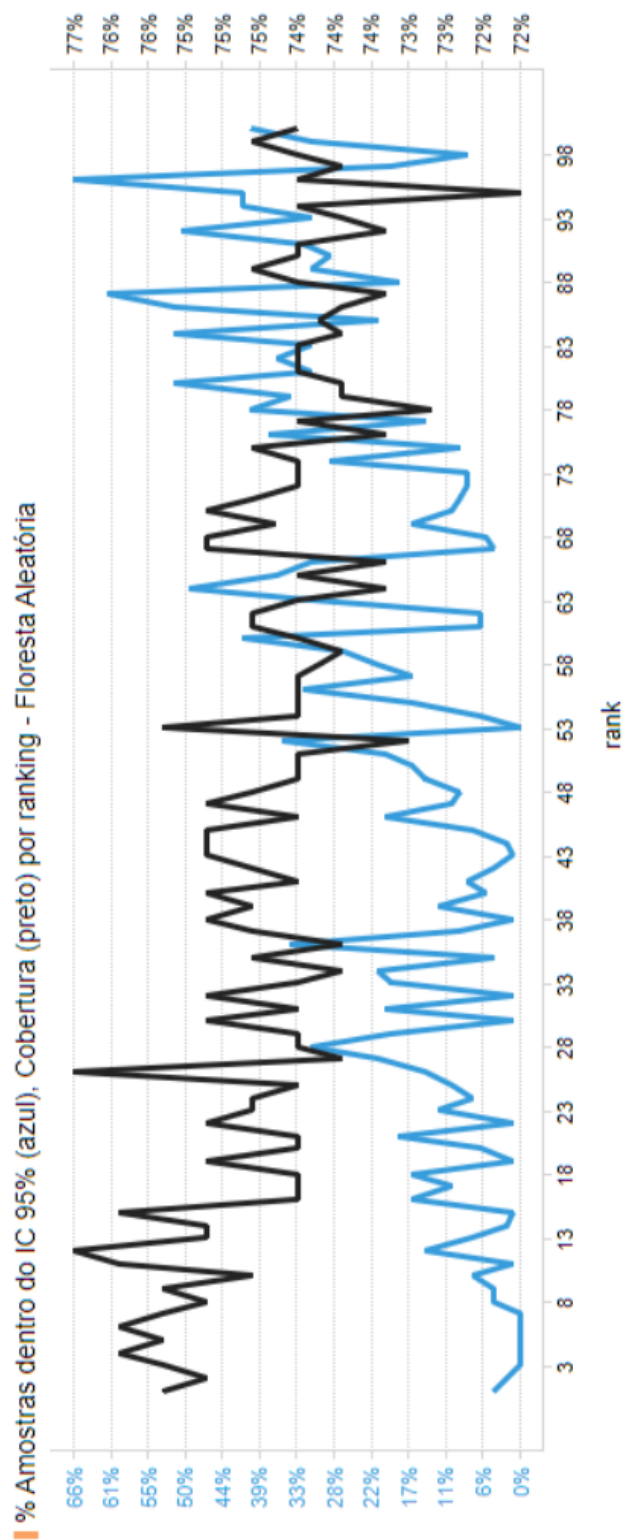


Figura 6.2: Modelos candidatos de RF por posição no *ranking* (eixo x) em duas escalas: à esquerda, % de amostras (em azul) que atenderam ao critério restritivo do IC de 95%; à direita, desempenho da medida de cobertura (em preto). Valores observados dos 100 modelos candidatos com melhor posição no *ranking* por cobertura, aplicando-se *k-fold* sobre a validação e *Bootstrap*, em 100 execuções sobre o teste por modelo

validação (barras em vermelho) e teste (barras em azul). A média, limites inferior (0,025 percentil) e superior (0,975 percentil) de cada distribuição também foram calculados. Os quartis Q1 (primeiro quartil) e Q3 (terceiro quartil) também são visualizados bem como a curva gaussiana, com o objetivo de auxiliar na interpretação do padrão de distribuição em cada conjunto de dados.

A Figura 6.3 exibe o histograma do modelo RF ajustado por *negative log loss*. Neste modelo, os limites inferior e superior da distribuição, indicados pela linha verde tracejada, ficaram situados entre -0,061 (2,5% percentil) e -0,037 (97,5% percentil); e os limites inferior e superior da distribuição sobre a validação ficaram situados entre - 0,074 e - 0,033. Em seguida, os histogramas foram sobrepostos (gráfico à direita) em pontos percentuais de forma a auxiliar na análise gráfica das distorções entre os dados de teste (barra em azul) e validação (barra em vermelho). Pode-se observar que as médias das distribuições de teste e validação estão levemente deslocadas entre si. Isto indica um comportamento distinto entre os dois conjuntos de dados, fato que resultou em apenas 63% das amostras de teste enquadradas dentro do IC de 95% de *negative log loss*. Neste cenário, há evidências de insuficiente robustez no modelo de RF ajustado. Isto sugere um possível *overfitting* aos dados de validação, o que pode ser investigado testando novos modelos. Como o modelo foi considerado inválido pela metodologia proposta, um novo modelo foi selecionado do ranking, de forma a atender ao critério do IC de 95%.

O modelo de RF que atendeu ao critério do IC foi o modelo RF IC 0,95. O histograma deste modelo é exibido na Figura 6.4. Pode-se observar que a média e mediana do modelo de RF IC 0,95 para os dados de teste (barras em azul) e validação (barras em vermelho) estão muito mais próximos entre si quando comparado ao modelo RF ajustado, o que comprova maior sobreposição (gráfico à direita) nos valores da distribuição. Este comportamento também pode ser observado por meio da curva gaussiana sobreposta entre as amostras. Já no modelo ajustado, as curvas gaussianas estão ligeiramente deslocadas entre si, sugerindo maior variabilidade no modelo e não sendo possível confirmar a verossimilhança nos resultados. Ou seja, que as médias reais são de fato o que se calculou.

De forma similar, foram gerados os histogramas para os modelos por XGBoost. A Figura 6.5 exibe o histograma do modelo XGBoost ajustado por *negative log loss*. Pode-se observar que a mediana se encontra em -0,024 e a média se encontra em -0,025 (dados de teste), enquanto que a mediana para os dados de validação se encontra em - 0,026 e a média em -0,029. O comportamento das distribuições pode ser visualizado por meio da sobreposição (gráfico à direita) dos histogramas do modelo XGBoost ajustado. Como este modelo foi considerado inválido pela metodologia proposta, um novo modelo XGBoost foi selecionado do ranking de forma a atender ao critério do IC de 95%.

O modelo de XGBoost que atendeu ao critério do IC foi o modelo XGBoost IC 0,95. A Figura 6.6 exibe o histograma deste modelo. Pode-se observar que a média e mediana para as métricas de *negative log loss* e de sensibilidade nos dados de teste foram inferiores ao desempenho do modelo XGBoost ajustado, sem o rigor do IC. A média *negative log loss* do modelo XGBoost IC 0,95 se encontra em -0,026 vs. -0,025 do modelo XGBoost ajustado (dados de teste). Na sobreposição (gráfico à direita) dos histogramas, pode-se observar que as medianas das distribuições de teste e validação estão mais próximas entre si em relação às medianas do modelo XGBoost ajustado, e com menor desvio padrão no conjunto de teste, com valor de 0,33% contra 0,93% do modelo ajustado. Portanto, a variância dos resultados foi menor no modelo XGBoost IC 0,95 em comparação ao modelo XGBoost ajustado.

6.2.2 *Bootstrap*: Matriz de Confusão Consolidada por *Negative Log Loss*

Em cada modelo, foi aplicada a técnica de *Bootstrap* em 100 execuções sobre a amostra de teste. Cada execução *Bootstrap* resultou na geração de um conjunto de teste e na estatística de interesse correspondente. Portanto, em 100 reamostragens com repetição, 100 conjuntos de teste amostrados por *Bootstrap* foram gerados para cada modelo. Como cada amostra de teste corresponde a 20% do total ($0,20 \times 3383$), cada amostra de teste contém 677 valores observados. A matriz de confusão consolidada, resultado das 100 execuções *Bootstrap* para os dados de teste, representa a classificação (VP, VN, FP, FN) dos 67700 (677×100) valores observados.

