UNIVERSIDADE FEDERAL DO MARANHÃO - UFMA
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA - CCET
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA - PPGEE

WESLEY BATISTA DOMINICES DE ARAUJO

Método de auxílio ao diagnóstico de câncer de próstata utilizando aprendizado de máquina e dados clínicos

São Luís
2024

WESLEY BATISTA DOMINICES DE ARAUJO

Método de auxílio ao diagnóstico de câncer de próstata utilizando aprendizado de máquina e dados clínicos

Tese apresentada à banca examinadora do Programa de Pós-graduação em Engenharia Elétrica da UFMA como um dos pré-requisitos para obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Dr. Ewaldo Eder Carvalho Santana

São Luís

2024

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Araujo, Wesley Batista Dominices de.
Método de auxílio ao diagnóstico de câncer de próstata utilizando aprendizado de máquina e dados clínicos / Wesley Batista Dominices de Araujo. - 2024.
141 f.

Orientador(a): Ewaldo Eder Carvalho Santana.
Tese (Doutorado) - Programa de Pós-graduação em Engenharia Elétrica/ccet, Universidade Federal do Maranhão, São Luís - Ma, 2024.

1. Câncer de Próstata. 2. Aprendizado de Máquina. 3. Triagem. 4. Diagnóstico. I. Santana, Ewaldo Eder Carvalho. II. Título.

WESLEY BATISTA DOMINICES DE ARAUJO

Método de auxílio ao diagnóstico de câncer de próstata utilizando aprendizado de máquina e dados clínicos

Tese apresentada à Banca examinadora do Programa de Pós-graduação em Engenharia Elétrica da UFMA como um dos pré-requisitos para obtenção do título de Doutor em Engenharia Elétrica.

Aprovada em de de 2024.

BANCA EXAMINADORA

_____

Prof. Dr. Ewaldo Éder Carvalho Santana (Orientador)
Universidade Estadual do Maranhão - UEMA

_____

Prof. Dr. Allan Kardec Duailibe Barros Filho (Examinador Interno)
Universidade Federal do Maranhão – UFMA

_____

Prof. Dr. Francisco José da Silva e Silva (Examinador Interno)
Universidade Federal do Maranhão – UFMA

_____

Prof.ª Dra. Claúdia Regina de Andrade Arrais Rosa (Examinadora Interna)
Universidade Federal do Maranhão – UFMA

_____

Prof. Dr. Fábio Manoel França Lobato (Examinador Externo)
Universidade Federal do Oeste do Pará – UFOPA

_____

Prof. Dr. Luís Cláudio Nascimento da Silva (Examinador Externo)
Universidade CEUMA - UNICEUMA

AGRADECIMENTOS

A Deus, pelo dom da vida e por estar sempre presente em minha vida.

A meu orientador, prof. Dr. Ewaldo Eder Carvalho Santana, pela paciência, dedicação, credibilidade e confiança destinadas a mim durante todo o período do doutorado.

À Prefeitura Municipal de São Luís e à Universidade Estadual do Maranhão por me concederem licença para dedicação exclusiva aos meus estudos do doutorado.

Ao Hospital Universitário da Universidade Federal do Maranhão, pela aprovação do projeto que concedeu acesso aos prontuários médicos dos pacientes.

Ao Dr. Giullianno Lopes de Moura e ao médico residente José Arnon Linhares Moraes dos Santos do HUUFMA, pelo auxílio ao projeto.

Aos integrantes da Liga Acadêmica de Urologia da UFMA, Paloma Larissa Arruda Lopes, Wesley do Nascimento Silva, João Pedro Pereira Gonçalves e Felipe Castelo Branco Rocha Silva, pela coleta dos dados utilizados neste trabalho.

À minha esposa, Raísa Oliveira da Silva Araujo e a meus filhos, Enzo e Yan Oliveira Dominices de Araujo, pelo amor, incentivo, compreensão e paciência durante todo o período do doutorado.

A meus pais, João Batista Serra de Araujo e Maria do Socorro Dominices de Araujo, e à minha irmã, Débora Dominices de Araujo, pelo afeto e assistência incondicional ao longo de uma vida, fazendo com que nunca desistisse dos meus sonhos e objetivos.

A todos os que direta e indiretamente contribuíram para a realização deste trabalho.

“O coração do inteligente adquire o conhecimento, e o ouvido dos sábios busca a sabedoria”.

(Provérbios 18:15)

RESUMO

O câncer de próstata, depois do câncer de pele não-melanoma, é o tipo de câncer mais comum entre os homens, e o que causa mais mortes. Para iniciar o diagnóstico de câncer de próstata são utilizados o exame físico (toque retal) e o exame laboratorial (antígeno específico da próstata). Se houver alterações nestes exames, outros podem ser solicitados, como ressonância magnética e biópsia. Atualmente, a biópsia é o único procedimento capaz de confirmar o câncer, tem um custo financeiro elevado, e é um procedimento muito invasivo. Esta Tese propõe um novo método para auxiliar na triagem de pacientes em risco de câncer de próstata. O método foi desenvolvido com base em variáveis clínicas (idade, raça, hipertensão arterial sistêmica, diabetes *mellitus*, tabagismo, etilismo, toque retal e PSA total) de 274 pacientes, dos quais 137 têm câncer e 137 não têm, conforme obtido dos prontuários médicos. Os dados foram analisados utilizando diversos algoritmos de aprendizado de máquina, como Redes Neurais Artificiais, Máquina de Vetor de Suporte, *Naive Bayes*, *K*-vizinhos mais próximos e árvore de decisão, para classificar as amostras quanto à presença ou ausência de câncer de próstata. O método foi avaliado com base em métricas de desempenho, incluindo acurácia, sensibilidade, especificidade e área sob a curva ROC. Para aumentar a confiabilidade dos resultados e a capacidade de generalização do classificador, foi utilizada a técnica de validação cruzada *10-fold*. O melhor desempenho foi obtido com o modelo *Naive Bayes*, resultando em uma acurácia de 89,09%, sensibilidade de 92%, especificidade de 86,67% e uma Área sob a curva ROC de 0,9187.

Palavras-chave: Câncer de próstata; Aprendizado de máquina; Triagem; Diagnóstico.

ABSTRACT

Prostate cancer, after non-melanoma skin cancer, is the most common type of cancer among men, and the one that causes the most deaths. To begin the diagnosis of prostate cancer, a physical examination (digital rectal exam) and laboratory exam (prostate-specific antigen) are used. If there are changes in these tests, other tests may be requested, such as resonance magnetic imaging and biopsy. Currently, biopsy is the only procedure capable of confirming cancer, it has a high financial cost and is a very invasive procedure. This thesis proposes a new method to aid in the screening of patients at risk for prostate cancer. The method was developed based on clinical variables (age, race, systemic arterial hypertension, diabetes mellitus, smoking, alcoholism, digital rectal examination, and total PSA) of 274 patients, of which 137 have cancer and 137 do not, as obtained from medical records. The data were analyzed using several machine learning algorithms, such as Artificial Neural Networks, Support Vector Machine, Naive Bayes, K-nearest neighbors, and decision tree, to classify the samples according to the presence or absence of prostate cancer. The method was evaluated based on performance metrics, including accuracy, sensitivity, specificity, and area under the ROC curve. To increase the reliability of the results and the generalization capacity of the classifier, the 10-fold cross-validation technique was used. The best performance was obtained with the Naive Bayes model, resulting in an accuracy of 89.09%, sensitivity of 92%, specificity of 86.67% and an Area under the ROC curve of 0.9187.

Keywords: Prostate cancer; Machine learning; Screening; Diagnosis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASAP	*Atypical Small Acinar Proliferation*
AUC	*Area Under the Curve*
BOW	*Bag-of-Words*
BRCA1	*Breast Cancer 1*
BRCA2	*Breast Cancer 2*
CAAE	Certificado de Apresentação de Apreciação Ética
CAD	*Computer-Aided Diagnosis*
CEP	Comitê de Ética em Pesquisa
COMIC	Comissão Científica
CSPCa	*Clinically Significant Prostate Cancer*
DCNN	*Deep Convolutional Neural Networks*
DM	*Diabetes Mellitus*
DT	*Decision Trees*
DWI	*Diffusion-Weighted Imaging*
EAS	Elementos Anormais do Sedimento
ERSPCRC	*European Randomized Study Prostate Cancer Risk Calculator*
FN	Falso Negativo
FP	Falso Positivo
GDI	*Gini Diversity Index*
HAS	Hipertensão Arterial Sistêmica
HPB	Hiperplasia Prostática Benigna
HU-UFMA	Hospital Universitário da Universidade Federal do Maranhão
IA	Inteligência Artificial
INCA	Instituto Nacional do Câncer
IRM	Imagem por Ressonância Magnética
IRMmp	Imagem por Ressonância Magnética multiparamétrica
KNN	*K-Nearest Neighbor*
KPCRC-HG	*Korean Prostate Cancer Risk Calculator for High-Grade*
LDA	*Linear Discriminant Analysis*
MAP	*Maximum A Posteriori*
ML	*Machine Learning*

PCA	*Prostate Cancer*
PCPTRC-HG	*Prostate Cancer Prevention Trial Risk Calculator for High-Grade*
PET	*Positron Emission Tomography*
PHI	*Prostate Health Index*
PIA	*Proliferative Inflammatory Atrophy*
PIN	*Prostatic Intraepithelial Neoplasia*
PI-RADS	*Prostate Imaging Reporting and Data System*
PSA	*Prostate-Specific Antigen*
PSAD	*Prostate-Specific Antigen Density*
PSMA	*Prostate-Specific Membrane Antigen*
QDA	*Quadratic Discriminant Analysis*
ReLU	*Rectified Linear Unit*
RNA	Rede Neural Artificial
ROC	*Receiver Operating Characteristic*
SIFT	*Scale-Invariant Feature Transform*
SVM	*Support Vector Machines*
TFP	Taxa de Falsos Positivos
TRUS	*Transrectal Ultrasound*
TVP	Taxa de Verdadeiros Positivos
TZV	*Transition Zone Volume*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

O câncer de próstata – *Prostate Cancer* (PCa) é um dos tipos de câncer mais prevalentes entre os homens em nível mundial, com cerca de 1 milhão de casos por ano (KIM et al., 2022). Adicionalmente, sua incidência em nível global está crescendo, sendo que mais de 80% dos homens diagnosticados ainda não apresentam metástase (LEE et al., 2021). A estimativa mais recente nos Estados Unidos para 2024 indicou que o PCa é o segundo tipo mais comum entre os homens, com aproximadamente 299.010 novos casos e 35.250 mortes previstas (NATIONAL CANCER INSTITUTE, 2024). No Brasil, a situação não é diferente, somente no ano de 2021 houve 16.301 mortes por PCa, e a estimativa de novos casos para o ano de 2024 é de 71.740, que corresponde a 30,0% dos tumores incidentes no gênero masculino (INCA, 2023). No Estado do Maranhão, o PCa também é altamente prevalente, resultando em um número significativo de óbitos. Entre 2010 e 2021, foram registradas 4.276 mortes. Apenas em 2021, ano da última estatística divulgada até o momento da redação desta Tese, houve 362 mortes atribuídas a esse tipo de câncer (INCA, 2021).

O aumento da expectativa de vida, a melhoria e a evolução dos métodos diagnósticos e da qualidade dos sistemas de informação do país, bem como a ocorrência de sobrediagnóstico, em função da disseminação do rastreamento do PCa com o antígeno específico da próstata - *Prostate-Specific Antigen* (PSA) e toque retal, podem explicar o aumento das taxas de incidência (observadas pela análise da série histórica de incidência dos registros de câncer de base populacional) ao longo dos anos (INCA, 2022).

Alguns fatores de risco podem influenciar para que uma pessoa possa contrair o PCa, como idade, raça/etnia, histórico familiar, alterações genéticas herdadas, dieta, obesidade, tabagismo, exposição química, inflamação na próstata, infecções transmitidas sexualmente e vasectomia. Ter um fator de risco, ou mesmo vários, não significa que a paciente terá a doença. Muitas pessoas com um ou mais fatores de risco podem nunca ter câncer, enquanto outras que sofrem de câncer podem ter poucos ou nenhum fator de risco conhecido. Pesquisadores descobriram vários fatores que podem afetar o risco de um homem contrair PCa (NATIONAL CANCER INSTITUTE, 2023).

Para iniciar o diagnóstico de PCa são utilizados basicamente o exame clínico,

chamado de exame de toque retal e o exame laboratorial, que é a dosagem do PSA, iniciado em 1986 (COSMA et al., 2021). Se houver alterações em ambos os exames, outros exames poderão ser solicitados, como endoscópios, radiológicos e Imagem por Ressonância Magnética (IRM). Mas nenhum desses exames tem 100% de precisão, levando ao paciente a ter que fazê-los, às vezes, desnecessariamente, e com um alto custo financeiro.

Atualmente, a biópsia prostática é o único procedimento capaz de diagnosticar o PCa, mas tem um custo financeiro elevado, é muito invasiva, dolorida, e tem 5% de chance de desenvolver infecções como urosepse. Mais de 60% dos resultados da biópsia da próstata são negativos nos Estados Unidos a cada ano, esta é uma razão para filtrar a indicação de biópsia, diminuindo as biópsias negativas e aumentando a acurácia do exame. Mesmo usando ultrassom transretal - *Transrectal Ultrasound* (TRUS), não detecta até 30% dos cânceres de próstata clinicamente significativos, e cerca de 55% dão resultado negativo, mesmo que o paciente esteja com câncer de próstata. Uma biópsia da próstata negativa não significa, portanto, necessariamente uma próstata livre de câncer, pois o câncer de próstata pode estar presente nas partes anteriores da zona periférica ou de transição que são inacessíveis por esta via (COSMA et al., 2021).

Outros exames de imagem, além do TRUS, que têm sido utilizados no diagnóstico de PCa, são a IRM e a tomografia computadorizada. Contudo, não fornecem informações sobre a composição química, apresentam baixa resolução e incorrem em altos custos (CORREAS et al., 2020; HICKS et al., 2018; STABILE et al., 2020). Assim, satisfazer os critérios de alta precisão, sensibilidade e invasão mínima para o diagnóstico do câncer de próstata ainda é difícil. A interpretação correta dos exames de imagem requer uma observação cuidadosa. Entender o resultado do TRUS, da IRM ou da tomografia requer experiência e, algumas vezes, o mesmo exame pode resultar em diagnósticos diferentes dependendo do especialista. Devido às complexas patologias e mudanças sutis de textura, os médicos podem cometer erros mesmo quando bem treinados.

Diante disso, é importante o uso de métodos alternativos, que servirão para auxiliar o médico na tomada de decisão em relação à indicação ou não da biópsia da próstata. Esses métodos são baseados no aprendizado de máquina - *Machine Learning* (ML). O ML é um ramo da Inteligência Artificial (IA) que se baseia na ideia

de o sistema aprender um padrão a partir de uma base de dados, usando ferramentas probabilísticas e estatísticas, e tomando decisões ou previsões sobre os novos dados (NASRABADI, 2007; GOLDBERG; HOLLAND, 1988; MICHALSKI; CARBONELL; MITCHELL, 2013).

Em ML, computadores são programados para aprender com a experiência passada. Para isso, empregam um princípio de inferência denominado indução, no qual se obtêm conclusões genéricas a partir de um conjunto particular de exemplos. Desta maneira, algoritmos de ML aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de dados que representam instâncias do problema a ser resolvido. Esses dados formam um conjunto, denominado conjunto de dados. ML é uma das áreas de pesquisa da computação que mais tem crescido nos últimos anos. Diferentes algoritmos de ML, diferentes formas de utilizar os algoritmos existentes e adaptações de algoritmos são continuamente propostos. Além disso, surgem a todo momento novas variações nas características dos problemas reais a serem tratados (FACELI et al., 2021).