A Figura 6.7 exibe a matriz de confusão para os modelos de RF ajustado (à esquerda) e RF IC 0,95 (à direita). No modelo RF ajustado, a medida de sensibilidade se encontra em 75,4% e especificidade em 99% para um valor de corte em 0,28. Das 2100 amostras com risco de ceratocone, o classificador previu corretamente 1585 instâncias e errou 515 amostras. E para as 65600 amostras sem risco de ceratocone, o classificador acertou 65253, ou seja, 99% das amostras. No modelo RF IC 0,95, a medida de sensibilidade de 72,9% confirma um desempenho inferior à sensibilidade de 75,4% alcançada no modelo RF ajustado. Entretanto, este foi o desempenho que melhor atendeu ao critério restritivo do IC de 95% em prol de um diagnóstico com maior robustez e previsibilidade, justificando um maior sacrifício nas métricas.

A Figura 6.8 exibe a matriz de confusão para os modelos XGBoost ajustado (à esquerda) e XGBoost IC 0,95 (à direita). No modelo XGBoost ajustado, a medida de cobertura (sensibilidade) se encontra em 86,8% e especificidade em 99,7%, para um valor de corte em 0,28. Das 65600 amostras sem risco de ceratocone, o classificador acertou

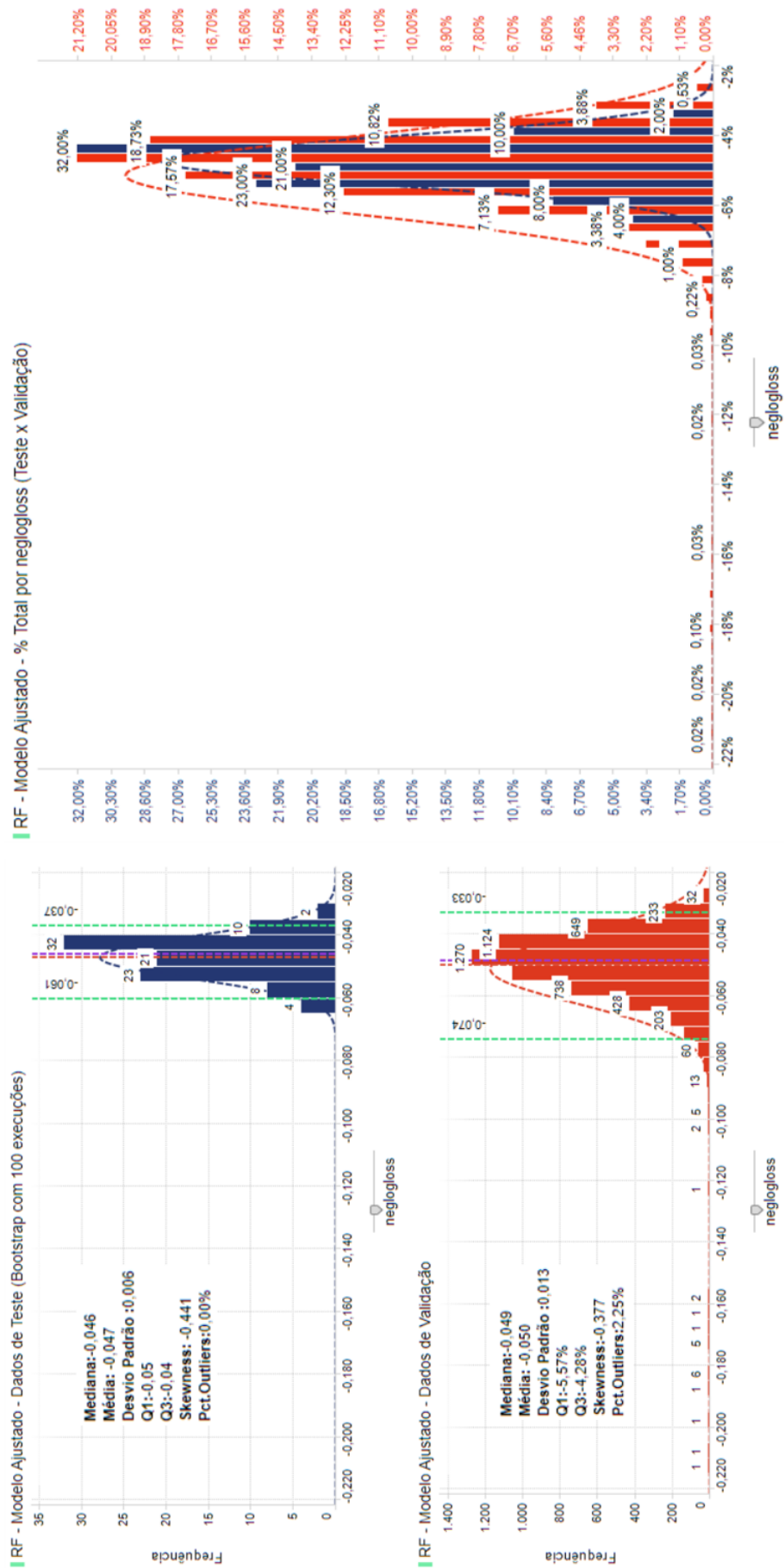


Figura 6.3: Histogramas com os valores de *negative log loss* sobre o teste (barras em azul) e validação (barras em vermelho) - Floresta Aleatória 100 rodadas *Bootstrap* sobre o teste - Modelo ajustado por *negative log loss*

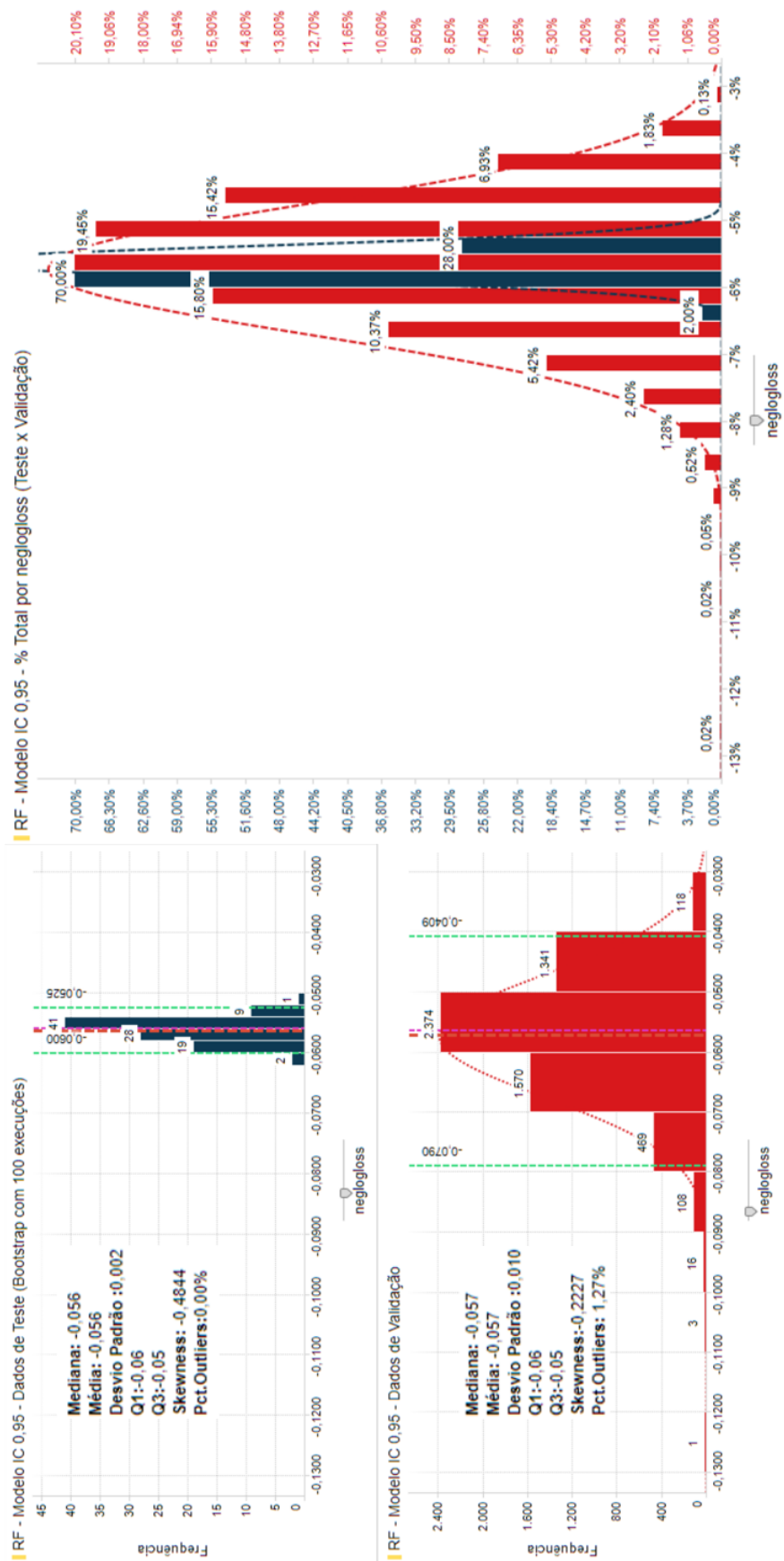


Figura 6.4: Histogramas com os valores de *negative log loss* sobre os dados de teste (barras em azul) e validação (barras em vermelho) - Floresta Aleatória 100 rodadas *Bootstrap* sobre o teste - Modelo IC 0,95 por *negative log loss*

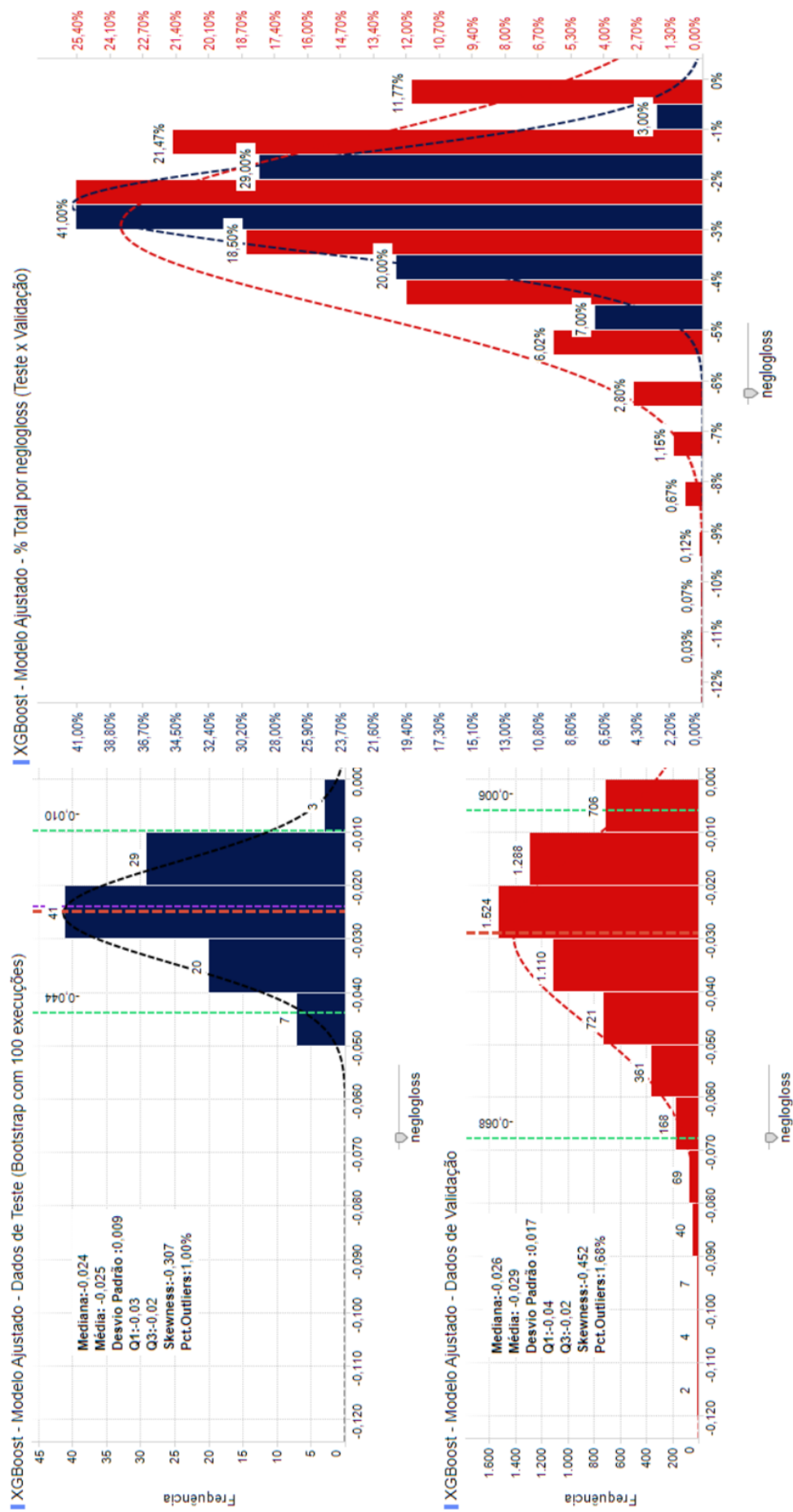


Figura 6.5: Histogramas com os valores de *negative log loss* sobre os dados de teste (barras em azul) e validação (barras em vermelho) - XGBoost 100 rodadas *Bootstrap* sobre o teste - Modelo ajustado por *negative log loss*

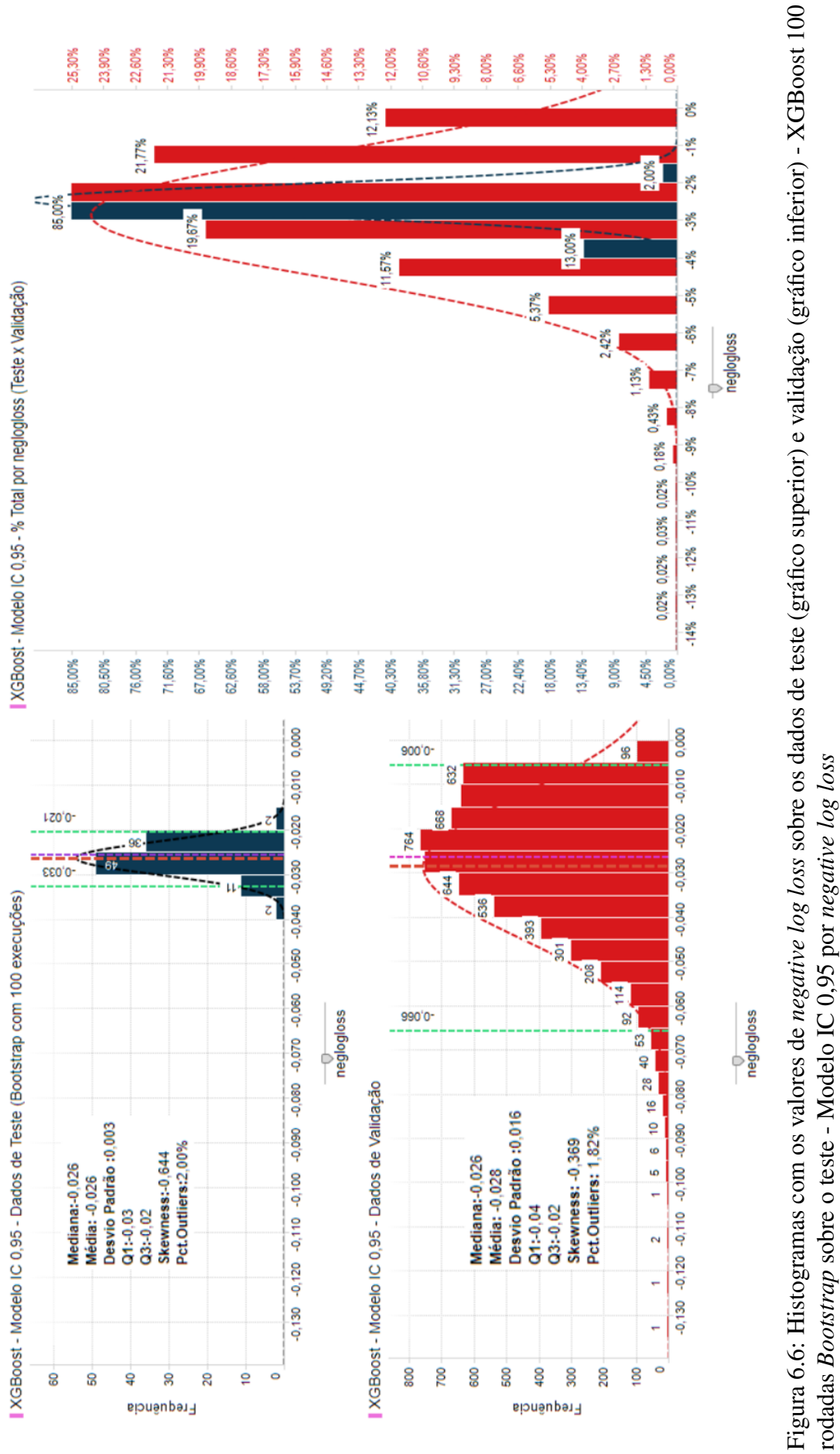


Figura 6.6: Histogramas com os valores de *negative log loss* sobre os dados de teste (gráfico superior) e validação (gráfico inferior) - XGBoost 100 rodadas *Bootstrap* sobre o teste - Modelo IC 0,95 por *negative log loss*