Em tarefa de previsão, a meta é encontrar uma função a partir dos dados de treinamento que possa ser utilizada para prever um rótulo ou valor que caracterize um novo exemplo, com base nos valores de seus atributos ou características de entrada. Para isso, cada objeto do conjunto de treinamento deve possuir atributos de entrada e de saída. Os algoritmos de ML utilizados nesta tarefa induzem modelos preditivos. Esses algoritmos seguem o paradigma de aprendizado supervisionado. O termo supervisionado vem da simulação da presença de um “supervisor externo”, que conhece a saída (rótulo) desejada para cada exemplo (conjunto de valores para os atributos de entrada). Com isso, o supervisor externo pode avaliar a capacidade da hipótese induzida de predizer o valor de saída para novos exemplos ou observações (FACELI et al., 2021). Os algoritmos de ML supervisionado mais comuns incluem Rede Neural Artificial (RNA), Máquinas de Vetores de Suporte - *Support Vector Machine* (SVM), K-vizinhos mais próximos - *K-Nearest Neighbor* (KNN), *Naive Bayes* e Árvore de decisão.

Métodos de diagnóstico utilizando algoritmos de ML têm sido propostos com o objetivo de auxiliar no diagnóstico precoce do câncer de próstata. Eles também são chamados de esquemas CAD (*Computer-Aided Diagnosis*), e têm sido desenvolvidos por vários grupos de pesquisa no mundo inteiro. O diagnóstico auxiliado por

computador é aquele no qual o profissional da saúde usa os resultados de uma análise computadorizada de imagens, sinais ou exames médicos como uma “segunda opinião” na detecção de lesões, características em sinais, imagens e exames, e na elaboração do diagnóstico de câncer. Os esquemas CAD, aliados aos algoritmos de ML, têm representado uma importante ferramenta no auxílio ao diagnóstico médico em diversas aplicações. A eficácia dessas técnicas tem mostrado sucesso no diagnóstico precoce de vários tipos de câncer, tais como câncer de mama (ALJUAID et al., 2022), câncer de pâncreas (CICHOSZ et al., 2024), câncer de pulmão (WANI; KUMAR; BEDI, 2024), câncer de ovário (MYSONA et al., 2023) e câncer de próstata (WANG et al., 2017; LIU et al., 2019; LIU et al., 2020; PARK et al., 2017; YOO et al., 2019; CHEN et al., 2021; LU et al., 2023; ZHENG et al., 2019).

A principal motivação desta Tese de doutorado é a escassez de métodos na literatura e em publicações que utilizem variáveis clínicas não-invasivas dos pacientes para estudar suas correlações e classificá-las, a fim de auxiliar na triagem para a biópsia do câncer de próstata. Além disso, devido à alta mortalidade associada à patologia, é crucial desenvolver um novo método de baixo custo para auxiliar no diagnóstico precoce e aumentar a sobrevida dos pacientes. O baixo custo refere-se à eliminação da necessidade de realizar certos exames, como hemograma, coagulograma, Elementos Anormais do Sedimento (EAS), urocultura, ureia, creatinina, biópsia de próstata guiada pelo TRUS e anatomopatológico de biópsia prostática.

Esta Tese sugere um método para auxiliar ao diagnóstico do câncer de próstata, baseado em um novo método que utiliza variáveis clínicas dos pacientes, em conjunto com algoritmos de ML, para saber qual será o mais eficiente em fazer a previsão de novos dados entre câncer ou normal, gerando assim, o modelo ótimo a ser utilizado pelo método.

1.1 Objetivos

1.1.1 Objetivo Geral

Desenvolver um método de auxílio ao diagnóstico de câncer de próstata utilizando modelos de aprendizado de máquina aplicados em variáveis clínicas.

1.1.2 Objetivos Específicos

- Aplicar algoritmos de aprendizado de máquina, como SVM, *Naive Bayes*, KNN, Árvores de Decisão e RNA, nas variáveis clínicas: idade, raça, diabetes *mellitus*, etilismo, tabagismo, hipertensão arterial sistêmica, exame de toque retal e PSA total;
- Avaliar o método proposto utilizando as medidas de desempenho: acurácia, sensibilidade, especificidade, e área sob a curva característica de operação do receptor - *Receiver Operating Characteristic* (ROC);
- Obter alto desempenho nas métricas de desempenho ao implementar modelos de aprendizado de máquina;
- Comparar e analisar o desempenho do método proposto com outros métodos já consolidados na literatura;
- Desenvolver um protótipo de aplicativo para utilização do método proposto.

1.2 Contribuições

Esta Tese apresenta como principais contribuições:

- Desenvolvimento de um novo método para auxiliar na triagem do câncer de próstata, utilizando variáveis clínicas, como idade, raça, diabetes *mellitus*, etilismo, tabagismo, hipertensão arterial sistêmica, exame de toque retal e PSA total, em conjunto com algoritmos de Aprendizado de Máquina;
- Elaboração de uma nova estratégia para implementar um método otimizado de auxílio ao diagnóstico precoce de câncer de próstata, que seja de baixo custo e utilize variáveis não-invasivas.

1.3 Organização do Trabalho

Esta Tese será composta de mais cinco capítulos, conforme descrição sumária a seguir.

No Capítulo 2 será mostrada a fundamentação teórica da literatura necessária ao desenvolvimento do método proposto. Serão apresentados os conceitos sobre o

câncer de próstata, aprendizado de máquina: algoritmos, seleção de modelos, otimização e medidas de desempenho.

No Capítulo 3 serão apresentados alguns trabalhos relacionados ao câncer de próstata.

No Capítulo 4 serão descritos a metodologia proposta juntamente com os materiais utilizados em cada etapa deste trabalho.

No Capítulo 5 apresentam-se os resultados obtidos, discussões, análise das técnicas utilizadas e a validação do método pelas medidas de desempenho: acurácia, sensibilidade, especificidade e área sob a curva ROC.

No Capítulo 6, as conclusões sobre o trabalho, mostrando a eficiência do método proposto e as propostas de trabalhos futuros.

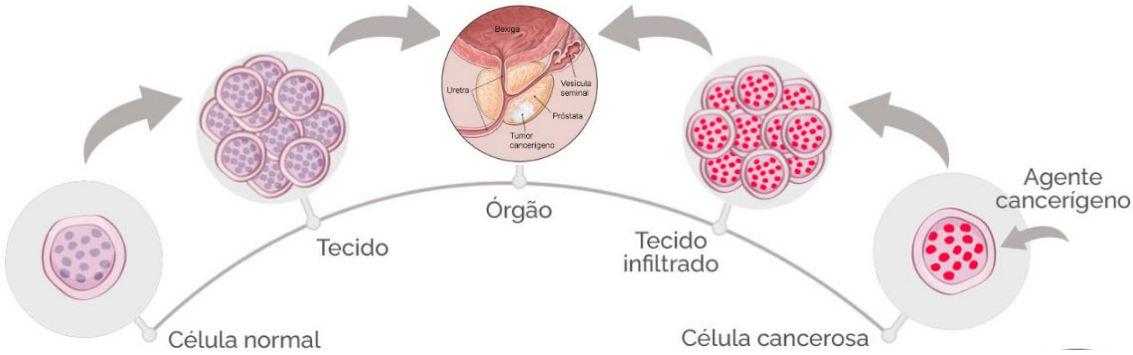
2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica utilizada no desenvolvimento desta Tese, sendo necessária para a compreensão das técnicas, materiais e conceitos a serem utilizados na metodologia proposta.

2.1 Câncer de próstata

O corpo humano é composto de trilhões de células que, ao longo de sua vida, normalmente crescem e se dividem conforme necessário. Quando as células são anormais ou envelhecem, elas geralmente morrem. O câncer começa quando algo dá errado nesse processo e suas células continuam produzindo novas células e as velhas ou anormais não morrem quando deveriam. À medida que as células cancerígenas crescem fora de controle, elas podem expulsar as células normais. Isso torna difícil para o seu corpo funcionar da maneira que deveria. A Figura 1, a seguir, mostra este processo.

Figura 1 – Processo de infiltração de um órgão por células cancerosas



Fonte: adaptado de INCA (2023)

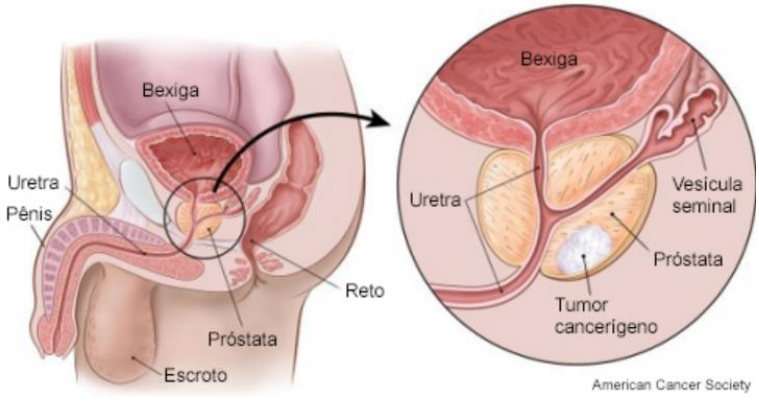
Existem duas categorias principais de câncer, os cânceres hematológicos (sangue) são cânceres das células sanguíneas, incluindo leucemia, linfoma e mieloma múltiplo. E os cânceres de tumor sólido que são cânceres de qualquer um dos outros órgãos ou tecidos do corpo. Os tumores sólidos mais comuns são os cânceres de mama, próstata, pulmão e colorretal. Esses cânceres são semelhantes em alguns aspectos, mas podem ser diferentes na maneira como crescem, se espalham e respondem ao tratamento (NATIONAL CANCER INSTITUTE, 2022).

Um tumor é um nódulo. Alguns nódulos são câncer, mas muitos não são. Nódulos que não são câncer são chamados de benignos, nódulos que são câncer são chamados de malignos. O que torna o câncer diferente é que ele pode se espalhar para outras partes do corpo, enquanto os tumores benignos não. As células cancerosas podem se desprender do local onde o câncer começou. Essas células podem viajar para outras partes do corpo e acabar nos gânglios linfáticos ou em outros órgãos do corpo, causando problemas nas funções normais. O câncer pode se espalhar de onde começou (local primário) para outras partes do corpo. Essa disseminação do câncer para uma nova parte do corpo é chamada de metástase.

As células cancerosas se desenvolvem devido a múltiplas mudanças em seus genes. Essas alterações podem ter muitas causas possíveis. Hábitos de estilo de vida, genes que a pessoa recebe de seus pais e exposição a agentes causadores de câncer no ambiente podem desempenhar um papel. Muitas vezes, não há uma causa óbvia.

A próstata está abaixo da bexiga (o órgão oco onde a urina é armazenada) e na frente do reto (a última parte do intestino). Logo atrás da próstata estão as glândulas chamadas vesículas seminais, que produzem a maior parte do fluido para o sêmen. A uretra, que é o tubo que transporta a urina e o sêmen para fora do corpo através do pênis, passa pelo centro da próstata. O tamanho da próstata pode mudar à medida que o homem envelhece. Em homens mais jovens, é do tamanho de uma noz, mas pode ser muito maior em homens mais velhos. A Figura 2, a seguir, exibe esses órgãos.

Figura 2 – Visualização da próstata e de órgãos próximos a ela



Fonte: adaptado de NATIONAL CANCER INSTITUTE (2021)

O câncer de próstata começa quando as células da próstata começam a

crescer fora de controle. A próstata é uma glândula encontrada apenas nos homens. Quase todos os cânceres de próstata são adenocarcinomas. Esses cânceres se desenvolvem a partir das células da glândula. Outros tipos de câncer, que são mais raros, e podem começar na próstata incluem (NATIONAL CANCER INSTITUTE, 2021):

- Carcinomas de pequenas células;
- Tumores neuroendócrinos (exceto carcinomas de pequenas células);
- Carcinomas de células transicionais;
- Sarcomas.

Alguns tipos de câncer de próstata crescem e se espalham rapidamente, mas a maioria cresce lentamente. De fato, estudos de autópsia mostram que muitos homens mais velhos, e até alguns homens mais jovens, que morreram de outras causas também tiveram câncer de próstata que nunca os afetou durante suas vidas. Em muitos casos, nem eles nem seus médicos sabiam que eles tinham.

Os cânceres de próstata mais avançados às vezes podem causar sintomas, como: problemas para urinar, incluindo fluxo urinário lento ou fraco ou necessidade de urinar com mais frequência, especialmente à noite; sangue na urina ou sêmen; problemas para obter uma ereção (disfunção erétil); dor nos quadris, costas (coluna), tórax (costelas) ou outras áreas de câncer que se espalhou para os ossos; fraqueza ou dormência nas pernas ou pés, ou até mesmo perda de controle da bexiga ou do intestino devido ao câncer pressionando a medula espinhal.

A maioria desses problemas é mais provável de ser causada por algo diferente do câncer de próstata. Por exemplo, a dificuldade para urinar é muito mais frequentemente causada por HPB, um crescimento não cancerígeno da próstata. Alguns homens podem precisar de mais exames para verificar se há câncer de próstata.

2.1.1 Fatores de risco

Um fator de risco é qualquer variável que aumente o risco de contrair uma doença como o câncer. Diferentes tipos de câncer têm diferentes fatores de risco. Alguns fatores de risco, como fumar ou consumir muita bebida alcoólica, podem ser

controlados, pois o indivíduo decide ser exposto ou não ao risco. Outros, como a idade, raça e histórico familiar de uma pessoa, não podem ser alterados. Mas ter um fator de risco, ou mesmo vários, não significa que a pessoa terá a doença. Muitas pessoas com um ou mais fatores de risco nunca sofrem de câncer, enquanto outras que sofrem de câncer podem ter poucos ou nenhum fator de risco conhecido. Pesquisadores descobriram vários fatores que podem afetar o risco de um homem contrair câncer de próstata (NATIONAL CANCER INSTITUTE, 2023), dentre os quais se destacam:

- **Idade**, o câncer de próstata é raro em homens com menos de 40 anos, mas a chance de ter câncer de próstata aumenta rapidamente após os 50 anos de idade. Cerca de 60% dos casos de câncer de próstata são encontrados em homens com mais de 65 anos de idade;
- **Raça ou etnia**, o câncer de próstata se desenvolve com mais frequência em homens afro-americanos e nos homens do Caribe de ascendência africana do que em homens de outras raças. E quando se desenvolve nesses homens, eles tendem a ser mais jovens. O câncer de próstata ocorre com menos frequência em homens asiático-americanos e hispânicos ou latinos do que em brancos não-hispânicos. As razões para essas diferenças raciais e étnicas ainda não são claras;
- **Histórico familiar**, pois, em alguns casos, pode haver um fator genético ou herdado. Ainda assim, a maioria dos cânceres de próstata ocorre em homens sem histórico familiar da doença. Ter um pai ou irmão com câncer de próstata mais que dobram o risco de um homem desenvolver esta doença. O risco é muito maior para homens com vários parentes afetados, principalmente se os parentes eram jovens quando o câncer foi descoberto;
- **Genética**, alterações genéticas herdadas (mutações) parecem aumentar o risco de câncer de próstata, mas provavelmente representam apenas uma pequena porcentagem dos casos em geral. Mutações herdadas dos genes BRCA1 ou BRCA2, que estão ligadas a um risco aumentado de câncer de mama e ovário em algumas famílias, também podem aumentar o risco de câncer de próstata em homens (especialmente mutações no BRCA2). Homens com síndrome de Lynch (também conhecido como câncer colorretal não polipose hereditário), uma condição causada por alterações

genéticas herdadas, têm um risco aumentado de vários tipos de câncer, incluindo câncer de próstata;