Confusion Matrix and Statistics			Confusion Matrix and Statistics		
Reference			Reference		
Prediction	NO	YES	Prediction	NO	YES
NO	65253	515	NO	65333	569
YES	347	1585	YES	267	1531
Accuracy : 0.9873			Accuracy : 0.9877		
95% CI : (0.9864, 0.9881)			95% CI : (0.9868, 0.9885)		
No Information Rate : 0.969			No Information Rate : 0.969		
P-Value [Acc > NIR] : < 2.2e-16			P-Value [Acc > NIR] : < 2.2e-16		
Kappa : 0.7797			Kappa : 0.7792		
McNemar's Test P-Value : 1.285e-08			McNemar's Test P-Value : < 2.2e-16		
Sensitivity : 0.75476			Sensitivity : 0.72905		
Specificity : 0.99471			Specificity : 0.99593		
Pos Pred Value : 0.82039			Pos Pred Value : 0.85150		
Neg Pred Value : 0.99217			Neg Pred Value : 0.99137		
Prevalence : 0.03102			Prevalence : 0.03102		
Detection Rate : 0.02341			Detection Rate : 0.02261		
Detection Prevalence : 0.02854			Detection Prevalence : 0.02656		
Balanced Accuracy : 0.87474			Balanced Accuracy : 0.86249		
'Positive' Class : YES			'Positive' Class : YES		

Figura 6.7: Matriz de confusão consolidada - Floresta Aleatória 100 rodadas - *Bootstrap* - Modelo ajustado (à esquerda) vs. Modelo IC 0,95 (à direita)

65467 e das 2100 amostras com risco de ceratocone, o classificador acertou 1823. Apesar do bom desempenho do modelo XGBoost ajustado, apenas 57% das amostras de teste estão enquadradas dentro do IC de 95% por *negative log loss*. Este comportamento sugere que o modelo de XGBoost ajustado ainda não garante a robustez e confiabilidade necessária para generalizar em novas amostras de teste, sendo considerado inválido e descartado. Já no modelo XGBoost IC 0,95, um pequeno sacrifício da medida de interesse é esperado e justificado em prol da robustez do modelo, resultado da introdução de viés para minimizar a variância. Das 2100 amostras com risco de ceratocone, o classificador XGBoost IC 0,95 acertou 1772 amostras. Das 65600 sem risco de ceratocone, o classificador acertou 65464. Portanto, a sensibilidade do modelo que atendeu ao critério do IC foi inferior à sensibilidade do modelo XGBoost ajustado (84,3% vs. 86,8%). A especificidade foi equivalente em ambos os modelos, com desempenho de 99%, acertando a classificação de 67236 amostras, em um total de 67700 valores observados sobre o teste.

6.2.3 *Bootstrap*: Estatísticas por *Negative Log Loss*

As seguintes estatísticas em cada distribuição foram coletadas: mediana, média, desvio padrão, Q1, Q3, desvio (em inglês, *skewness*), que indica o grau de assimetria, e % de *outliers* (valores extremos). A *skewness* informa a direção e o grau de assimetria de uma distribuição. O valor foi calculado a partir do resultado de: $3 \times (\text{média} -$

Confusion Matrix and Statistics				Confusion Matrix and Statistics			
		Reference				Reference	
Prediction		NO	YES	Prediction		NO	YES
NO	65467	277		NO	65464	328	
YES	133	1823		YES	136	1772	
Accuracy : 0.9939				Accuracy : 0.9931			
95% CI : (0.9933, 0.9945)				95% CI : (0.9925, 0.9938)			
No Information Rate : 0.969				No Information Rate : 0.969			
P-Value [Acc > NIR] : < 2.2e-16				P-Value [Acc > NIR] : < 2.2e-16			
Kappa : 0.8958				Kappa : 0.8807			
McNemar's Test P-Value : 1.638e-12				McNemar's Test P-Value : < 2.2e-16			
Sensitivity : 0.86810				Sensitivity : 0.84381			
Specificity : 0.99797				Specificity : 0.99793			
Pos Pred Value : 0.93200				Pos Pred Value : 0.92872			
Neg Pred Value : 0.99579				Neg Pred Value : 0.99501			
Prevalence : 0.03102				Prevalence : 0.03102			
Detection Rate : 0.02693				Detection Rate : 0.02617			
Detection Prevalence : 0.02889				Detection Prevalence : 0.02818			
Balanced Accuracy : 0.93303				Balanced Accuracy : 0.92087			
'Positive' Class : YES				'Positive' Class : YES			

Figura 6.8: Matriz de confusão consolidada - XGBoost 100 rodadas - *Bootstrap* - Modelo ajustado (à esquerda) vs. Modelo IC 0,95 (à direita)

mediana)/desvio_padrão, limitada entre -3 e 3. Valores positivos indicariam uma distribuição enviesada para direita e valores negativos indicariam uma distribuição enviesada para esquerda. A distribuição é perfeitamente simétrica para valores iguais a zero embora este comportamento seja improvável de ocorrer em situação reais. Como regra geral, infere-se que:

- Se a assimetria for menor que -1 ou maior que 1, a distribuição é altamente assimétrica.
- Se a assimetria estiver situada entre -1 e -0,5 ou entre 0,5 e 1, a distribuição é moderadamente assimétrica.
- Se a assimetria estiver situada entre -0,5 e 0,5, a distribuição é aproximadamente simétrica.

A Tabela 6.1 exhibe algumas destas estatísticas coletadas para os modelos sobre as distribuições de validação e teste.

Em uma distribuição perfeitamente normal, a média e a mediana retornariam o mesmo valor. Porém, nos histogramas apresentados dos modelos ajustados, pode-se verificar que a média é menor que a mediana. Isto indica uma assimetria enviesada à esquerda na distribuição, refletida pelo sinal negativo indicado na medida de *skewness*. Esta assimetria

modelo	amostra	média	mediana	desvio padrão	skewness	% outliers
RF ajustado	teste	-0,047	-0,046	0,006	-0,441	0
RF ajustado	validação	-0,05	-0,049	0,013	-0,377	2,25
RF IC 0,95	teste	-0,056	-0,056	0,002	-0,480	0
RF IC 0,95	validação	-0,057	-0,057	0,001	-0,220	1,27
XGBoost ajustado	teste	-0,025	-0,024	0,009	-0,300	1
XGBoost ajustado	validação	-0,029	-0,026	0,017	-0,450	1,68
XGBoost IC 0,95	teste	-0,026	-0,026	0,003	-0,640	2
XGBoost IC 0,95	validação	-0,028	-0,026	0,016	-0,360	1,82

Tabela 6.1: Estatísticas para as distribuições de validação e teste - Modelos por *negative log loss*

é resultado da presença de *outliers* (valores extremos) nos dados de validação e o fato de a medida de interesse ser limitada por zero como valor máximo possível. Embora estas observações extremas não afetem o cálculo da mediana, a média é afetada, e consequentemente o cálculo dos limites do IC. Portanto, é importante verificar o quão severo é a assimetria da amostra para garantir a credibilidade do experimento. Um comportamento altamente assimétrico seria indesejável e poderia afetar diretamente na interpretação das conclusões com base nos critérios da proposta metodológica. Um valor de *skewness* próximo a -0,5 é um comportamento comum e esperado em dados reais.

De acordo com o teorema do limite central, a distribuição amostral aproxima-se da distribuição normal (curva gaussiana), e de fato foi o que se observou na análise das simetrias. A distribuição resultante seguiu um padrão aproximadamente normal de simetria, orbitando sobre a média da medida de interesse e com poucos valores extremos observados. Apesar da influência dos *outliers* para o cálculo do IC, sua presença não significou um impedimento para que as amostras de teste atendessem ao critério proposto do IC de 95% calculado sobre a validação. Eliminar a presença de *outliers* não é uma tarefa simples e demandaria um estudo mais aprofundado sobre as origens deste fenômeno. Fatores genéticos ou ambientais não mapeados nos dados de origem podem ter influenciado no classificador. Também há a possibilidade de erro durante a mensuração e registro dos dados, que poderia ocasionar uma discrepância entre as medidas coletadas.

Pode-se observar também que os valores da mediana e média nos modelos IC 0,95 foram inferiores à mediana e média dos modelos ajustados. Este comportamento indica um desempenho médio inferior dos modelos IC 0,95 em relação aos modelos ajustados. Porém, as medianas dos modelos IC 0,95 estão mais próximas entre si em relação aos modelos ajustados, e com menor desvio padrão no conjunto de teste. Este comportamento sugere uma menor variabilidade nos resultados dos modelos IC 0,95, decorrente da intro-

dução do viés. Portanto, os critérios baseados no IC em prol da robustez se mostraram aderentes para o alcance de um resultado satisfatório do experimento computacional.

A Seção 6.2.4 apresenta uma análise gráfica em diagramas de caixa para comparar o desempenho de cada modelo por *negative log loss* (Seção 6.2).

6.2.4 Diagrama de Caixa das Distribuições

Para comparar os resultados da Seção 6.2, gerou-se um consolidado com os valores de *negative log loss* em diagrama de caixa (em inglês, *boxplot*). A Figura 6.9 exibe o resultado dos modelos IC 0,95 e ajustado, aplicando os algoritmos de RF e XGBoost. O resultado é exibido para o conjunto de teste (em azul) e validação (em vermelho).

As medidas estatísticas do modelo de RF IC 0,95 são rotuladas por “testRF” (teste) e “valRF” (validação). Para o modelo XGBoost IC 0,95, as medidas estatísticas são rotuladas por “testxg” (teste) e “valxg” (validação). Para o modelo de Floresta Aleatória ajustado, as medidas estatísticas são rotuladas por “testRFAJ” (teste), e “valRFAJ” (validação). E para o modelo de XBoost ajustado, as medidas estatísticas são rotuladas por “testxgAJ” (teste), e “valxgAJ” (validação).

A partir dos resultados, pode-se observar que os modelos IC 0,95 aplicados sobre os dados de teste geraram resultados enquadrados dentro do IC de 95% calculado para os dados de validação, em 95% das amostras. Já os modelos ajustados pela medida de desempenho, extrapolaram o IC, não garantindo uma boa generalização em seus resultados. Apenas 63% das amostras de teste se enquadraram dentro do IC 0,95 para o modelo RF ajustado e 57% das amostras de teste se enquadraram dentro do IC 0,95 para o modelo XGBoost ajustado, embora o desempenho desses modelos ajustados tenha sido superior em comparação aos modelos IC 0,95. Ao comparar o desempenho dos modelos IC 0,95, o modelo XGBoost obteve melhor desempenho das medidas de interesse em relação ao modelo RF sobre o teste. O modelo de XGBoost IC 0,95 obteve uma mediana de -0,026 contra -0,056 no modelo de RF IC 0,95.

Com base na matriz de confusão dos modelos de IC 0,95 (Seção 6.2.2), observa-se que o modelo de XGBoost IC 0,95 também alcançou melhor sensibilidade (0,843) quando comparado ao modelo de RF IC 0,95 (0,729), sendo o experimento que alcançou melhor desempenho e robustez nos resultados para o problema da pesquisa.

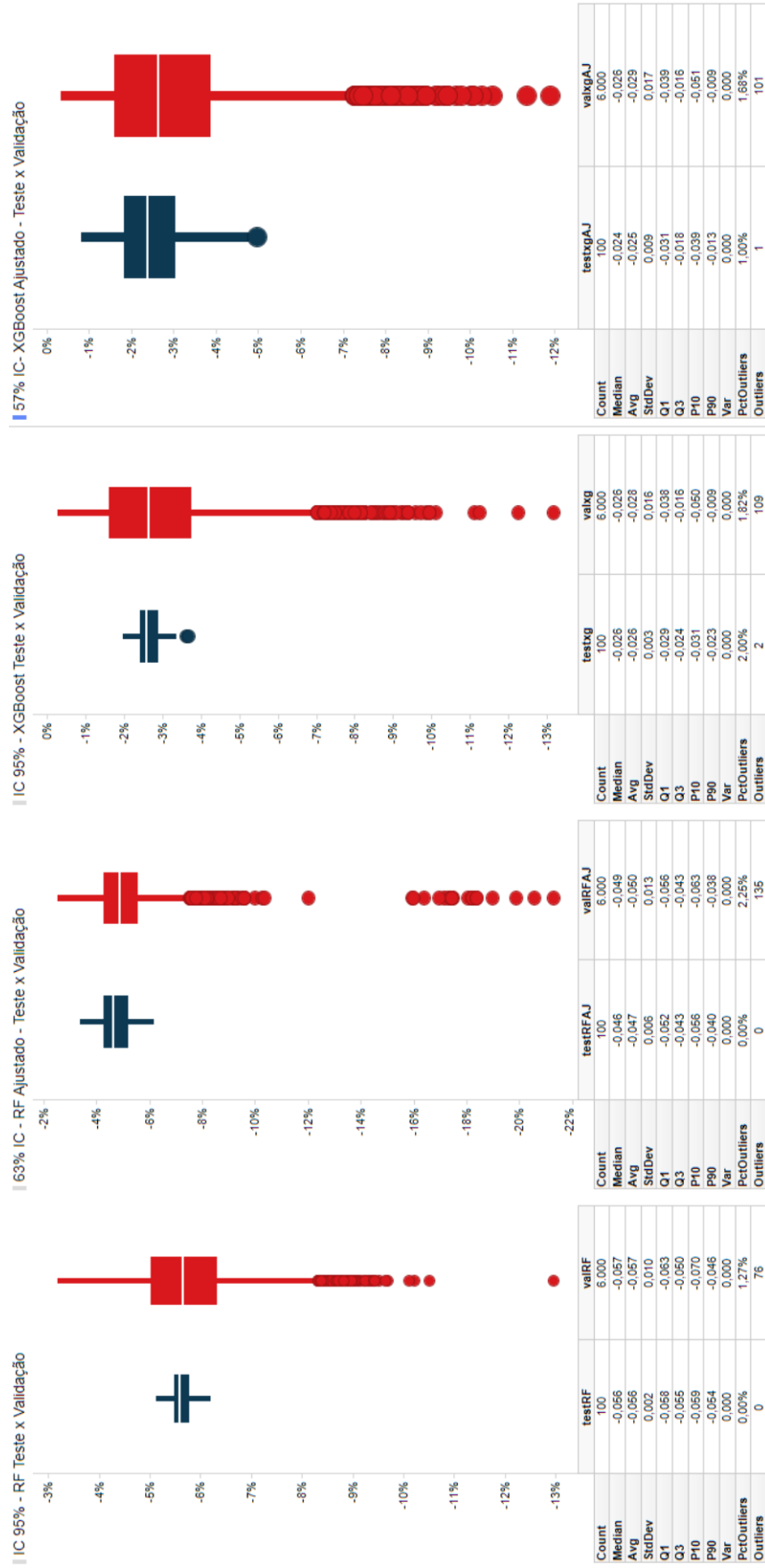


Figura 6.9: *Boxplot* com os valores de *negative log loss* sobre os dados de teste (barras em azul) e validação (barras em vermelho) - *Bootstrap* de 100 rodadas - Floresta Aleatória (RF) e XGBoost - Modelo IC 0,95 e Modelo ajustado

6.3 Discussão dos Resultados

A seguir, alguns pontos relevantes sobre as distribuições observadas:

1. Assimetria das distribuições por cobertura (*recall*):

Pode-se observar que a arbitragem por um critério de otimização como o *recall*, que não é um critério natural de otimização do algoritmo, produziu uma acentuada distorção nas distribuições entre as amostras. Esta distorção foi consequência das classes desbalanceadas, influenciando de forma acentuada na simetria entre amostras. A assimetria também dificultou na escolha dos hiperparâmetros de forma a atender ao critério de restrição do IC. Também pode-se observar que modelos candidatos melhor posicionados no *ranking* de modelos otimizados pela medida de cobertura, resultaram em ser os modelos com menor percentual de aderência ao critério restritivo proposto na metodologia. Enquanto que modelos com desempenho inferior apresentaram maior percentual de amostras dentro do IC, sugerindo maior robustez nos resultados alcançados pelo critério proposto do IC de 95%.

Desta forma, foi oportuno investigar modelos com desempenho levemente sacrificados pela medida de interesse, porém, de maior robustez para a detecção de risco de ceratocone. Estes apresentaram menor distorção e variabilidade, corroborando com a introdução do critério do IC de 95% no esquema metodológico proposto. A estratégia do ranking para seleção de modelos foi eficiente, contribuindo na seleção de modelos com melhor capacidade de generalização, apesar da dificuldade no atendimento da restrição do IC.