- **Dieta**, o papel exato da dieta no câncer de próstata não está claro, mas vários fatores foram estudados. Homens que comem muita carne vermelha ou alimentos ricos em gordura (especialmente produtos lácteos) parecem ter uma chance ligeiramente maior de contrair câncer de próstata. Alguns estudos sugeriram que homens que consomem muito cálcio (através de alimentos ou suplementos) podem ter um risco maior de desenvolver câncer de próstata. Os laticínios (com alto teor de cálcio) também podem aumentar o risco. Mas a maioria dos estudos não encontrou essa ligação com os níveis de cálcio encontrados na dieta média, e é importante observar que é conhecido que o cálcio tem outros benefícios importantes para a saúde;
- **Obesidade**, ser obeso não parece aumentar o risco geral de contrair câncer de próstata. Alguns estudos descobriram que homens obesos têm um risco menor de contrair uma forma de baixo grau (crescimento mais lento) da doença, mas um risco maior de contrair câncer de próstata mais agressivo (crescimento mais rápido). As razões para isso não são claras. Outros estudos também descobriram que homens obesos podem estar em maior risco de ter câncer de próstata mais avançado e de morrer de câncer de próstata, mas nem todos os estudos descobriram isso;
- **Tabagismo**, a maioria dos estudos não encontrou uma ligação entre fumar e desenvolver câncer de próstata. Algumas pesquisas associaram o tabagismo a um possível pequeno aumento do risco de morrer de câncer de próstata, mas essa descoberta precisa ser confirmada por outros estudos;
- **Exposição química,** existem algumas evidências de que os bombeiros podem ser expostos a produtos químicos que podem aumentar o risco de câncer de próstata. Alguns estudos sugeriram uma possível ligação entre a exposição ao Agente Laranja, uma substância química amplamente utilizada durante a Guerra do Vietnã, e o risco de câncer de próstata, embora nem todos os estudos tenham encontrado essa ligação. A Academia Nacional de Medicina considera que há “evidência

limitada/sugestiva” de uma ligação entre a exposição ao Agente Laranja e o câncer de próstata;

- **Inflamação da próstata**, alguns estudos sugeriram que a prostatite (inflamação da próstata) pode estar ligada a um risco aumentado de câncer de próstata, mas outros estudos não encontraram tal ligação. A inflamação é frequentemente observada em amostras de tecido da próstata que também contêm câncer. A ligação entre os dois ainda não está clara, e esta é uma área ativa de pesquisa;
- **Infecções sexualmente transmissíveis,** os pesquisadores verificaram se as infecções sexualmente transmissíveis (como gonorreia ou clamídia) podem aumentar o risco de câncer de próstata, porque podem levar à inflamação da próstata. Até agora, os estudos não concordaram e nenhuma conclusão firme foi alcançada;
- **Vasectomia**, alguns estudos sugeriram que homens que fizeram vasectomia (pequena cirurgia para tornar os homens inférteis) têm um risco ligeiramente aumentado de câncer de próstata, mas outros estudos não encontraram isso. A pesquisa sobre essa possível ligação ainda está em andamento.

2.1.2 Diagnóstico precoce

A maioria dos cânceres de próstata é descoberta precocemente, por meio de triagem, que é um teste para encontrar câncer nas pessoas antes que elas apresentem sintomas. Para alguns tipos de câncer, essa triagem pode ajudar a encontrar cânceres em um estágio inicial, quando é provável que sejam mais fáceis de tratar. Para a triagem é verificado o nível do PSA no sangue de um homem, e através do exame de toque retal.

O PSA é uma proteína produzida pelas células da próstata (tanto células normais quanto células cancerosas). O PSA é encontrado principalmente no sêmen, mas uma pequena quantidade também é encontrada no sangue.

O nível de PSA no sangue é medido em unidades chamadas nanogramas por mililitro (ng/ml). A chance de ter câncer de próstata aumenta à medida que o nível de PSA aumenta, mas não há um ponto de corte definido que possa dizer com certeza

se um homem tem ou não câncer de próstata. Muitos médicos usam um ponto de corte de PSA de 4 ng/ml ou superior ao decidir se um homem pode precisar de mais testes, enquanto outros podem recomendar que comece em um nível mais baixo, como 2,5 ng/ml ou 3 ng/ml.

Quando o câncer de próstata se desenvolve, o nível de PSA geralmente fica acima de 4 ng/ml. Ainda assim, um nível abaixo de 4 ng/ml não é garantia de que um homem não tenha câncer. Cerca de 15% dos homens com PSA abaixo de 4 ng/ml terão câncer de próstata se uma biópsia for feita. Homens com um nível de PSA entre 4 ng/ml e 10 ng/ml (muitas vezes chamado de “faixa limítrofe”) têm cerca de 25% de chance de ter câncer de próstata. Se o PSA for superior a 10 ng/ml, a chance de ter câncer de próstata é superior a 50%.

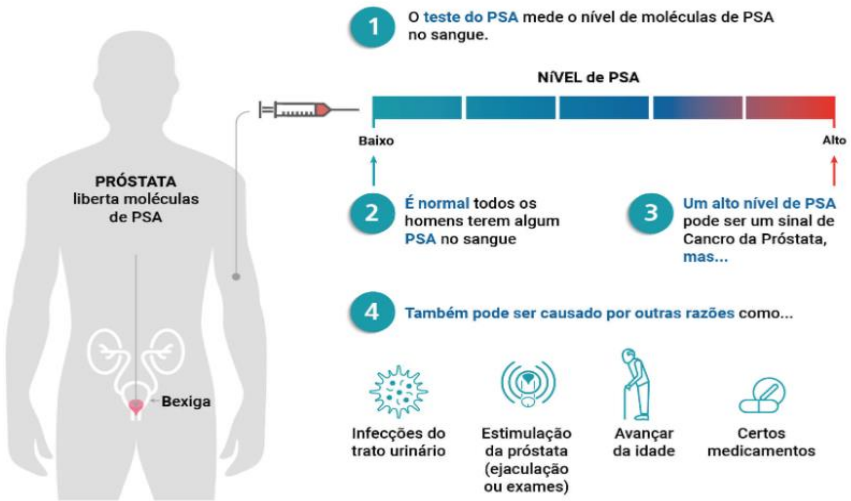
Uma razão pela qual é difícil usar um ponto de corte definido com o teste de PSA ao procurar câncer de próstata é que vários fatores além do câncer também podem afetar os níveis de PSA. A Figura 3 mostra o nível de PSA e alguns fatores que podem alterá-lo. Alguns fatores que podem elevar os níveis de PSA, incluem (NATIONAL CANCER INSTITUTE, 2020):

- **Próstata aumentada**, condições como a Hiperplasia Prostática Benigna (HPB), um aumento não cancerígeno da próstata que afeta muitos homens à medida que envelhecem, podem aumentar os níveis de PSA;
- **Idade avançada**, os níveis de PSA normalmente aumentam lentamente à medida que o indivíduo envelhece, mesmo que não tenha nenhuma anormalidade na próstata;
- **Prostatite**, é uma infecção ou inflamação da próstata, que pode elevar os níveis de PSA;
- **Ejaculação**, isso pode fazer o PSA subir por um curto período de tempo. É por isso que alguns médicos sugerem que os homens se abstenham de ejacular por um ou dois dias antes do teste;
- **Andar de bicicleta**, alguns estudos sugeriram que andar de bicicleta pode aumentar os níveis de PSA por um curto período de tempo, possivelmente porque o assento pressiona a próstata, embora nem todos os estudos tenham encontrado isso;
- **Procedimentos urológicos**, alguns procedimentos realizados em um consultório médico que afetam a próstata, como uma biópsia de próstata

ou cistoscopia, podem elevar os níveis de PSA por um curto período de tempo. Alguns estudos sugeriram que um exame de toque retal pode aumentar ligeiramente os níveis de PSA, embora outros estudos não tenham encontrado isso. Ainda assim, se um teste de PSA e um exame de toque retal estiverem sendo feitos durante uma consulta médica, alguns médicos aconselham a coleta de sangue para o PSA antes de fazer o exame de toque retal, apenas por precaução;

- **Medicamentos**, tomar hormônios masculinos como a testosterona, ou outros medicamentos que aumentam os níveis de testosterona, pode causar um aumento no PSA.

Figura 3 – Nível e alteração de PSA



Fonte: www.institutodaprostata.com

Existem alguns fatores que podem diminuir os níveis de PSA, mesmo que um homem tenha câncer de próstata, (NATIONAL CANCER INSTITUTE, 2020):

- **Inibidores da 5-alfa redutase**, certos medicamentos usados para tratar HPB ou sintomas urinários, como finasterida (Proscar ou Propecia) ou dutasterida (Avodart), podem diminuir os níveis de PSA. Esses medicamentos também podem afetar o risco de câncer de próstata;
- **Misturas de ervas**, algumas misturas vendidas como suplementos dietéticos podem mascarar um alto nível de PSA;
- **Outros medicamentos**, algumas pesquisas sugeriram que o uso prolongado de certos medicamentos, como aspirina, estatinas

(medicamentos para baixar o colesterol) e diuréticos tiazídicos, como a hidroclorotiazida, podem reduzir os níveis de PSA. Mais pesquisas são necessárias para confirmar esses achados.

Para homens que podem ser rastreados para câncer de próstata, nem sempre está claro se a redução do PSA é útil. Em alguns casos, o fator que reduz o PSA também pode reduzir o risco de câncer de próstata em um homem. Mas em outros casos, pode diminuir o nível de PSA sem afetar o risco de câncer de um homem. Na verdade, isso pode ser prejudicial, se reduzir o PSA de um nível anormal para um normal, pois pode resultar na não detecção de um câncer. É por isso que é importante informar ao médico sobre qualquer coisa que possa afetar o nível de PSA.

O nível de PSA de um teste de triagem às vezes é chamado de PSA total, porque inclui as diferentes formas de PSA. Existem alguns tipos especiais de testes de PSA, que incluem (NATIONAL CANCER INSTITUTE, 2020):

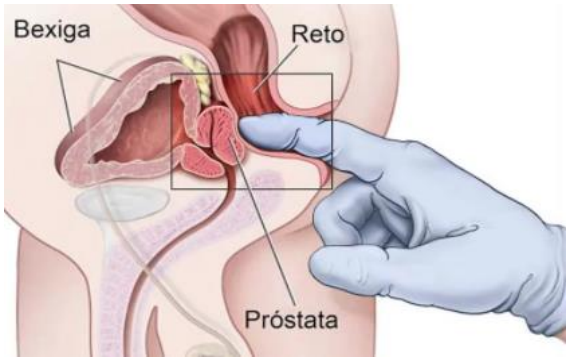
- **PSA livre**, o PSA ocorre em duas formas principais no sangue. Uma forma está ligada às proteínas do sangue, enquanto a outra circula livre (desvinculada). O PSA livre é a proporção de quanto PSA circula livre em comparação com o nível total de PSA. A porcentagem de PSA livre é menor em homens com câncer de próstata do que em homens que não têm. Se o resultado do teste de PSA estiver na faixa limítrofe (entre 4 ng/ml e 10 ng/ml), o percentual de PSA livre pode ser usado para ajudar a decidir se deve-se fazer uma biópsia de próstata. Um PSA livre menor significa que a chance de ter câncer de próstata é maior e provavelmente deverá ser feita uma biópsia;
- **PSA Complexado**, este teste mede diretamente a quantidade de PSA que está ligada a outras proteínas (a porção do PSA que não está “livre”). Este teste poderia ser feito em vez de verificar o PSA total e livre, e poderia fornecer a mesma quantidade de informações, mas não é amplamente utilizado;
- **Índice de Saúde da Próstata -** *Prostate Health Index* (PHI), que combina os resultados do PSA total, PSA livre e proPSA (se refere a uma série de precursores inativos de PSA secretados pelas células prostáticas);
- **Teste 4K score**, que combina os resultados de PSA total, PSA livre, PSA

intacto e calicreína humana 2, juntamente com alguns outros fatores;

- **Velocidade do PSA**, a velocidade do PSA não é um teste separado. É uma medida de quão rápido o PSA aumenta ao longo do tempo. Normalmente, os níveis de PSA sobem lentamente com a idade. Algumas pesquisas descobriram que esses níveis sobem mais rapidamente se um homem tem câncer, mas os estudos não mostraram que a velocidade do PSA é mais útil do que o próprio nível de PSA para encontrar o câncer de próstata;
- **Densidade do PSA -** *Prostate-Specific Antigen Density* (PSAD), os níveis de PSA são mais altos em homens com próstatas maiores. A PSAD às vezes é usada para homens com próstatas grandes para tentar se ajustar a isso. O médico mede o volume da próstata com ultrassom transretal e divide o nível do PSA pelo volume da próstata. Uma maior densidade de PSA indica uma maior probabilidade de câncer. A densidade do PSA não demonstrou ser tão útil quanto o teste de percentual de PSA livre.

Para um exame de toque retal, o médico insere um dedo enluvado e lubrificado no reto para sentir qualquer inchaço ou área dura na próstata que possa ser câncer. Os cânceres de próstata geralmente começam na parte posterior da glândula e às vezes podem ser sentidos durante um exame de toque retal. Este exame pode ser desconfortável (especialmente para homens com hemorroidas), mas geralmente não é doloroso e leva pouco tempo. A Figura 4 mostra como é realizado um exame de toque retal.

Figura 4 – Exame de toque retal



Fonte: adaptado de NATIONAL CANCER INSTITUTE (2020)

Outra questão importante é que, mesmo que o rastreamento detecte o câncer

de próstata, os médicos, às vezes, não conseguem dizer se o câncer é realmente perigoso, portanto, precisa ser tratado. Encontrar e tratar todos os cânceres de próstata precocemente pode parecer fazer sentido, mas alguns cânceres de próstata crescem tão lentamente que nunca causariam problemas a um homem durante sua vida.

Por causa da triagem, alguns homens podem ser diagnosticados com um câncer de próstata que, de outra forma, nunca teriam conhecido. Isso nunca teria levado à morte deles, ou mesmo causado qualquer sintoma. Encontrar uma "doença" como essa que nunca causaria problemas é conhecido como sobrediagnóstico.

Um problema com o sobrediagnóstico de câncer de próstata é que muitos desses homens ainda podem ser tratados com cirurgia ou radiação, porque o médico não pode ter certeza da rapidez com que o câncer pode crescer e se espalhar, ou porque o homem fica desconfortável sabendo que tem câncer de próstata e não está recebendo nenhum tratamento. O tratamento de um câncer que nunca teria causado nenhum problema é conhecido como tratamento excessivo. A principal desvantagem disso é que, mesmo que não fossem necessários, tratamentos como cirurgia e radiação ainda podem ter efeitos colaterais urinários, intestinais e/ou sexuais que podem afetar seriamente a qualidade de vida de um homem.

Os homens e seus médicos podem optar por determinar se o tratamento é necessário imediatamente ou se o câncer pode ser monitorado de perto sem intervenção imediata, em uma abordagem chamada espera vigilante ou vigilância ativa. Mesmo quando não há tratamento imediato, é essencial que os homens façam exames regulares de PSA no sangue e biópsias de próstata para avaliar a necessidade de tratamento futuro. Esses exames estão associados a riscos como ansiedade, dor, infecção e sangramento.

Pesquisadores ainda estão estudando se os testes de triagem reduzirão o risco de morte por câncer de próstata. Os resultados mais recentes de dois grandes estudos (PINSKY et al., 2019; HUGOSSON et al., 2019) foram conflitantes e não ofereceram respostas claras.

Nos primeiros resultados de um grande estudo feito nos Estados Unidos por Pinsky et al. (2019), descobriu-se que a triagem anual com PSA e o exame de toque retal, detectou mais cânceres de próstata do que em homens não rastreados, mas essa triagem não reduziu a taxa de mortalidade por câncer de próstata. No entanto,

questões foram levantadas sobre este estudo, porque alguns homens do grupo de não-rastreados foram rastreados durante o estudo, o que pode ter afetado os resultados. Outro estudo europeu, conduzido por Hugosson et al. (2019), identificou um menor risco de morte por câncer de próstata com o rastreamento do PSA realizado aproximadamente a cada 4 anos. No entanto, os pesquisadores estimaram que seria necessário rastrear cerca de 781 homens e detectar 27 casos de câncer para evitar uma única morte por câncer de próstata. Nenhum desses estudos mostrou que a triagem de PSA ajuda os homens a viver mais em geral, ou seja, reduz a taxa de mortalidade geral.

O câncer de próstata costuma crescer lentamente, então os efeitos da triagem nesses estudos podem se tornar mais claros nos próximos anos. Ambos os estudos estão sendo continuados para ver se um acompanhamento mais longo dará resultados mais claros. O toque retal é menos eficaz do que o exame de sangue do PSA na detecção de câncer de próstata, mas às vezes pode detectar câncer em homens com níveis normais de PSA. Por esse motivo, pode ser incluído como parte do rastreamento do câncer de próstata.

Nem o teste de PSA nem o exame de toque retal são 100% precisos. Às vezes, esses testes podem ter resultados positivos mesmo quando um homem não tem câncer (conhecidos como falso-positivos) ou resultados negativos mesmo quando um homem tem câncer (conhecido como falso-negativos). Resultados de testes pouco claros podem causar confusão e ansiedade. Resultados falso-positivos podem levar alguns homens a fazerem biópsias de próstata (com pequenos riscos de dor, infecção e sangramento) quando não têm câncer. E resultados falso-negativos podem dar a alguns homens uma falsa sensação de segurança, mesmo que eles possam realmente ter câncer.

Se os resultados de qualquer um desses testes forem anormais, testes adicionais, como uma biópsia da próstata, geralmente são feitos para verificar se um homem tem câncer. Este teste é a única maneira de saber com certeza se um homem tem câncer de próstata. Se o câncer de próstata for encontrado em uma biópsia, esse teste também pode ajudar a determinar a probabilidade de o câncer crescer e se espalhar rapidamente.

2.1.3 Exames de imagem

Os exames de imagem usam raios-x, campos magnéticos, ondas sonoras ou substâncias radioativas para criar imagens do interior do corpo. Um ou mais exames de imagem podem ser usados para procurar câncer na próstata, para ajudar o médico a ver a próstata durante certos procedimentos (como uma biópsia de próstata ou certos tipos de tratamento de câncer de próstata), e para procurar a disseminação do câncer de próstata para outras partes do corpo.

A escolha do exame de imagem a ser utilizado depende do procedimento a ser realizado. Por exemplo, uma biópsia de próstata geralmente é feita com a utilização do TRUS e/ou IRM para ajudar a orientar a biópsia. Se for constatado câncer de próstata, outros exames de imagem de outras partes do corpo podem ser necessários para procurar uma possível disseminação do câncer. Homens com resultado normal no exame de toque retal, PSA baixo e pontuação de *Gleason* baixa podem não precisar de nenhum outro teste porque a chance de o câncer se espalhar é muito baixa.

Para realizar um exame utilizando TRUS, uma pequena sonda da largura de um dedo é lubrificada e colocada no reto do paciente. Esta sonda emite ondas sonoras que penetram na próstata e geram ecos. Os ecos são captados pela sonda e transformados, por um computador, em uma imagem em preto e branco da próstata. O TRUS pode ser utilizado em diferentes situações (NATIONAL CANCER INSTITUTE, 2019):

- Procurar áreas suspeitas na próstata em homens com resultados anormais no teste de toque retal ou PSA (embora possa não detectar alguns tipos de câncer);
- Pode ser usado durante uma biópsia da próstata para guiar as agulhas na área correta da próstata;
- Pode ser usado para medir o tamanho da próstata, o que pode ajudar a determinar a densidade do PSA;
- Pode ser usado como um guia durante algumas formas de tratamento, como braquiterapia (radioterapia interna) ou crioterapia.

Os exames de ressonância magnética criam imagens detalhadas dos tecidos

moles do corpo usando ondas de rádio e ímãs fortes. Os exames de ressonância magnética podem dar aos médicos uma imagem muito clara da próstata e áreas próximas. Um material de contraste chamado gadolínio pode ser injetado em uma veia antes da varredura para ver melhor os detalhes. A ressonância magnética pode ser usada em diferentes situações (NATIONAL CANCER INSTITUTE, 2019):

- Determinar se um homem com um teste de triagem anormal ou com sintomas que possam ser de câncer de próstata deve fazer uma biópsia de próstata. O tipo de ressonância magnética frequentemente usado para isso é conhecido como Imagem por Ressonância Magnética multiparamétrica (IRMmp);
- Ajudar a localizar e direcionar as áreas da próstata com maior probabilidade de conter câncer. Isso geralmente é feito como uma biópsia de fusão de ressonância magnética/ultrassom;
- Pode ser usada durante uma biópsia da próstata para ajudar a guiar as agulhas na próstata;
- Determinar a extensão (estágio) do câncer. Podem mostrar se o câncer se espalhou para fora da próstata nas vesículas seminais ou outras estruturas próximas. Isso pode ser muito importante para determinar as opções de tratamento. Mas exames de ressonância magnética geralmente não são necessários para cânceres de próstata recém-diagnosticados que provavelmente estão confinados à próstata com base em outros fatores.

A técnica mais recente de ressonância magnética é a IRMmp. Esta técnica pode ser usada para ajudar a definir melhor possíveis áreas de câncer na próstata e avaliar a rapidez com que um câncer pode crescer. Além disso, pode ajudar a identificar se o câncer cresceu além da próstata ou se espalhou para outras partes do corpo. Para este exame, inicialmente é realizada uma ressonância magnética padrão para analisar a anatomia da próstata. Em seguida, pelo menos um outro tipo de ressonância magnética é realizado, como a imagem ponderada por difusão - *Diffusion-Weighted Imaging* (DWI), ressonância magnética aprimorada com contraste dinâmico, ou espectroscopia de ressonância, para avaliar outros parâmetros do tecido prostático. Os resultados dessas diferentes varreduras são então comparados para identificar áreas anormais.

Quando este exame é realizado para ajudar a determinar se um homem pode ter câncer de próstata, os resultados são geralmente reportados utilizando o Sistema de Dados e Relatórios de Imagens da Próstata - *Prostate Imaging Reporting and Data System* (PI-RADS). Nesse sistema, as áreas anormais na próstata são classificadas em categorias que variam de PI-RADS 1 (muito improvável de ser um câncer clinicamente significativo) a PI-RADS 5 (muito provável de ser um câncer clinicamente significativo). Essa classificação ajuda os médicos a tomarem decisões informadas sobre a necessidade de biópsias ou outros procedimentos diagnósticos adicionais. O uso do PI-RADS facilita a padronização e a comparação dos resultados de ressonância magnética entre diferentes instituições, garantindo uma abordagem mais consistente na detecção e avaliação do câncer de próstata. Além disso, esse sistema contribui para o monitoramento da progressão da doença e a avaliação da resposta ao tratamento, auxiliando na escolha das estratégias terapêuticas mais apropriadas para cada paciente.

Outra técnica é a biópsia de próstata guiada por fusão de IRM/TRUS. Nesta abordagem, o paciente realiza uma ressonância magnética alguns dias ou semanas antes da biópsia para procurar áreas anormais na próstata. Durante a própria biópsia, o TRUS é usado para visualizar a próstata e um programa de computador especial é usado para fundir as imagens de ressonância magnética e o TRUS em uma tela de computador. Isso pode ajudar a garantir que o médico obtenha amostras de biópsia de todas as áreas suspeitas vistas nas imagens.

Outro exame de imagem utilizado é a cintilografia óssea. Se o câncer de próstata se espalhar para partes distantes do corpo, geralmente atinge primeiro os ossos. Uma cintilografia óssea pode ajudar a mostrar se o câncer atingiu os ossos. Para este teste, é injetado uma pequena quantidade de material radioativo de baixo nível, que se deposita em áreas danificadas do osso em todo o corpo. Uma câmera especial detecta a radioatividade e cria uma imagem do seu esqueleto. Uma cintilografia óssea pode sugerir câncer no osso, mas para fazer um diagnóstico preciso, podem ser necessários outros exames, como radiografia simples, tomografia computadorizada ou ressonância magnética, ou mesmo uma biópsia óssea.

A Tomografia por Emissão de Pósitrons - *Positron Emission Tomography* (PET) é semelhante a uma cintilografia óssea, em que uma substância levemente radioativa (conhecida como traçador) é injetada no sangue, que pode ser detectada

com uma câmera especial. Mas as varreduras de PET usam marcadores diferentes que se acumulam principalmente nas células cancerosas. O traçador mais comum para exames de PET padrão é o *Flúor-Deoxi-Glicose*, que é um tipo de açúcar. Infelizmente, esse tipo de PET *scan* não é muito útil para encontrar células de câncer de próstata no corpo. No entanto, marcadores mais recentes, como fluciclovina F18, fluoreto de sódio F18 e colina C11, demonstraram ser melhores na detecção de células de câncer de próstata. Outros marcadores mais recentes, como Ga 68 PSMA-11, 18F-DCFPyl (também conhecido como piflufolastat F18 ou Pylarify) e Ga 68 gozetotide (Locametz), ligam-se ao antígeno de membrana específico da próstata - *Prostate-Specific Membrane Antigen* (PSMA), uma proteína frequentemente encontrada em grandes quantidades em células de câncer de próstata. Os testes que usam esses tipos de rastreadores às vezes são chamados de varreduras PSMA PET.

Esses tipos mais recentes de PET são frequentemente utilizados quando não está claro se o câncer de próstata se espalhou ou para determinar exatamente onde ele se espalhou. Por exemplo, um desses testes pode ser feito se os resultados de uma cintilografia óssea não forem claros ou se um homem tiver um aumento no nível de PSA após o tratamento inicial, mas não estiver claro onde o câncer está no corpo.

As imagens de um PET *scan* não são tão detalhadas quanto as imagens de ressonância magnética ou tomografia computadorizada, mas muitas vezes podem mostrar áreas de câncer em qualquer parte do corpo. Algumas máquinas podem fazer uma varredura PET e uma ressonância magnética PET-IRM ou uma tomografia computadorizada PET-CT ao mesmo tempo, o que pode fornecer mais detalhes sobre as áreas que aparecem na varredura PET.

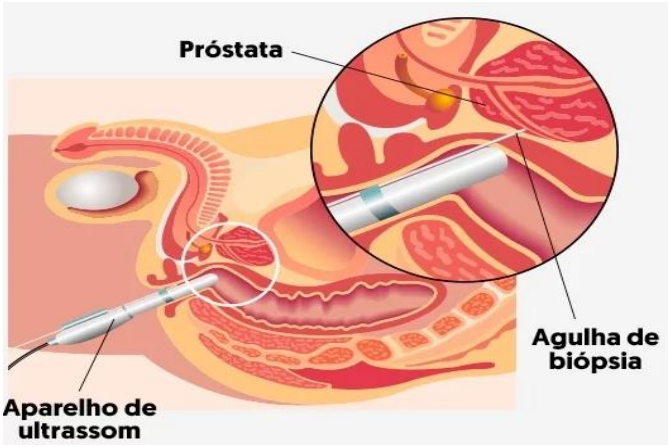
A Tomografia computadorizada usa raios-x para fazer imagens detalhadas e transversais do seu corpo. Este teste geralmente não é necessário para câncer de próstata recém-diagnosticado se o câncer provavelmente estiver confinado à próstata com base em outros achados (resultado de toque retal, nível de PSA e pontuação de *Gleason*). Ainda assim, às vezes pode ajudar a dizer se o câncer de próstata se espalhou para os gânglios linfáticos próximos. Se o câncer de próstata voltou após o tratamento, a tomografia computadorizada geralmente pode dizer se ele está crescendo em outros órgãos ou estruturas na pélvis. A tomografia computadorizada não é tão útil quanto a ressonância magnética para observar a própria próstata.

2.1.4 Biópsia da próstata

Se os resultados de um exame de sangue de PSA, exame de toque retal ou outros testes sugerirem a presença de câncer de próstata, o paciente provavelmente precisará passar por uma biópsia da próstata. Este procedimento envolve a remoção de pequenas amostras da próstata, que são posteriormente examinadas ao microscópio para detectar células cancerosas.

Durante a biópsia, o médico geralmente examina a próstata com um exame de imagem, como o TRUS ou ressonância magnética, ou uma "fusão" dos dois. O médico insere rapidamente uma agulha fina e oca na próstata. Isso é feito através da parede do reto (uma biópsia transretal) ou através da pele entre o escroto e o ânus (uma biópsia transperineal). Quando a agulha é retirada, ela remove um pequeno cilindro (núcleo) de tecido prostático. Isso é repetido várias vezes. Na maioria das vezes, o médico coletará cerca de 12 amostras principais de diferentes partes da próstata. A Figura 5 mostra uma biópsia da próstata.

Figura 5 – Biópsia da próstata



Fonte: adaptado de NATIONAL CANCER INSTITUTE (2019)

As amostras de biópsia serão enviadas para um laboratório, onde serão examinadas com um microscópio para ver se contêm células cancerosas. Os resultados podem ser relatados como:

- **Positivo para câncer**, células cancerosas foram observadas nas amostras de biópsia;
- **Negativo para câncer**, nenhuma célula cancerosa foi observada nas

amostras de biópsia;

- **Suspeito**, algo anormal foi visto, mas pode não ser câncer.

Mas, mesmo que muitas amostras sejam coletadas, as biópsias podem, às vezes, não detectar um câncer se nenhuma das agulhas da biópsia passar por ele. Isso é conhecido como resultado falso-negativo. Se o médico ainda suspeitar que o paciente possa ter câncer de próstata, devido a um nível de PSA muito alto, por exemplo, ele pode sugerir outros exames laboratoriais (de sangue, urina ou amostras de biópsia da próstata) ou outros testes, como: PHI, teste 4Kscore, testes PCA3 (como Progensa) e ConfirmMDx. O médico também pode repetir a biópsia da próstata para obter amostras adicionais de partes da próstata não examinadas anteriormente, ou utilizar exames de imagem, como a ressonância magnética, para examinar mais detalhadamente as áreas anormais a serem avaliadas.

Se o câncer de próstata for detectado em uma biópsia, ele receberá uma classificação de grau. O grau do câncer é determinado com base em quão anormal as células cancerosas aparecem ao microscópio. Cânceres de alto grau parecem mais anormais e têm maior probabilidade de crescer e se espalhar rapidamente. Existem duas principais formas de descrever o grau de um câncer de próstata: a primeira é utilizando o sistema *Gleason*, e a segunda é através de grupo de notas ou graus.

O sistema Gleason, amplamente utilizado há muitos anos, atribui notas ao câncer de próstata com base na similaridade das células cancerosas ao tecido normal da próstata. Este sistema avalia o grau de anormalidade celular, fornecendo uma pontuação que ajuda a prever o comportamento do câncer. Cânceres com notas mais altas apresentam células que se desviam mais do aspecto do tecido prostático normal, indicando uma maior agressividade e uma maior probabilidade de crescimento e disseminação rápida (NATIONAL CANCER INSTITUTE, 2019):

- Se o câncer se parece muito com o tecido normal da próstata, é atribuído o grau 1;
- Se o câncer parecer muito anormal, ele receberá a nota 5;
- Os graus 2 a 4 têm características entre esses extremos. Quase todos os cânceres são de grau 3 ou superior; os graus 1 e 2 não são frequentemente usados.