2. Assimetria das distribuições por *negative log loss*:

A aplicação do experimento para a medida de *negative log loss* foi oportuna para investigar o comportamento das distribuições para um problema de classificação binária com dados reais e validar a metodologia. As distribuições permitiram auferir a capacidade de atendimento ao critério restritivo introduzido na metodologia. Pela análise gráfica (Figuras 6.3, 6.4, 6.5, 6.6), as amostras de validação aparentaram ser levemente enviesadas para esquerda. Como o valor máximo do *negative log loss* é zero, é de se esperar que as amostras não sejam perfeitamente simétricas. A leve assimetria negativa e próxima a zero nos modelos de RF e XGBoost (Tabela 6.1) está dentro do padrão de comportamento esperado em distribuições aproximadamente simétricas (distribuição normal). Outra medida que reforçou este comportamento foi o percentual baixo de *outliers* para ambos os modelos. Portanto, a convergência observada nas distribuições por *negative log loss* sugeriu maior previsibilidade

e capacidade de generalização em novas amostras, apesar de não ser uma medida sensível para o desbalanceamento de classes.

3. Justificativa para os *outliers*:

Em todo fenômeno real, que precisa ser reduzido em uma estrutura de dados de forma a representar um comportamento de uma população, pode ocorrer perda de informações que não foram mapeadas. Fatores extrínsecos também podem influenciar na classificação, mas que não são dependentes das variáveis amostradas. Uma visão simplificada dos dados pode sugerir que todas as relações entre as variáveis são conhecidas e foram estruturadas. Porém, variáveis extrínsecas nem sempre estão mapeadas no dado original. Como exemplo, pode-se citar a presença de fatores genéticos ou ambientais que podem ter deliberado no desenvolvimento do ceratocône (coceira nos olhos, acidente, etc.). Em geral, tende-se a ignorar explicações alternativas e valoriza-se a forma como a robustez pode ser quantificada por probabilidades. Estas hipóteses demonstram o quão difícil é interpretar um resultado estatístico sem explorar todas as nuances envolvidas.

4. Conclusões equivocadas:

Falsas interpretações podem levar a conclusões equivocadas. Em geral, para quantificar a robustez do modelo, recorre-se ao rigor estatístico a fim de se evitar falsas conclusões, e não se ludibriar com aparentes padrões que possam causar uma confiança superestimada acerca da real capacidade do modelo. Na estatística, para fazer inferências sobre uma população ou modelo, percorre-se o caminho de um cenário específico para o genérico. Porém, nem sempre essas inferências na forma de probabilidades e representadas por intervalos de confiança correspondem a uma verdade absoluta sobre os fatos.

5. As estratégias de reamostragem e validação cruzada:

Toda vez que se realiza uma reamostragem nos dados, altera-se a estatística das distribuições amostradas. Mesmo realizando uma divisão estratificada, ainda assim as medidas coletadas podem variar de forma acentuada. Uma estratégia favorável para diminuir essa variabilidade foi repetir a execução, processando o *Bootstrap* n vezes. Além disso, Kohavi [33] sugeriu que dados reais com 10-*fold* oferecem o melhor *trade-off* viés-variância. Kim [32] também afirmou que o número de repetições no k -*fold* aumenta a precisão das estimativas. Em geral, este procedimento repetitivo contribui com melhores estimativas, além de trazer maior estabilidade para o modelo. Esta estratégia foi adotada nos experimentos computacionais dos modelos por *recall* e *negative log loss*. Apesar da variância ter sido controlada com

as técnicas de reamostragem *Bootstrap* e validação cruzada *k-fold* com repetição, esta estratégia não foi suficiente para controlar as distorções causadas nos modelos por *recall*. Portanto, pode-se observar que o desbalanceamento elevado da classe alvo é um fator importante a se considerar que pode desencadear uma instabilidade nas distribuições e dificuldade em manter a variância e o viés sob controle.

6. A contaminação das amostras:

A aleatoriedade presente entre execuções *Bootstrap* não permitiu garantir com absoluta certeza a não sobreposição das amostras.

7. A estratégia da proposta metodológica:

A estratégia da proposta metodológica é uma possibilidade de lidar com o desafio apresentado. Outras técnicas poderiam ter sido aplicadas na proposta, como a utilização da matriz de custo [35] para lidar com a penalização do falso negativo, e a técnica de *out of bag* [28], eficiente para tratar o desbalanceamento de dados em árvores de decisão. Pode-se concluir que o critério restritivo introduzido na proposta metodológica alcançou o objetivo inicial e se mostrou aderente para o problema da pesquisa.

8. Erros nos registros:

Como todo fenômeno da natureza, o registro dos dados pode ser suscetível a erros involuntários. Erros de medição ou até mesmo de armazenagem, e/ou falha momentânea dos equipamentos utilizados no diagnóstico podem intercorrer a qualquer momento, implicando em erros nas medidas originalmente coletadas para a condução do experimento computacional.

9. Interpretabilidade dos dados:

Na literatura, é frequente encontrar modelos que supostamente resultaram em acurácia elevada mas sem explicação acerca dos principais fatores que acarretaram em suas respostas. Para uma aplicação médica, a transparência é essencial para o médico e para o paciente. Sem uma clareza acerca dos motivos que levaram a um determinado diagnóstico de uma maneira que um ser humano entenda, decisões de tratamento ficam comprometidas. Para auxiliar na interpretabilidade dos dados, técnicas como LIME e Shapley poderiam ter sido aplicadas para explicar as previsões geradas.

6.4 Resumo

Em resumo, este capítulo apresentou os resultados e discussão sobre os dados amostrados. Inicialmente, foi apresentado o resultado preliminar com o modelo de RF por cobertura. Foram selecionados dois modelos candidatos para análise comparativa: o primeiro, com melhor desempenho porém com baixo percentual de amostras dentro do IC; e o segundo, com desempenho inferior porém melhor ajustado para atendimento ao critério proposto do IC. Em seguida, foram apresentados os resultados para quatro cenários de modelo por *negative log loss*: RF ajustado, RF IC 0,95, XGBoost ajustado e XGBoost IC 0,95. Os resultados foram consolidados para auxiliar na análise comparativa entre cada modelo. Pode-se observar, com base na variabilidade das métricas e comportamento dos intervalos de confiança, que os modelos de melhor desempenho (modelos ajustados) nem sempre garantiram a assertividade nos resultados: as métricas sobre o teste extrapolaram os limites do IC calculados sobre a validação. Este comportamento de assimetria e distorção nas distribuições se mostrou ainda mais acentuado quando a medida de cobertura foi selecionada, embora esta tenha sido a medida escolhida para minimizar os falsos negativos representados por casos positivos de risco de ceratocone. Em contrapartida, quando foi selecionada a medida de (*negative log loss*), os modelos IC 0,95 produziram distribuições de maior simetria sobre o teste, atendendo ao critério restritivo introduzido na proposta metodológica. Esta maior verossimilhança entre as medidas de validação e teste fortalece a robustez e previsibilidade dos resultados gerados, justificando sua adoção no esquema metodológico proposto. O capítulo também compara o desempenho dos modelos e conclui que os modelos balizados pelos limites do IC de 95% são mais interessantes para o contexto da pesquisa, dada a sua capacidade de generalização em novas amostras de dados, apesar de apresentarem um desempenho inferior dentre aqueles mais otimizados pela medida de interesse. Também pode-se concluir que modelos otimizados pela medida de desempenho, em geral, apresentam elevada distorção nas distribuições e um baixo percentual de amostras atendendo à restrição do IC quando comparados aos modelos de menor desempenho. Este comportamento observado corrobora com a estratégia de restrição introduzida na proposta metodológica.