Como os cânceres de próstata geralmente têm áreas com graus diferentes, um grau é atribuído às duas áreas que compõem a maior parte do câncer. Essas duas notas são adicionadas para produzir a pontuação de *Gleason* (também chamada de soma de *Gleason*).

O primeiro número atribuído é o grau mais comum no tumor. Por exemplo, se a pontuação de *Gleason* for escrita como “3 + 4 = 7”, isso significa que a maior parte do tumor é de grau 3 e a menor parte é de grau 4, e eles são somados para uma pontuação de *Gleason* igual a 7.

Embora na maioria das vezes a pontuação de *Gleason* seja baseada nas duas áreas que compõem a maior parte do câncer, há algumas exceções quando uma amostra de biópsia tem, ou muito câncer de alto grau ou há 3 graus, incluindo câncer de alto grau. Nesses casos, a forma como a pontuação de *Gleason* é determinada é modificada para refletir a natureza agressiva (de crescimento rápido) do câncer.

Em teoria, a pontuação de *Gleason* pode estar entre 2 e 10, mas pontuações abaixo de 6 raramente são usadas. Com base na pontuação de *Gleason*, os cânceres de próstata são frequentemente divididos em três grupos (NATIONAL CANCER INSTITUTE, 2019):

- Cânceres com pontuação de *Gleason* de 6 ou menos podem ser chamados de bem diferenciados ou de baixo grau;
- Cânceres com pontuação de *Gleason* de 7 podem ser chamados de moderadamente diferenciados ou de grau intermediário;
- Cânceres com pontuação de *Gleason* de 8 a 10 podem ser chamados de pouco diferenciados ou de alto grau.

Nos últimos anos, os médicos perceberam que a pontuação de *Gleason* nem sempre é a melhor maneira de descrever o grau do câncer, por alguns motivos: Os resultados do câncer de próstata podem ser divididos em mais do que apenas os três grupos mencionados acima. Por exemplo, homens com câncer de pontuação de *Gleason* “3 + 4 = 7” tendem a se sair melhor do que aqueles com câncer de “4 + 3 = 7”. E os homens com um câncer de pontuação de *Gleason* igual a 8 tendem a se sair melhor do que aqueles com uma pontuação de *Gleason* de 9 ou 10.

A escala da pontuação de *Gleason* pode ser confusa para os pacientes. Por exemplo, um homem com um câncer de pontuação de *Gleason* 6 pode assumir que

seu câncer está no meio da faixa de graus (que teoricamente vai de 2 a 10), embora, na prática, os cânceres de grau 6 sejam, na verdade, os de menor agressividade. Essa suposição pode levar o paciente a pensar que seu câncer tem uma probabilidade maior de crescer e se espalhar rapidamente do que realmente tem, o que pode influenciar suas decisões sobre o tratamento.

Por causa disso, para melhorar a compreensão, os médicos desenvolveram grupos de graus, que variam de 1 (mais propensos a crescer e se espalhar lentamente) a 5 (mais propensos a crescer e se espalhar rapidamente) (NATIONAL CANCER INSTITUTE, 2019):

- Grupo de notas 1: *Gleason* 6 (ou menos);
- Grupo de notas 2: *Gleason* 3 + 4 = 7;
- Grupo de notas 3*: Gleason 4 + 3 = 7;*
- Grupo de notas 4: *Gleason* 8;
- Grupo de notas 5: *Gleason* 9 a 10.

Os grupos de classificação provavelmente substituirão a pontuação de *Gleason* ao longo do tempo, mas atualmente pode ser visto um, ou ambos, em um relatório de patologia de biópsia.

Às vezes, quando as células da próstata são vistas, elas não se parecem com câncer, mas também não são normais. Esses casos são denominados de neoplasia intraepitelial prostática - *Prostatic Intraepithelial Neoplasia* (PIN). Na PIN, há alterações na aparência das células da próstata, mas as células anormais não parecem ter crescido em outras partes da próstata (como as células cancerosas fariam). PIN é frequentemente dividida em dois grupos (NATIONAL CANCER INSTITUTE, 2019):

- PIN de baixo grau: os padrões das células da próstata parecem quase normais;
- PIN de alta qualidade: os padrões das células parecem mais anormais.

Muitos homens começam a desenvolver a PIN de baixo grau em tenra idade, mas acredita-se que a PIN de baixo grau não esteja relacionado ao risco de câncer de próstata. Se a PIN de baixo grau for relatada em uma biópsia de próstata, o acompanhamento dos pacientes geralmente é o mesmo como se nada de anormal

fosse observado.

Se a PIN de alto grau for encontrado em uma biópsia, há uma chance maior do paciente desenvolver câncer de próstata ao longo do tempo. É por isso que os médicos geralmente observam cuidadosamente os homens com a PIN de alto grau e podem aconselhar outra biópsia da próstata ou exames laboratoriais para ajudar a determinar o risco de ter câncer, como o PHI, teste 4Kscore, testes PCA3 ou ConfirmMDx. Isso é especialmente verdadeiro se a PIN de alto grau for encontrada em diferentes partes da próstata (PIN multifocal de alto grau) ou se a biópsia original não tiver coletado amostras de todas as partes da próstata.

Outro termo utilizado como resultado suspeito é a proliferação atípica de pequenos acinos - *Atypical Small Acinar Proliferation* (ASAP), que também pode ser chamado de atipia glandular ou proliferação glandular atípica. Significa que as células parecem cancerosas quando vistas ao microscópio, mas são muito poucas para ter certeza. Portanto, há uma grande chance de que também haja câncer na próstata, por isso muitos médicos recomendam repetir a biópsia em alguns meses.

A Atrofia inflamatória proliferativa - *Proliferative Inflammatory Atrophy* (PIA) é outro termo utilizado quando encontrado algo suspeito na biópsia. Na PIA, as células da próstata parecem menores que o normal e há sinais de inflamação na área. A PIA não é câncer, mas os pesquisadores acreditam que a PIA às vezes pode levar a PIN de alto grau ou diretamente ao câncer de próstata.

2.2 Aprendizado de máquina

O aprendizado de máquina inclui várias técnicas que permitem aos computadores aprenderem e tomarem excelentes decisões (SARKER, 2021). O aprendizado de máquina - *Machine Learning* (ML) iniciou-se na década de 1960 como um campo da inteligência artificial que tinha o objetivo de aprender padrões com base em dados. Originalmente, as aplicações de ML eram de cunho estritamente computacional. Contudo, desde o final dos anos 1990, essa área expandiu seus horizontes e começou a se estabelecer como um campo por si mesma. Em particular, as aplicações de ML começaram a ter muitas intersecções com a área de diagnóstico de doenças.

O tipo mais comum de aprendizado de máquina é aprender o mapeamento

$Y = f(X)$ para fazer previsões de Y para um novo X. Isso é chamado de modelagem preditiva ou análise preditiva e tem como objetivo fazer as previsões mais precisas possíveis. Como tal, não se estar realmente interessado no formato da função f que se estar aprendendo, apenas que ela faça previsões precisas. Pode-se até aprender o mapeamento de $Y = f(X)$ para entender mais sobre o relacionamento nos dados, isso é chamado de inferência estatística. Se esse fosse o objetivo, se usariam métodos mais simples e se valorizava a compreensão do modelo aprendido e da forma da função f, fazendo previsões precisas.

Quando se aprende uma função f, está estimando sua forma a partir dos dados disponíveis. Como tal, esta estimativa terá algum erro. Não será uma estimativa perfeita para o melhor mapeamento hipotético subjacente de Y dado X. Muito tempo aplicado no aprendizado de máquina é gasto tentando melhorar a estimativa da função subjacente, e melhorar o desempenho das previsões feitas pelo modelo. Os modelos gerados são ainda capazes de lidar com situações não apresentadas durante seu desenvolvimento, sem necessariamente necessitar de uma nova fase de projeto.

Os algoritmos de aprendizado de máquina são técnicas para estimar a função alvo f para prever a variável de saída Y, dadas as variáveis de entrada X. Representações diferentes fazem suposições diferentes sobre a forma da função que está sendo aprendida, como por exemplo, se ela é linear ou não linear. Os algoritmos de aprendizado de máquina fazem suposições diferentes sobre a forma e a estrutura da função e sobre a melhor forma de otimizar uma representação para aproximá-la. É por isso que é tão importante testar um conjunto de algoritmos diferentes em um mesmo problema de aprendizado de máquina, porque não se pode saber de antemão qual abordagem será melhor para estimar a estrutura da função subjacente que se está tentando aproximar. Esses algoritmos podem reconhecer padrões, extrapolar a partir de exemplos e mudar ao longo do tempo (JUTEL et al., 2023).

Além do ML ser associado à Inteligência Artificial, outras áreas de pesquisa são importantes e têm contribuído diretamente e significativamente no avanço do ML, tais como: probabilidade e estatística, teoria da computação, neurociência, teoria da informação, entre outras. ML é uma das áreas de pesquisa da computação ou engenharia que têm mais crescido nos últimos anos. Diferentes algoritmos de ML, diferentes formas de utilizar os algoritmos existentes e adaptações de algoritmos são continuamente propostos. Além disso, surgem a todo instante novas variações nas

características dos problemas reais a serem tratados.

Existem várias aplicações dos algoritmos de ML na solução de problemas reais, tais como: reconhecimento de palavras faladas, detecção do uso fraudulento de cartões de crédito, condução de automóveis de forma autônoma, predição de taxa de cura de pacientes com diferentes doenças, diagnóstico de câncer por meio da análise de dados, entre outras.

Além do grande volume de aplicações que se beneficiam das características da área de ML, outros fatores têm favorecido a expansão dessa área, como o desenvolvimento de algoritmos cada vez mais eficazes e eficientes, e a elevada capacidade dos recursos computacionais atualmente disponíveis. Ao se aplicar aprendizado de máquina a enormes conjuntos de dados, permitindo que os computadores aprendam à medida que progridem, a programação eliminou muitos perigos e becos sem saída (HUA, 2022).

Existe uma relação entre ML e hipótese de indução. Por exemplo, cada dado correspondente a um paciente é uma tupla formada pelos valores das características referentes ao paciente, que descrevem seus principais aspectos. As características, também chamadas de atributos ou variáveis, utilizadas para cada paciente podem ser, por exemplo, sua identificação, nome, idade, sexo, sintomas e exames clínicos.

Para algumas tarefas de ML, uma das características é considerada um atributo de saída, também chamado de atributo alvo, cujos valores podem ser estimados utilizando os valores das demais características, denominados características de entrada. O objetivo de um algoritmo de ML é aprender, a partir de um subconjunto dos dados, denominado conjunto de treinamento, um modelo ou hipótese capaz de relacionar os valores das características de entrada de um objeto do conjunto de treinamento ao valor de seu atributo de saída.

O que se deseja é induzir uma hipótese capaz de diagnosticar corretamente novos pacientes diferentes daqueles que foram utilizados para aprender a regra de decisão. Assim, uma vez induzida uma hipótese, é desejável que ela também seja válida para outros objetos do mesmo problema quem não fizeram parte do conjunto de treinamento. A essa propriedade dá-se o nome de capacidade de generalização. A generalização é um conceito-chave no aprendizado de máquina, pois se refere à capacidade de uma máquina de extrapolar informações a partir de um conjunto limitado de dados de treinamento e aplicá-las a novos dados não vistos anteriormente.

Em outras palavras, é a habilidade de uma máquina aprender padrões e tendências a partir de exemplos específicos e aplicá-los a situações semelhantes.

Quando uma hipótese apresenta uma baixa capacidade de generalização, a razão pode ser que ela esteja sobreajustada (*overfitting*) aos dados. Nesse caso, também é dito que a hipótese memorizou ou se especializou nos dados de treinamento. No caso inverso, o algoritmo de ML pode induzir hipóteses que apresentem uma baixa taxa de acerto, mesmo no subconjunto de treinamento, configurando uma condição de subajuste (*underfitting*). Essa situação pode ocorrer quando exemplos de treinamento disponíveis são pouco representativos ou o modelo usado é muito simples e não captura os padrões existentes nos dados.

Quando um algoritmo de ML está aprendendo a partir de um conjunto de dados de treinamento, ele está procurando uma hipótese, no espaço de hipóteses possíveis, capaz de descrever as relações entre os objetos e que melhor se ajuste aos dados de treinamento. Cada algoritmo utiliza uma forma de representação para descrever a hipótese induzida. A representação define a preferência ou viés (*bias*) de representação do algoritmo e pode restringir o conjunto de hipóteses que podem ser induzidas pelo algoritmo.

Além do viés de representação, os algoritmos de ML possuem também um viés de busca. O viés de busca de um algoritmo é a forma como o algoritmo busca a hipótese que melhor se ajusta aos dados de treinamento. Ele define como as hipóteses são pesquisadas no espaço de hipóteses. Ou seja, o viés restringe as hipóteses a serem visitadas no espaço de busca. Sem viés não haveria aprendizado ou generalização, pois os modelos seriam especializados para exemplos individuais.

Os dois paradigmas mais comuns de aprendizado de máquina são o aprendizado supervisionado e o aprendizado não supervisionado. Em tarefas de previsão, a meta é encontrar uma função a partir dos dados de treinamento que possa ser utilizada para prever um rótulo que caracterize uma nova observação, com base nos valores dos atributos de entrada e saída. Os algoritmos de ML utilizados nessa tarefa induzem modelos preditivos. Esses algoritmos seguem o paradigma de aprendizado supervisionado. Os modelos são aprendidos a partir de dados com rótulos por meio do aprendizado supervisionado, o que lhes permite fazer previsões precisas (KOUROU et al., 2015).

O aprendizado supervisionado é a forma predominante de aprendizado de

máquina, e funciona em duas etapas. O algoritmo identifica padrões no conjunto de dados de treinamento, consistindo em conjuntos de amostras e rótulos (alvo). Em seguida, transforma esses padrões em uma expressão matemática conhecida como modelo durante a etapa de treinamento. Este modelo é utilizado para fazer previsões sobre amostras que não encontrou durante a fase de inferência (KULKARNI; KULKARNI; PANT, 2021).

Várias técnicas fornecem uma função que atribui entradas às saídas correspondentes desejadas durante o aprendizado supervisionado. O problema de classificação é um exemplo típico de tarefa de aprendizado supervisionado. Neste cenário, deve-se aproximar o comportamento de uma função que atribui um vetor a uma das classes. Isso é conseguido analisando inúmeras amostras de entrada-saída da função (NASTESKI, 2017).

O processo de aprendizagem de um modelo básico de ML envolve duas etapas: treinamento e teste. Na etapa de treinamento, o modelo é construído usando exemplos dos dados de treinamento como entrada, permitindo que o algoritmo de aprendizagem adquira conhecimento (DHAGE; RAINA, 2016). O aprendizado supervisionado é usado quando dados rotulados estão disponíveis para treinamento, e é aplicado em tarefas de classificação, por exemplo, detecção de spam, classificação de imagens, análise de sinais e processamento de imagens, e regressão, por exemplo, previsão de preços de casas (SHETTY et al., 2022).

Em tarefas de descrição, a meta é explorar um conjunto de dados. Os algoritmos de ML utilizados nessas tarefas não utilizam o atributo de saída. Por isso, seguem o paradigma de aprendizado não supervisionado. O aprendizado não supervisionado encontra estruturas e padrões em dados não rotulados, trazendo à luz informações escondidas (WICKRAMASINGHE et al., 2021). As tarefas desse tipo de aprendizado se referem à identificação de informações relevantes nos dados sem a presença de um supervisor externo para guiar o aprendizado, ou seja, não existe o atributo alvo ou meta. Essencialmente, o aprendizado reside na identificação de propriedades intrínsecas aos dados de entrada, de maneira a construir representações desses dados que servirão a diversos propósitos, tais como: auxílio a tomada de decisões ou descoberta de conhecimento. Essas técnicas são utilizadas principalmente quando o objetivo do aprendizado seja encontrar padrões ou tendências que auxiliem no entendimento dos dados.