7. Considerações Finais

Esta pesquisa teve como objetivo propor uma metodologia experimental para geração de modelos supervisionados de aprendizado de máquina, aplicado ao contexto de detecção de risco de ceratocone. Para validação da proposta metodológica, foram gerados modelos treinados sobre uma amostra de medidas oculares disponibilizada por uma clínica oftalmológica, aplicando os algoritmos de Floresta Aleatória e XGBoost. Como contribuição, geraram-se modelos robustos que alcançaram um desempenho razoável nas medidas de interesse para classificação de olhos normais e com risco de ceratocone. Dada a criticidade do problema, a proposta teve como estratégica a geração de distribuições aplicando as técnicas de *Bootstrap* e *k-fold* para o cálculo dos intervalos de confiança de 95%, sendo este o principal critério restritivo introduzido no esquema metodológico. A medida de *recall* foi a medida prioritariamente selecionada de forma a minimizar a presença de falsos negativos, representada pelos casos com risco de ceratocone incorretamente classificados como olhos normais. Porém, o resultado sugeriu que medidas não naturalmente sensíveis à questão do desbalanceamento, como *negative log loss*, produziram distribuições mais simétricas e com menor distorção quando comparadas aos resultados por *recall*. O resultado também comprovou que modelos com um desempenho inferior, mas que apresentaram maior sobreposição nas distribuições, garantiram maior capacidade de generalização e previsibilidade. Portanto, o critério restritivo introduzido com o IC manifestou-se aderente para a seleção de modelos válidos, justificando suas vantagens em termos da credibilidade alcançada. Como pontos de ressalva, pode-se observar: (i) a introdução de viés para minimizar a variabilidade dos resultados; (ii) o fato de a metodologia proposta ter sido validada a partir de uma única fonte de dados, não sendo possível confirmar a sua generalização em novas amostras; (iii) a ausência de análises comparativas com outras propostas metodológicas. Para trabalhos futuros, a proposta metodológica experimental da pesquisa pode ser ampliada para outros diagnósticos médicos, bem como validada por distintas técnicas e algoritmos de aprendizado de máquina: rede neural profunda, métodos *ensemble*, entre outros.

Referências Bibliográficas

- [1] AFONJA, T, 2017. “Accuracy Paradox. Towards Data Science”. Disponível em: <<https://towardsdatascience.com/tagged/accuracy-paradox/>>. Acesso em: 01 de fev. de 2022.
- [2] AMBRÓSIO, R., JR, LOPES, B. T., FARIA-CORREIA, F., *et al.* “Integration of Scheimpflug-Based Corneal Tomography and Biomechanical Assessments for Enhancing Ectasia Detection”. *Journal of refractive surgery (Thorofare, N.J. : 1995)*, v. 33, n. 7, pp. 434-443, 2017.
- [3] ARBELAEZ, M. C., VERSACI, F., VESTRI, G., *et al.* “Use of a support vector machine for keratoconus and subclinical keratoconus detection by topographic and tomographic data”. *Ophthalmology*, v. 119, n. 11, pp. 2231–2238, 2012.
- [4] ARCE, C. G., FRANCESCONI, C. M. “Topografia da Córnea por Varredura de Fenda de Luz –Orbscan I, II e IIZ. Exames Complementares”. In: *Conselho Brasileiro de Oftalmologia (CBO) - Cirurgia Refrativa - Série Oftalmologia Brasileira*, Rio de Janeiro, Guanabara Koogan, cap. 7, pp. 81-99, 2008.
- [5] ÁVILA, M., SALOMÃO, M., JUNIOR AMBROSIO, S. “Tomografia da Córnea e do Segmento Anterior. Exames Complementares”. In: *Conselho Brasileiro de Oftalmologia (CBO) - Cirurgia Refrativa - Série Oftalmologia Brasileira*, Rio de Janeiro, Guanabara Koogan, cap. 7, pp. 99-106, 2013.
- [6] BATES, S., HASTIE, T., TIBSHIRANI, R. “Cross-validation: what does it estimate and how well does it do it?”, *arXiv preprint arXiv:2104.00673*, 2021.
- [7] BLACKBURN, B. J., JENKINS, M. W., ROLLINS, *et al.* “A Review of Structural and Biomechanical Changes in the Cornea in Aging, Disease, and Photochemical Crosslinking”. *Frontiers in bioengineering and biotechnology*, v. 7, n. 66, 2019.
- [8] BREIMAN, L. “Random Forests”. *Machine Learning*. v. 45, n. 1, pp. 5–32, 2001.

- [9] BUDHOLIYA, K., SHRIVASTAVA, S.K., SHARMA, V. “An optimized XGBoost based diagnostic system for effective prediction of heart disease”. *Journal of King Saud University - Computer and Information Sciences*, 2020.
- [10] CASSIN, B., SOLOMON, S., RUBIN, M. L. “Dictionary of eye terminology”. Gainesville, Triad Publishing Company, pp. 304, 1990.
- [11] CHASTANG, P. J., BORDERIE, V. M., CARVAJAL-GONZALEZ, S., *et al.* “Automated keratoconus detection using the EyeSys videokeratoscope”. *Journal of cataract and refractive surgery*, v. 26, n. 5, pp. 675–683, 2000.
- [12] CHEN, T., GUESTRIN, C. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, 2016.
- [13] CLAESEN, M., DE MOOR, B. “Hyperparameter search in machine learning”. *arXiv preprint arXiv:1502.02127*, 2015.
- [14] DAVSON, H., PERKINS, EDWARD S. ““human eye“. Encyclopedia Britannica”, 2021. Disponível em: <<https://www.britannica.com/science/human-eye/>>. Acesso em: 01 de fev. de 2022.
- [15] DEMIR, S., ORTAK, H., YETER, V., *et al.* “Mapping corneal thickness using dual-Scheimpflug imaging at different stages of keratoconus”. *Cornea*, v.32, n.11, pp. 1470-1474, 2013.
- [16] DOSHI-VELEZ, F., KIM, B. “Towards A Rigorous Science of Interpretable Machine Learning”. *arXiv preprint arXiv:1702.08608*, 2017.
- [17] EFRON, B., TIBSHIRANI, R. “An Introduction to the bootstrap”. CRC press, 1994.
- [18] EFRON, B., TIBSHIRANI, R. “Improvements on Cross-Validation: The .632+ Bootstrap Method”. *Journal of the American Statistical Association*, v. 92, n. 438, pp. 548-560, 1997.
- [19] GEISSER, S. “The Predictive sample reuse method with applications”. *Journal of the American statistical Association*, v. 70, n. 350, pp. 320-328, 1975.
- [20] GINI, C. “Measurement of inequality of incomes”. *The economic journal*, v. 31, n. 121, pp. 124-126, 1921.
- [21] GOKHALE, N. S. “Epidemiology of keratoconus”. *Indian journal of ophthalmology*, v. 61, n. 8, p. 382, 2013.

- [22] GOLDMANN, H., SCHMIDT, T. “Über Applanationstonometrie [Applanation tonometry]”. *Ophthalmologica*, v. 134, n. 4, pp. 221-242, 1957.
- [23] GOLDSTEIN, E. B., ed. “Blackwell handbook of sensation and perception”. John Wiley & Sons, 2008.
- [24] GORDON-SHAAG, A., MILLODOT, M., SHNEOR, E., *et al.* “The genetic and environmental factors for keratoconus”. *BioMed research international*, 2015.
- [25] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. “The elements of statistical learning: data mining, inference, and prediction”, New York: springer, v. 2, pp. 1-758, 2009.
- [26] HIDALGO, I. R., RODRIGUEZ, P., ROZEMA, J.J., *et al.* “Evaluation of a Machine-Learning Classifier for Keratoconus Detection Based on Scheimpflug Tomography”. *Cornea*, v. 35, n. 6, pp. 827-832, 2016.
- [27] HOGAN, M. J., ALVARADO, J. A., WEDDELL, J. E. “Histology of the human eye”. *An atlas and textbook*. Philadelphia: Saunders, 1971.
- [28] JAMES, G., WITTEN, D., HASTIE, *et al.* “An Introduction to Statistical Learning: with Applications in R”, New York: springer, v. 112, 2013.
- [29] JAVADI, M. A., MOTLAGH, B. F., JAFARINASAB, M., *et al.* “Outcomes of penetrating keratoplasty in keratoconus”. *Cornea*, v. 24, n. 8, pp. 941-946, 2005.
- [30] JIMÉNEZ, J. R., ALARCÓN, A., ANERA, R. G., *et al.* “Q-optimized algorithms: theoretical analysis of factors influencing visual quality after myopic corneal refractive surgery”. *Journal of Refractive Surgery*, v. 32, n. 9, pp. 612-617, 2016.
- [31] KENNEDY, H., BOURNE, W. M., DYER, J. A. “A 48-year clinical and epidemiologic study of keratoconus”, *American journal of ophthalmology*, v. 101, n. 3, pp. 267-273, 1986.
- [32] KIM, J. H. “Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap”. *Computational statistics & data analysis*, v. 53, n. 11, pp. 3735-3745, 2009.
- [33] KOHAVI, R. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: *International joint Conference on artificial intelligence. [S.l.: s.n.].* v. 14, pp. 1137–1145, 1995.