2.2.1 Algoritmos de classificação

Considerando uma amostra com observações independentes $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$ com objetivo de construir uma função $g(x)$ que possa ser usada para fazer bem a predição de novas observações $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$, isto é, se quer que $g(X_{n+1}) \approx Y_{n+1}, \ldots, g(X_{n+m}) \approx Y_{n+m}$. Em um problema de classificação a variável resposta Y é uma variável qualitativa. Por exemplo, prever se um paciente tem uma certa doença com base em variáveis clínicas X é um problema de classificação.

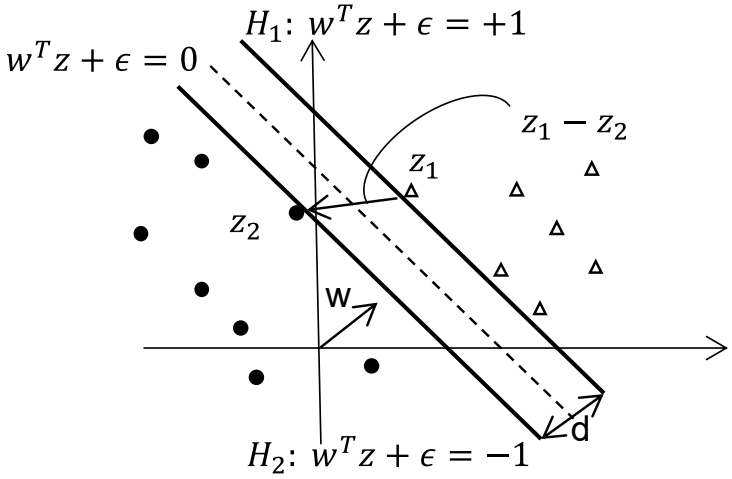
2.2.1.1 *Support Vector Machines* (SVM)

As Máquinas de Vetores de Suporte - *Support Vector Machines* (SVM) são um método de aprendizado supervisionado, capaz de classificar a partir de ***n*** indivíduos observados pertencentes a diversos subgrupos, a que classe um indivíduo pertence, e tem como base a teoria de aprendizado estatístico, desenvolvida por Cortes e Vapnik (1995). Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa capacidade de generalização.

A ideia da SVM é construir um hiperplano como superfície de decisão, de tal forma que a margem de separação entre as classes seja máxima possível. O objetivo do treinamento através da SVM é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização.

A SVM é considerada um sistema de aprendizagem que utiliza um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador SVM é capaz de encontrar um hiperplano baseado em um conjunto de pontos, denominados vetores de suporte, o qual maximiza a margem de separação entre as classes. Mesmo quando as duas classes não são separáveis, a SVM é capaz de encontrar um hiperplano através do uso de conceitos pertencentes à teoria da otimização (DING; PENG, 2005). A Figura 6 ilustra um hiperplano ótimo para padrões linearmente separáveis.

Figura 6 - Hiperplano ótimo, com dois vetores de suporte H_1 e H_2



Fonte: MITCHELL (1997)

Considerando o caso de classificação utilizando duas classes, usando um modelo linear descrito por:

$$y(z) = w^T z + \epsilon = 0 \tag{2.1}$$

Sendo w um vetor de pesos ajustados, $\boldsymbol{\epsilon}$ um viés e z um vetor de treinamento de características, com seus respectivos rótulos $y_i \in Y$, em que $Y = \{-1, +1\}$. O modelo definido na equação (2.1) define um hiperplano ótimo, que classifica todos os vetores de treinamento, e z é dito linearmente separável, se é possível separar os dados das classes –1 e +1 por este hiperplano (SCHÖLKOPF; SMOLA, 2018).

Baseado no modelo da Equação (2.1), tem-se:

$$\begin{cases} w^T z + \epsilon \geq 0, \text{ para } y(z) = +1 & (2.2) \\ w^T z + \epsilon < 0, \text{ para } y(z) = -1 & (2.3) \end{cases}$$

Sendo z_1 um ponto pertencente ao hiperplano $H_1 = w^T z + \epsilon = +1$, e z_2 um ponto pertencente ao hiperplano $H_2 = w^T z + \epsilon = -1$, conforme ilustrado, anteriormente, na Figura 6.

Ao se projetar $z_1 - z_2$ na direção de w, perpendicular ao hiperplano ótimo, é possível obter a distância entre os hiperplanos H_1 e H_2, dada pela Equação (2.4), a seguir:

$$(z_1 - z_2)\left(\frac{w}{||w||}.\frac{(z_1 - z_2)}{||z_1 - z_2||}\right) \tag{2.4}$$

A diferença entre as Equações (2.2) e (2.3) tem como resultado $w(z_1 - z_2) = 2$. Substituindo o resultado encontrado na equação (2.4), tem-se:

$$\frac{2(z_1 - z_2)}{||w||.||z_1 - z_2||} \tag{2.5}$$

Tomando-se a norma da equação (2.5) acima, tem-se:

$$\frac{2}{||w||} \tag{2.6}$$

Esta é a distância $\boldsymbol{d}$, já ilustrada na Figura 6, entre os hiperplanos H_1 e H_2, paralelos ao hiperplano ótimo separado. Minimizando $\|w\|$, pode-se minimizar a margem de separação dos dados em relação ao $w^T z + \epsilon = 0$. Assim, tem-se o problema de otimização descrito por:

$$\min_{w,\epsilon} \frac{1}{2}||w||^2 \tag{2.7}$$

Com a restrição $y_i(w^T z_i + \epsilon) - 1 \geq 0, \forall_i = 1,2,\ldots,n$, que é imposta para assegurar que não haja dados de treinamento entre as margens de separação das classes. Este é um problema de otimização quadrática, cuja função objetivo é convexa e os pontos que satisfazem as restrições formam um conjunto convexo, logo possui um único mínimo global.

Para solucionar tal problema, aplica-se um operador Lagrangiano, capaz de englobar as restrições às funções objetivo, associadas aos multiplicadores de Lagrange ρ_i, conforme a equação a seguir:

$$L(w, \epsilon, \rho) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \rho_i(y_i(w^T z_i + \epsilon) - 1) \tag{2.8}$$

Onde o operador Lagrangiano deve ser minimizado, implicando na maximização das variáveis ρ_i, enquanto w e ϵ devem ser minimizados.

Derivando L em relação a ϵ e w, e igualando a 0 (zero), obtêm-se as equações (2.9) e (2.10):

$$w = \sum_{i=1}^{n} \rho_i y_i z_i \tag{2.9}$$

$$\sum_{i=1}^{n} \rho_i y_i = 0 \tag{2.10}$$

Substituindo as equações (2.9) e (2.10) na equação (2.8), tem-se o seguinte problema de otimização:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j (z_i . z_j) \tag{2.11}$$

$$\text{com as restrições: } \begin{cases} \rho_i \geq 0, \forall_i = 1, \dots, n \\ \sum_{i=1}^{n} \rho_i y_i = 0 \end{cases} \tag{2.12}$$

Sendo que $(z_i . z_j)$ corresponde ao produto interno entre $z_i \, e \, z_j$. Essa formulação é denominada forma dual, enquanto o problema original é referenciado como forma primal. A forma dual possui os atrativos de apresentar restrições mais simples e permitir a representação do problema de otimização em termos de produtos internos entre objetos. É interessante observar também que o problema dual é formulado utilizando apenas dados de treinamento e os seus rótulos.

Como trata-se de padrões não separáveis linearmente, não é possível construir um hiperplano ótimo de separação sem encontrar erros de classificação (MITCHELL, 1997). Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis. Isso se deve a alguns fatores, tais como: presença de ruídos, *outliers* nos objetos ou à própria natureza do problema, que pode ser não linear. Para o caso de pontos de dados não separáveis, é introduzido um conjunto de variáveis escalares não negativas, ε_i, na definição do hiperplano de separação, para resolver este problema, assim tem-se:

$$y_i(w^T z_i + \epsilon) \geq 1 - \varepsilon_i, \; i = 1, 2, \cdots, n \tag{2.13}$$

As variáveis ε_i são chamadas variáveis de folga e medem o desvio de um ponto de dado na condição ideal de separabilidade de padrões. Estas variáveis relaxam as restrições impostas ao problema de otimização primal. A aplicação deste procedimento suaviza as margens do classificador linear, permitindo que alguns objetos permaneçam entre os hiperplanos H_1 e H_2, e também a ocorrência de alguns erros de classificação.

Um erro no conjunto de treinamento é indicado por um valor de ε_i maior que 1. Logo, a soma dos ε_i representa um limite no número de erros de treinamento. Para esse termo ser levado em consideração, minimizando o erro sobre os dados de treinamento, a função objetivo da equação (2.7) é reformulada como (BURGES, 1998):

$$\min_{w,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \tag{2.14}$$

Sujeito a: $y_i(w.z_i + \epsilon) \geq 1 - \varepsilon_i, \;\; \varepsilon_i \geq 0, \; \forall_i = 1, \cdots, n$ (2.15)

Onde C é um termo de regularização que impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo. O termo $\sum_{i=1}^{n} \varepsilon_i$ também pode ser visto como uma minimização de erros marginais, pois um valor de $\varepsilon_i \in (0,1]$ indica um objeto entre as margens.

Novamente, trata-se de um problema de otimização que é quadrático, com as restrições lineares dadas pela equação (2.15). Aplicando o operador Lagrangiano, tem-se:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j \left(z_i . z_j\right) \tag{2.16}$$

Sujeito as restrições: $\begin{cases} 0 \leq \rho_i \geq C, \forall_i = 1, \ldots, n \\ \sum_{i=1}^{n} \rho_i y_i = 0 \end{cases}$ (2.17)

As variáveis ε_i podem ser calculadas pela equação (2.18)

$$\varepsilon_i = \max \{0,1 - y_i \sum_{j=1}^{n} y_j \rho_j z_j \, . \, z_i + \epsilon\} \tag{2.18}$$

Os pontos z_i são denominados vetores de suporte, e são os objetos que participam da formação do hiperplano separador.

Os problemas não lineares de classificação são resolvidos mapeando o conjunto de treinamento, saindo de seu espaço de entrada para um novo espaço com maior dimensão, denominado espaço de características. Seja $\theta(z)$, uma função que mapeia o espaço de entrada sobre o espaço de características. Esta função, também chamada de função de núcleo (*kernel*), é usada para mapear z_i e z_j para o espaço de características, antes da realização do produto interno entre eles:

$$K(z_i, y_i) = \theta(z_i).\theta(z_j) \tag{2.19}$$

Modificando desta forma, o problema de maximização proposto, será:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_i \rho_j y_i y_j K(z_i . z_j) \tag{2.20}$$

2.2.1.2 *Naive Bayes*

O raciocínio bayesiano fornece uma abordagem probabilística para inferência. Baseia-se na suposição de que as quantidades de interesse são governadas por distribuições de probabilidade e que decisões ótimas podem ser tomadas através do raciocínio sobre essas probabilidades juntamente com os dados observados. É importante para o aprendizado de máquina porque fornece uma abordagem quantitativa para avaliar as evidências que apoiam hipóteses alternativas. O raciocínio bayesiano fornece a base para o aprendizado de algoritmos que manipulam probabilidades diretamente, bem como uma estrutura para analisar a operação de outros algoritmos que não manipulam explicitamente as probabilidades (MITCHELL, 1997).

Os métodos de aprendizagem Bayesianos são relevantes para o estudo de aprendizagem de máquina por duas razões diferentes. Primeiro, os algoritmos de aprendizagem Bayesianos que calculam probabilidades explícitas para hipóteses, como o classificador *Naive Bayes*, estão entre as abordagens mais práticas para certos tipos de problemas de aprendizagem. Por exemplo, Michie, Spiegelhalter e

Taylor (1994) forneceram um estudo detalhado comparando o classificador *Naive Bayes* com outros algoritmos de aprendizagem, incluindo árvore de decisão e algoritmos de rede neural. Esses pesquisadores mostram que o classificador *Naive Bayes* é competitivo com esses outros algoritmos de aprendizagem em muitos casos e que, em alguns casos, supera esses outros métodos. Para tais tarefas de aprendizagem, o classificador *Naive Bayes* está entre os algoritmos mais eficazes conhecidos.

A segunda razão pela qual os métodos bayesianos são importantes para o estudo de aprendizado de máquina é que eles fornecem uma perspectiva útil para a compreensão de muitos algoritmos de aprendizado que não manipulam probabilidades explicitamente. Também se utiliza uma análise Bayesiana para justificar escolhas importantes de projeto em algoritmos de aprendizagem de redes neurais, como por exemplo: escolher minimizar a soma dos erros quadráticos ao pesquisar o espaço de possíveis redes neurais. Também se deriva uma função de erro alternativa, a entropia cruzada, que é mais apropriada do que a soma dos erros quadráticos ao aprender funções alvo que predizem probabilidades.

Usa-se uma perspectiva bayesiana para analisar o viés indutivo dos algoritmos de aprendizagem de árvores de decisão que favorecem árvores de decisão pequenas e examinar o princípio da descrição de comprimento mínimo intimamente relacionado. Uma familiaridade básica com os métodos Bayesianos é importante para compreender e caracterizar a operação de muitos algoritmos em aprendizado de máquina. Os recursos dos métodos de aprendizagem bayesianos incluem (MITCHELL, 1997):

- Cada exemplo de treinamento observado pode diminuir ou aumentar gradativamente a probabilidade estimada de que uma hipótese esteja correta. Isso fornece uma abordagem de aprendizagem mais flexível do que algoritmos que eliminam completamente uma hipótese se ela for considerada inconsistente com um único exemplo.
- O conhecimento prévio pode ser combinado com dados observados para determinar a probabilidade final de uma hipótese. Na aprendizagem bayesiana, o conhecimento prévio é fornecido pela afirmação de que há uma probabilidade anterior para cada hipótese candidata e de que há uma distribuição de probabilidade sobre os dados observados para cada

hipótese possível.

- Os métodos bayesianos podem acomodar hipóteses que fazem previsões probabilísticas.
- Novas instâncias podem ser classificadas combinando as previsões de múltiplas hipóteses, ponderadas por suas probabilidades.
- Mesmo nos casos em que os métodos bayesianos se mostram computacionalmente intratáveis, eles podem fornecer um padrão de tomada de decisão ideal contra o qual outros métodos práticos possam ser medidos.

No aprendizado de máquina, muitas vezes se estar interessado em determinar a melhor hipótese de algum espaço $\boldsymbol{H}$ de hipóteses, tendo os dados de treinamento observados $\boldsymbol{D}$. Quando se diz melhor hipótese se quer dizer que é exigida a hipótese mais provável, dados os dados $\boldsymbol{D}$, além de qualquer conhecimento inicial sobre as probabilidades anteriores das várias hipóteses no teorema de $\boldsymbol{H}$. Bayes fornece um método direto para calcular tais probabilidades. Mais precisamente, o teorema de Bayes fornece uma maneira de calcular a probabilidade de uma hipótese com base em sua probabilidade a *priori*, nas probabilidades de observar vários dados, dada a hipótese, e nos próprios dados observados.

Para definir o teorema de Bayes com precisão, será introduzida uma notação. $\boldsymbol{P}(\boldsymbol{h})$ será utilizada para denotar a probabilidade inicial que a hipótese $\boldsymbol{h}$ possui, antes que os dados de treinamento sejam observados. $\boldsymbol{P}(\boldsymbol{h})$ é frequentemente chamada de probabilidade a *priori* de $\boldsymbol{h}$ e pode refletir qualquer conhecimento prévio que se tenha sobre a chance de $\boldsymbol{h}$ ser uma hipótese correta. Se esse conhecimento prévio não existir, pode-se simplesmente atribuir a mesma probabilidade a *priori* a cada hipótese candidata. Da mesma forma, será escrito $\boldsymbol{P}(\boldsymbol{D})$ para denotar a probabilidade a *priori* dos dados de treinamento $\boldsymbol{D}$ que serão observados (ou seja, a probabilidade de $\boldsymbol{D}$, sem conhecimento prévio sobre qual hipótese é válida). A seguir, será escrito $\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{h})$ para denotar a probabilidade de observar os dados $\boldsymbol{D}$ dado algum espaço no qual a hipótese $\boldsymbol{h}$ é válida. De forma mais geral, será escrito $\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{y})$ para denotar a probabilidade de $\boldsymbol{x}$ dado $\boldsymbol{y}$. Em problemas de aprendizado de máquina, se estar interessado na probabilidade $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$ que $\boldsymbol{h}$ é válida dado os dados de treinamento observados $\boldsymbol{D}$. $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$ é chamada de probabilidade a *posteriori* de $\boldsymbol{h}$, porque reflete a

confiança de que $\boldsymbol{h}$ é válida depois que os dados de treinamento $\boldsymbol{D}$ forem observados. Observe que a probabilidade *a posteriori* $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$ reflete a influência dos dados de treinamento $\boldsymbol{D}$, em contraste com a probabilidade *a priori* $\boldsymbol{P}(\boldsymbol{h})$, que é independente de $\boldsymbol{D}$.

O teorema de Bayes é a pedra angular dos métodos de aprendizagem Bayesianos porque fornece uma maneira de calcular a probabilidade *a posteriori* $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$, a partir da probabilidade *a priori* $\boldsymbol{P}(\boldsymbol{h})$, juntamente com $\boldsymbol{P}(\boldsymbol{D})$ e $\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{h})$. Sendo assim, o teorema de Bayes pode ser visto na equação (2.21):

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{2.21}$$

Como se poderia esperar intuitivamente, $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$ aumenta com o aumento de $\boldsymbol{P}(\boldsymbol{h})$ e com $\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{h})$ de acordo com o teorema de Bayes, ou seja, são diretamente proporcionais. Também é razoável ver que $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$ diminui à medida que $\boldsymbol{P}(\boldsymbol{D})$ aumenta, pois são inversamente proporcionais, porque quanto mais provável for que $\boldsymbol{D}$ será observado independentemente de $\boldsymbol{h}$, menos evidências $\boldsymbol{D}$ fornecerá a $\boldsymbol{h}$.

Qualquer hipótese com maior probabilidade é chamada de estimativa máxima *a posteriori* - *Maximum A Posteriori* (MAP). Pode-se determinar as hipóteses MAP usando o teorema de Bayes para calcular a probabilidade *a posteriori* de cada hipótese candidata. Mais precisamente, diz-se que o y_{MAP} é uma hipótese MAP desde que

$$\begin{aligned} h_{MAP} &\equiv \arg\max_{h \in H} P(h|D) \\ h_{MAP} &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ h_{MAP} &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned} \tag{2.22}$$

Na qual $\arg\max_{h \in H}$ retorna a classe $\boldsymbol{h}$, com maior probabilidade de estar associada aos dados observados $\boldsymbol{D}$, que é aquela que possui o valor máximo para $\boldsymbol{P}(\boldsymbol{h}|\boldsymbol{D})$. Como o termo $\boldsymbol{P}(\boldsymbol{D})$ é uma constante independente de $\boldsymbol{h}$, pode ser ignorado. Ele é o mesmo para todas as classes, não afetando os valores relativos de suas

probabilidades. Em alguns casos, assumindo que as probabilidades *a priori* das hipóteses $\boldsymbol{h_i}$ são iguais, a equação (2.21) pode ser simplificada considerando apenas o termo das hipóteses $\boldsymbol{P(D|h)}$ para calcular a hipótese mais provável. O termo $\boldsymbol{P(D|h)}$ é frequentemente chamado por verossimilhança dos dados $\boldsymbol{D}$ dado $\boldsymbol{h}$, e qualquer hipótese que maximiza $\boldsymbol{P(D|h)}$ é chamada de hipótese de máxima verossimilhança, que pode ser escrita como:

$$h_{MV} \equiv \arg\max_{h \in H} P(D|h) \tag{2.22}$$

Um método de aprendizagem bayesiana altamente prático é o *Naive Bayes*, muitas vezes chamado de classificador *Naive Bayes*. Em alguns domínios, seu desempenho demonstrou ser comparável ao da aprendizagem de redes neurais e de árvores de decisão. O classificador *Naive Bayes* se aplica a tarefas de aprendizagem onde cada instância $\boldsymbol{x}$ é descrita por uma combinação de valores de atributos e onde a função alvo $\boldsymbol{f(x)}$ pode assumir qualquer valor de algum conjunto finito V. Um conjunto de exemplos de treinamento da função alvo é fornecido e uma nova instância é apresentada, descrita pela tupla de valores dos atributos $a_1, a_2 \ldots a_n$. O classificador é solicitado a prever o valor alvo, ou classificação, para esta nova instância.

A abordagem bayesiana para classificar a nova instância é atribuir o valor alvo mais provável, v_{MAP}, dado os valores dos atributos $a_1, a_2 \ldots a_n$ que descrevem a instância.

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j|a_1, a_2 \ldots a_n) \tag{2.23}$$

Usando o teorema de Bayes para reescrever a expressão, tem-se:

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n|v_j)P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2 \ldots a_n|v_j)P(v_j) \tag{2.24}$$

Poder-se-ia estimar os dois termos da equação (2.24) com base nos dados de treinamento, para cada um dos $P(v_j)$ simplesmente contando a frequência com

que cada valor alvo v_j ocorre nos dados de treinamento. No entanto, estimar os diferentes termos $P(a_1, a_2 \ldots a_n|v_j)$ desta forma não é viável, a menos que se tenha um conjunto muito, muito grande de dados de treinamento. O problema é que o número desses termos é igual ao número de instâncias possíveis multiplicado pelo número de valores alvo possíveis. Portanto, é preciso ver cada instância no espaço de instâncias muitas vezes para obter estimativas confiáveis.

O classificador *Naive Bayes* é baseado na suposição de que os valores dos atributos são independentes entre si, dado o valor alvo. Em outras palavras, a suposição é que, dado o valor alvo da instância, a probabilidade de observar os termos $a_1, a_2 \ldots a_n$ é apenas o produto das probabilidades dos atributos individuais: $P(a_1, a_2 \ldots a_n|v_j) = \prod_i P(a_i|v_j)$. Substituindo na equação (2.24), tem-se a abordagem utilizada pelo classificador *Naive Bayes* dada pela equação:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \tag{2.25}$$

onde v_{NB} denota o valor alvo gerado pelo classificador *Naive Bayes*. Observe que em um classificador *Naive Bayes*, o número de termos distintos $P(a_i|v_j)$ que devem ser estimados a partir dos dados de treinamento é apenas o número de valores de atributos distintos multiplicado pelo número de valores alvo distintos, um número muito menor do que se fosse para estimar os termos $P(a_1, a_2 \ldots a_n|v_j)$ conforme contemplados pela primeira vez.

Para resumir, o método de classificação *Naive Bayes* envolve uma etapa de aprendizagem na qual os vários termos $P(v_j)$ e $P(a_i|v_j)$ são estimados, com base em suas frequências nos dados de treinamento. O conjunto dessas estimativas corresponde à hipótese aprendida. Esta hipótese é então usada para classificar cada nova instância aplicando a regra da Equação (2.25). Sempre que a suposição de independência condicional é satisfeita, o classificador *Naive Bayes* v_{NB} é idêntica à classificação MAP.

Uma diferença interessante entre o método de aprendizagem *Naive Bayes* e outros métodos de aprendizagem é que não há busca explícita no espaço de hipóteses possíveis (neste caso, o espaço de hipóteses possíveis é o espaço de valores possíveis que podem ser atribuídos aos vários termos $P(v_j)$ e $P(a_i|v_j)$. Em

vez disso, a hipótese é formada sem pesquisa, simplesmente contando a frequência de várias combinações de dados nos exemplos de treinamento.

O desempenho do modelo pode ser ajustado de acordo com as preferências individuais com base na aplicação. Pesquisa em grade, pesquisa aleatória e otimização baseada em modelo sequencial podem ser implementadas para otimização de hiperparâmetros (ERICKSON et al., 2017; SARTIAS; YASAR, 2019; GANDHI, 2018).

2.2.1.3 *K-Nearest Neighbor* (KNN)

Os métodos de aprendizagem baseados em instâncias consistem simplesmente em armazenar os dados de treinamento apresentados. Quando uma nova instância de consulta é encontrada, um conjunto de instâncias relacionadas semelhantes é recuperado da memória e usado para classificar a nova instância de consulta. Esses métodos podem construir uma aproximação diferente para a função alvo para cada instância de consulta distinta que deve ser classificada. Na verdade, muitas técnicas constroem apenas uma aproximação local para a função alvo que se aplica na vizinhança da nova instância de consulta, e nunca constroem uma aproximação projetada para ter um bom desempenho em todo o espaço de instância.

O método mais básico baseado em instância é o algoritmo KNN. Este algoritmo assume que todas as instâncias correspondem a pontos no espaço n-dimensional R^n, ou seja, que seus atributos são numéricos (contínuos). Os vizinhos mais próximos de uma instância são definidos em termos da distância euclidiana padrão. Mais precisamente, sendo uma instância arbitrária x ser descrita pelo vetor de características $\{a_1(x), a_2(x), \ldots a_n(x)\}$, onde $a_r(x)$ denota o valor do *r*-ésimo atributo da instância x. Então a distância entre as duas instâncias x_i e x_j é definida como $d(x_i, x_j)$, onde

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{2.26}$$

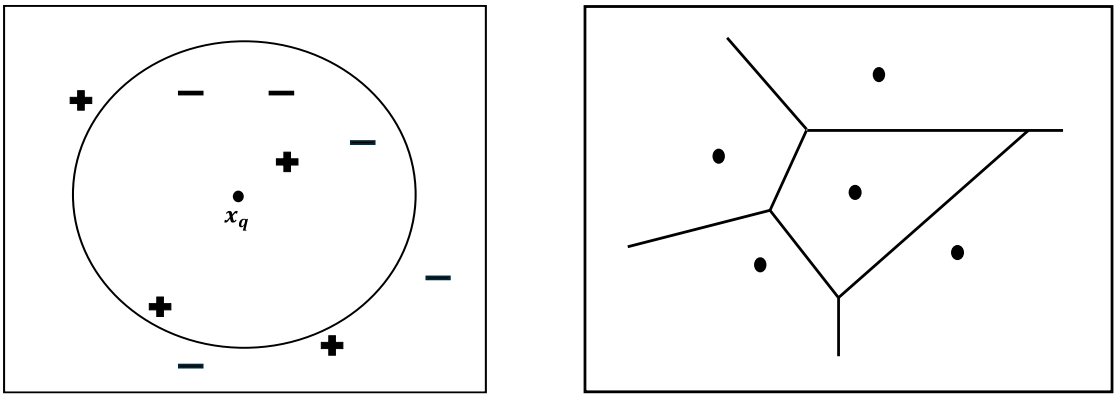
Na aprendizagem pelo vizinho mais próximo, a função alvo pode ter valor discreto ou valor real. Considerando primeiramente o aprendizado de funções alvo

com valor discreto da forma $f: R^n \rightarrow V$, onde V é o conjunto finito $\{v_1, v_2 \ldots v_s\}$. O algoritmo KNN para aproximar de uma função alvo com valor discreto é descrito da seguinte maneira. Durante o treinamento, para cada exemplo de treinamento $\{x, f(x)\}$ adicione o exemplo à lista de exemplos de treinamento. Durante a classificação, dada uma instância x_q a ser classificada, e considerando $\{x_1, x_2 \ldots x_k\}$ como as *k* instâncias dos exemplos de treinamento que estão mais próximas de x_q, então

$$\hat{f}(x_q) \leftarrow arg \max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i)) \tag{2.27}$$

onde $\delta(v, f(x_i))$ = 1 se $v = f(x_i)$ e $\delta(v, f(x_i))$ = 0 caso contrário. A Equação (2.27) retorna a estimativa de $f(x_q)$ que é apenas o valor mais comum de f entre os *k* exemplos de treinamento mais próximos de x_q. A Figura 7 ilustra a operação do algoritmo KNN para o caso em que as instâncias são pontos em um espaço bidimensional e onde a função alvo tem valor booleano. Os exemplos de treinamento positivos e negativos são mostrados por "+" e "-" respectivamente. Um ponto de consulta x_q também é mostrado. Observe que o algoritmo 1-NN classifica x_q, como um exemplo positivo na Figura 7, enquanto o algoritmo 5-NN o classifica como um exemplo negativo.

Figura 7 – Superfície de decisão: diagrama de *Voronoi*



Fonte: FACELI *et al.* (2021)

O algoritmo KNN nunca forma uma hipótese geral explícita $\hat{f}$ em relação à função alvo f. Ele simplesmente calcula a classificação de cada nova instância de consulta conforme necessário. No entanto, ainda se pode perguntar qual é a função

geral implícita, ou que classificações seriam atribuídas se fossem mantidos os exemplos de treinamento constantes e executasse o algoritmo com todas as instâncias possíveis em ***X***. O diagrama no lado direito da Figura 7 mostra a forma desta superfície de decisão induzida por 1-NN em todo o espaço de instância. A superfície de decisão é uma combinação de poliedros convexos que circundam cada um dos exemplos de treinamento. Para cada exemplo de treinamento, o poliedro indica o conjunto de pontos de consulta cuja classificação será completamente determinada por aquele exemplo de treinamento. Os pontos de consulta fora do poliedro estão mais próximos de algum outro exemplo de treinamento. Esse tipo de diagrama costuma ser chamado de *diagrama de Voronoi* para o conjunto de exemplos de treinamento.

O algoritmo KNN é facilmente adaptado para aproximar funções alvo de valor contínuo. Para conseguir isso, faz-se com que o algoritmo calcule o valor médio dos *k* exemplos de treinamento mais próximos, em vez de calcular seu valor mais comum. Mais precisamente, para aproximar uma função alvo com valor real $f: R^n \rightarrow R^n$ é dado pela Equação (2.28):

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} f(x_i)}{k} \tag{2.28}$$

A escolha do valor de *k* mais apropriado para um problema de decisão específico pode não ser trivial. O valor de *k* é definido pelo usuário. Frequentemente, o valor de *k* é pequeno e ímpar. Em problemas de classificação, não é usual utilizar valores pares, para evitar empates.