- [34] KOHLHAAS, M., SPOERL, E., SCHILDE, T., *et al.* “Biomechanical evidence of the distribution of cross-links in corneas treated with riboflavin and ultraviolet A light”. *Journal of Cataract & Refractive Surgery*, v. 32, n. 2, pp. 279-283, 2006.
- [35] KOLMAN, B., BECK, R. E. “Special Types of Linear Programming Problems”. *Elementary Linear Programming with Applications (Second Edition)*, pp. 295-395, 1995.
- [36] KRACHMER, J. H., FEDER, R. S., BELIN, M. W. “Keratoconus and related noninflammatory corneal thinning disorders”. *Survey of ophthalmology*, v. 28, n. 4, pp. 293-322, 1984.
- [37] KRUMEICH, J. H., DANIEL, J., KNÜLLE, A. “Live-epikeratophakia for keratoconus”. *Journal of Cataract and Refractive Surgery*, v. 24, n. 4, pp. 456-463, 1998.
- [38] KUNCHEVA, L., RODRÍGUEZ, J. “A weighted voting framework for classifiers ensembles”. *Knowledge and Information Systems*, v. 38, n. 2, pp. 259-275, 2014.
- [39] LAVRIC, A., VALENTIN, P. “KeratoDetect: Keratoconus Detection Algorithm Using Convolutional Neural Networks”. *Computational Intelligence and Neuroscience*, 2019.
- [40] LEÃO, E., REN, T. I., LYRA, J. M., *et al.* “Corneal deformation amplitude analysis for keratoconus detection through compensation for intraocular pressure and integration with horizontal thickness profile”. *Computers in Biology and Medicine*, v. 109, pp. 263-271, 2019.
- [41] LI, X., RABINOWITZ, Y. S., RASHEED, K., *et al.* “Longitudinal study of the normal eyes in unilateral keratoconus patients”. *Ophthalmology*, v. 111, n. 3, pp. 440-446, 2004.
- [42] MA, X., SHA, J., WANG, D., *et al.* “Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, v. 31, pp. 24-39, 2018.
- [43] MAEDA, N., KLYCE, S. D., SMOLEK, M. K. “Comparison of methods for detecting keratoconus using videokeratography”. *Archives of Ophthalmology*, v. 113, n. 7, pp. 870-874, 1995.
- [44] MITCHELL, T. M. “Machine Learning”, McGraw-Hill, New York, pp. 154-200, 1997.

- [45] MOHAMADPOUR, M., KHORRAMI-NEJAD, M., KIARUDI, M. Y., *et al.* “Evaluating the ectasia risk score system in cancelled laser In situ keratomileusis candidates”. *Journal of Ophthalmic & Vision Research*, v. 15, n. 4, pp. 481, 2020.
- [46] MOSHIRFAR, M., GARDINER, J. P., SCHLIESSER, J. A., *et al.* “Laser in situ keratomileusis flap complications using mechanical microkeratome versus femtosecond laser: retrospective comparison. *Journal of Cataract & Refractive Surgery*, v. 36, n. 11, pp. 1925-1933, 2010.
- [47] MUNNERLYN, C. R., KOONS, S. J., MARSHALL, J. “Photorefractive keratectomy: a technique for laser refractive surgery”. *Journal of Cataract & Refractive Surgery*, v. 14, n. 1, pp. 46-52, 1988.
- [48] NEYMAN, J. “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability”. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, v. 236, n. 767, pp. 333-380, 1937.
- [49] NOBRE, J., NEVES, R. F. “Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets”. *Expert Systems with Applications*, v. 125, pp. 181-194, 2019.
- [50] OH, E., YOO, T. K., HONG, S. “Artificial Neural Network Approach for Differentiating Open-Angle Glaucoma From Glaucoma Suspect Without a Visual Field Test”. *Investigative Ophthalmology & Visual Science*, v. 56, n. 6, pp. 3957-3966, 2015.
- [51] OSTAPOWICZ, M., ŻBIKOWSKI, K. “Detecting fraudulent accounts on blockchain: a supervised approach”. In: *Web Information Systems Engineering – WISE 2019: 20th International Conference, Hong Kong, China, January 19–22, 2020, Proceedings*. Springer, Cham, p. 18–31, 2020.
- [52] RANDLEMAN, J. B., WOODWARD, M., LYNN, M. J., *et al.* “Risk assessment for ectasia after corneal refractive surgery”. *Ophthalmology*, v. 115, n. 1, pp. 37-50. e4, 2008.
- [53] RANDLEMAN, J. B., DAWSON, D. G., GROSSNIKLAUS, H. E., *et al.* “Depth-dependent cohesive tensile strength in human donor corneas: implications for refractive surgery”. *Journal of refractive surgery*, v. 24, n. 1, pp. S85, 2008.
- [54] RANDLEMAN, J. B., TRATTLER, W. B., STULTING, R. D. “Validation of the Ectasia Risk Score System for preoperative laser in situ keratomileusis screening”. *American journal of ophthalmology*, v. 145, n. 5, p. 813-818. e2, 2008.

- [55] RIBEIRO, M. T., SINGH, S., GUESTIN, C. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144, 2016.
- [56] ROHM, M., TRESP, V., MÜLLER, M., *et al.* “Predicting Visual Acuity by Using Machine Learning in Patients Treated for Neovascular Age-Related Macular Degeneration”. *Ophthalmology*, v. 125, n. 7, pp. 1028-1036, 2018.
- [57] ROZEMA, J. J., RODRÍGUEZ, P., RUIZ HIDALGO, I., *et al.* “SyntEyes KTC: higher order statistical eye model for developing keratoconus”. *Ophthalmic and Physiological Optics*, v. 37, n. 3, pp. 358-365, 2017.
- [58] SAMUEL, A. L. “Some studies in machine learning using the game of checkers”. *II—Recent progress. IBM Journal of research and development*, v. 11, n. 6, pp. 601-617, 1967.
- [59] SANTHIAGO, M. R., SMADJA, D., GOMES, B. F., *et al.* “Association between the percent tissue altered and post-laser in situ keratomileusis ectasia in eyes with normal preoperative topography”. *American journal of ophthalmology*, v. 158, n. 1, pp. 87-95. e1, 2014.
- [60] SHALEV-SHWARTZ, S., BEN-DAVID, S. “Understanding machine learning: From theory to algorithms”. Cambridge university press, 2014.
- [61] SHAPLEY, L. S. “A Value for n-Person Games”. In: *H. Kuhn & A. Tucker (Ed.), Contributions to the Theory of Games (AM-28), Volume II (pp. 307-318)*. Princeton: Princeton University Press, 2016.
- [62] TAN, A. C., GILBERT, D. “An empirical comparison of supervised machine learning techniques in bioinformatics”, 2003.
- [63] TIBSHIRANI, R. J., TIBSHIRANI, R. “A bias correction for the minimum error rate in cross-validation”. *The Annals of Applied Statistics*, v. 3, n. 2, pp. 822-829, 2009.
- [64] TOPOL, E. “Deep medicine: how artificial intelligence can make healthcare human again”. Hachette UK, 2019.
- [65] TORTORA, G. J., DERRICKSON, B. “Corpo Humano-: Fundamentos de Anatomia e Fisiologia”. Artmed Editora, 2016.

- [66] TOUTOUNCHIAN, F., SHANBEHZADEH, J., KHANLARI, M., *et al.* “Detection of keratoconus and suspect keratoconus by machine vision”. In: *International Multiconference of Engineers and Computer Scientists*, pp. 14-16, 2012.
- [67] TWA, M. D., PARTHASARATHY, S., ROBERTS, C., *et al.* “Automated decision tree classification of corneal shape. *Optometry and vision science: official publication of the American Academy of Optometry*, v. 82, n. 12, p. 1038, 2005.
- [68] VINCIGUERRA, R., AMBRÓSIO JR, R., ELSHEIKH, A., *et al.* “Detection of keratoconus with a new biomechanical index”. *Journal of refractive surgery*, v. 32, n. 12, pp. 803-810, 2016.
- [69] WAGER, S. “Cross-validation, risk estimation, and model selection: Comment on a paper by Rosset and Tibshirani”. *Journal of the American Statistical Association*, v. 115, n. 529, pp. 157-160, 2020.
- [70] WOLPERT, D. H. “The lack of a priori distinctions between learning algorithms”. *Neural computation*, v. 8, n. 7, pp. 1341-1390, 1996.
- [71] YOO, T.K., RYU, I.H., LEE, G., *et al.* “Adopting machine learning to automatically identify candidate patients for corneal refractive surgery”. *NPJ digital medicine*, v. 2, n. 1, pp. 1-9, 2019.
- [72] YOUSEFI, S., YOUSEFI, E., TAKAHASHI, H., *et al.* “Keratoconus severity identification using unsupervised machine learning”. *PLoS One*, v. 13, n. 11, pp. e0205998, 2018.