Um refinamento óbvio do algoritmo KNN é ponderar a contribuição de cada um dos *k* vizinhos de acordo com sua distância até o ponto de consulta x_q, dando maior peso aos vizinhos mais próximos. Por exemplo, na Equação (2.27), que aproxima funções alvo de valor discreto, o voto de cada vizinho pode ser ponderado (w_i) pelo inverso do quadrado da sua distância de x_q. Desta forma, substituindo na Equação (2.27), ficaria da seguinte maneira:

$$\hat{f}(x_q) \leftarrow arg \max_{v \in V} \sum_{i=1}^{k} w_i \delta(v, f(x_i)) \tag{2.28}$$

onde

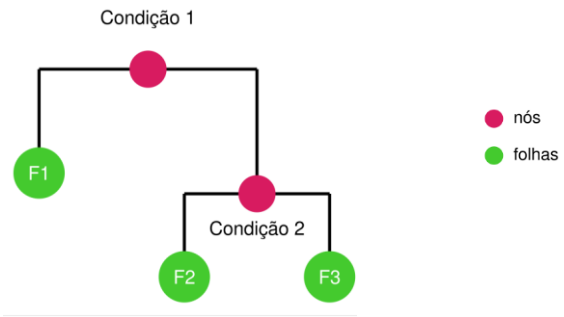
$$w_i \equiv \frac{1}{d(x_q, x_i)^2} \tag{2.29}$$

2.2.1.4 Árvores de Decisão

O aprendizado da árvore de decisão é um método para aproximar funções alvo de valor discreto, no qual a função aprendida é representada por uma árvore de decisão. Esses métodos de aprendizagem estão entre os mais populares algoritmos de inferência indutiva e são aplicados com sucesso a uma ampla gama de tarefas, desde aprender a diagnosticar casos médicos até aprender a avaliar o risco de crédito de solicitantes de empréstimos.

As árvores de decisão classificam as instâncias, desde a raiz até algum nó folha. Cada nó da árvore especifica um teste de algum atributo da instância, e cada ramo descendente desse nó corresponde a um dos valores possíveis para este atributo. Este processo é então repetido para a sub-árvore enraizada no novo nó. A Figura 8 ilustra um exemplo de uma árvore de decisão.

Figura 8 – Árvore de decisão



Fonte: Elaborada pelo autor

A utilização da árvore para prever uma nova observação é feita do seguinte modo: começando pelo topo (raiz), é verificado se a condição descrita no topo (primeiro nó) é satisfeita. Caso seja, segue-se para a esquerda. Caso contrário, segue-se para a direita. Assim prossegue-se até atingir uma folha. No caso ilustrativo da Figura 8, se a condição 1 for satisfeita, a predição é dada por F1. Caso não seja satisfeita, é encontrada outra condição (condição 2). Caso ela seja satisfeita, a

observação é prevista como F2 e, caso contrário, é prevista como F3.

A criação da estrutura de uma árvore é feita através de duas grandes etapas: (i) a criação de uma árvore completa e complexa e (ii) a poda dessa árvore, com a finalidade de evitar *overfitting*.

Questões práticas no aprendizado de árvores de decisão incluem determinar a profundidade do crescimento da árvore de decisão, lidar com atributos contínuos, escolher uma medida de seleção de atributos apropriada, lidar com dados de treinamento com valores de atributos ausentes, lidar com atributos com custos diferentes e melhorar a eficiência computacional.

Embora uma variedade de métodos de aprendizado de árvores de decisão tenha sido desenvolvida com capacidades e requisitos um tanto diferentes, o aprendizado de árvores de decisão geralmente é mais adequado para problemas com as seguintes características (MITCHELL, 1997):

- As instâncias são representadas por pares atributo-valor. As instâncias são descritas por um conjunto fixo de atributos e seus valores. A situação mais fácil para o aprendizado da árvore de decisão é quando cada atributo assume um pequeno número de valores possíveis disjuntos.
- A função alvo possui valores de saída discretos. A árvore de decisão atribui uma classificação booleana (por exemplo, sim ou não) a cada exemplo. Os métodos de árvore de decisão estendem-se facilmente a funções de aprendizagem com mais de dois valores de saída possíveis. Uma extensão mais substancial permite aprender funções alvo com saídas de valor real, embora a aplicação de árvores de decisão neste cenário seja menos comum.
- Os dados de treinamento podem conter erros. Os métodos de aprendizagem de árvores de decisão são robustos a erros, tanto erros nas classificações dos exemplos de treinamento quanto erros nos valores dos atributos que descrevem esses exemplos.
- Os dados de treinamento podem conter valores de atributos ausentes. Os métodos de árvore de decisão podem ser usados mesmo quando alguns exemplos de treinamento têm valores desconhecidos.

A maioria dos algoritmos desenvolvidos para aprender árvores de decisão são

variações de um algoritmo central que emprega uma busca ambiciosa ou gulosa de cima para baixo no espaço de possíveis árvores de decisão. Esta abordagem é exemplificada pelo algoritmo ID3 (QUINLAN, 1986).

O algoritmo básico, ID3, aprende árvores de decisão construindo-as de cima (raiz) para baixo (folha), começando com a pergunta "qual atributo deve ser testado na raiz da árvore?". O melhor atributo é selecionado e usado como teste no nó raiz da árvore. Um descendente do nó raiz é então criado para cada valor possível deste atributo, e os exemplos de treinamento são classificados para o nó descendente apropriado (ou seja, abaixo do ramo correspondente ao valor do exemplo para este atributo). Todo o processo é então repetido usando os exemplos de treinamento associados a cada nó descendente para selecionar o melhor atributo a ser testado naquele ponto (nó) da árvore. Isso forma uma busca gulosa por uma árvore de decisão aceitável, na qual o algoritmo nunca recua para reconsiderar escolhas anteriores.

A escolha central no algoritmo ID3 é selecionar qual atributo testar em cada nó da árvore. Seria bom selecionar o atributo mais útil para classificar os dados. Desta maneira, uma forma para fazer isso é definir uma propriedade estatística, chamada *ganho de informação*, que mede quão bem um determinado atributo separa os dados de treinamento de acordo com sua classificação alvo (no caso, a classe à qual pertence). O ID3 usa essa medida de ganho de informação para selecionar entre os atributos candidatos em cada etapa durante a construção da árvore.

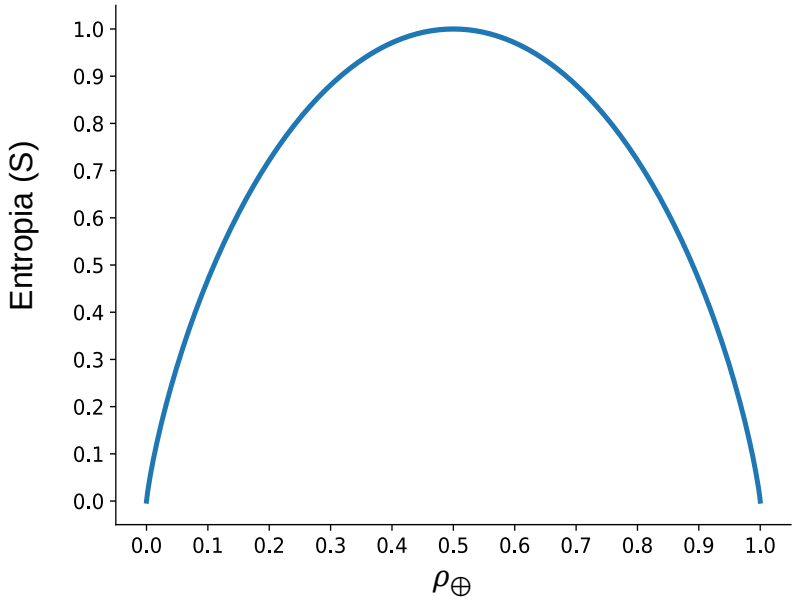
Para definir com precisão o ganho de informação, é necessário, primeiramente, definir uma medida comumente usada na teoria da informação, chamada *entropia*, que caracteriza a pureza ou impureza de uma coleção arbitrária de exemplos. A Entropia mede a aleatoriedade (dificuldade para predizer) de uma variável aleatória. Dada uma coleção S, contendo exemplos positivos e negativos de algum conceito alvo, a entropia de S relativa a esta classificação booleana é dada pela equação:

$$Entropia(S) \equiv -\rho_{\oplus} \log_2 \rho_{\oplus} - \rho_{\ominus} \log_2 \rho_{\ominus} \tag{2.30}$$

Onde $\rho_{\oplus}$ é a proporção de exemplos positivos em S e $\rho_{\ominus}$ é a proporção de exemplos negativos em S. Em todos os cálculos que envolvem entropia é definido $0\log 0$ como 0 (zero). A entropia é 0 se todos os membros de S pertencem à mesma

classe. A entropia é 1 quando a coleção contém um número igual de exemplos positivos e negativos. Se a coleção contiver números desiguais de exemplos positivos e negativos, a entropia estará entre 0 e 1. A entropia é medida em *bits* usando logaritmos na base 2. A Figura 9 ilustra a forma da função entropia relativa a uma classificação booleana, pois $\rho_{\oplus}$ varia entre 0 e 1.

Figura 9 - A função de entropia relativa a uma classificação booleana, com a proporção, $\rho_{\oplus}$, de exemplos positivos variando entre 0 e 1



Fonte: Elaborada pelo autor

Uma interpretação da entropia da teoria da informação é que ela especifica o número mínimo de bits de informação necessários para codificar a classificação de um membro arbitrário de S (isto é, um membro de S sorteado aleatoriamente com probabilidade uniforme). Por exemplo, se $\rho_{\oplus}$ for 1, o receptor sabe que o exemplo desenhado será positivo, portanto, nenhuma mensagem precisa ser enviada e a entropia é zero. Por outro lado, se $\rho_{\oplus}$ for 0,5, é necessário um bit para indicar se o exemplo desenhado é positivo ou negativo. Se $\rho_{\oplus}$ for 0,8, então uma coleção de mensagens pode ser codificada usando, em média, menos de 1 bit por mensagem, atribuindo códigos mais curtos a coleções de exemplos positivos e códigos mais longos a exemplos negativos menos prováveis.

Uma árvore de decisão divide os nós com base na impureza ou no erro do nó. Existem alguns critérios de divisão em árvores de decisão para calcular a impureza.

Dentre os quais se destacam o índice de diversidade de Gini - *Gini Diversity Index* (GDI) e o desvio (*deviance*).

O GDI de um nó é dado pela equação

$$1 - \sum_i \rho^2(i) \tag{2.31}$$

Onde a soma é das classes i no nó, e $\rho(i)$ é a fração observada de classes, das classes i que chegam ao nó. Um nó com apenas uma classe (um nó “puro”) possui GDI igual a 0 (zero); caso contrário, o GDI é positivo. Portanto, o GDI é uma medida da impureza do nó.

O desvio (*deviance*) é dado pela equação

$$-\sum_i \rho(i) \log_2 \rho(i) \tag{2.32}$$

Onde $\rho(i)$ é definido da mesma forma que o GDI. Um nó puro possui desvio igual a 0 (zero); caso contrário, o desvio é positivo.

Existe uma medida diferente para decidir como dividir um nó, chamada de regra de dois (*twoing*), mas ela não é uma medida de pureza de um nó. Sendo $L(i)$ a fração de membros da classe i no nó filho esquerdo após uma divisão, e $R(i)$ a fração de membros da classe i no nó filho direito após uma divisão. Escolha o critério de divisão para maximizar a equação

$$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2 \tag{2.33}$$

Onde $P(L)$ e $P(R)$ são as frações de observações que se dividem à esquerda e à direita, respectivamente. Se a expressão for grande, a divisão tornará cada nó filho mais puro. Da mesma forma, se a expressão for pequena, a divisão tornará cada nó filho semelhante entre si e, portanto, semelhante ao nó pai. A divisão não aumentou a pureza do nó.

O erro do nó é a fração de classes classificadas incorretamente em um nó. Se j for a classe com o maior número de amostras de treinamento em um nó, o erro do

nó será dado por:

$$1 - \rho(j) \tag{2.34}$$

A poda tem como objetivo tornar a árvore menor e menos complexa, de modo a diminuir a variância do estimador, e também evita o *overfitting*. Nessa etapa do processo cada nó é retirado, um por vez, e observa-se como o erro de predição varia no conjunto de validação. Com base nisso, decide-se quais 'nós' permanecerão na árvore.

A poda de um nó de decisão consiste em remover a subárvore enraizada naquele nó, tornando-a um nó folha e atribuindo-lhe a classificação mais comum dos dados de treinamento vinculados a esse nó. Os nós serão removidos somente se a árvore podada resultante não tiver um desempenho pior do que o original no conjunto de validação. Isso tem o efeito de que qualquer nó folha adicionado devido a regularidades coincidentes no conjunto de treinamento provavelmente será removido, porque é improvável que essas mesmas coincidências ocorram no conjunto de validação. Os nós são podados iterativamente, sempre escolhendo o nó cuja remoção mais aumenta a precisão da árvore de decisão em relação ao conjunto de validação. A poda de nós continua até que uma poda adicional seja prejudicial (ou seja, diminui a precisão da árvore em relação ao conjunto de validação). Podar, em geral, leva a erros menores de generalização.

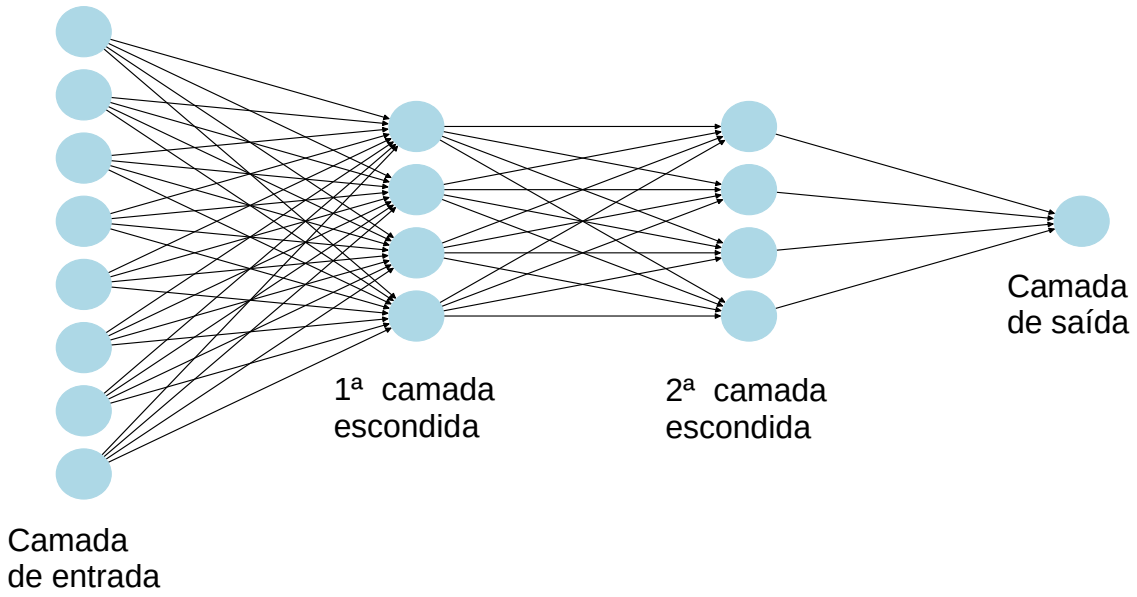
2.2.1.5 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são dispositivos não-lineares, inspirados na funcionalidade dos neurônios biológicos, aplicados no reconhecimento de padrões, problemas de regressão, classificação e compactação de dados, na otimização e na previsão de sistemas complexos. Além disso, são aplicadas frequentemente em situações em que existem interações não lineares entre as variáveis dependentes e independentes.

As RNAs são compostas por diversas unidades computacionais paralelas, interconectadas parcial ou totalmente. Cada uma dessas unidades, chamadas de neurônios artificiais, efetuam um certo número de operações simples e transmite seus

resultados às unidades vizinhas com as quais possui conexão. Os neurônios são dispostos em uma ou mais camadas e interligadas por muitas conexões, geralmente unidirecionais. Na maioria das arquiteturas, essas conexões, que simulam as sinapses biológicas, possuem pesos associados, que ponderam a entrada recebida por cada neurônio da rede. Os pesos podem assumir valores positivos ou negativos, e têm seus valores ajustados em um processo de aprendizado e codificam o conhecimento adquirido pela rede (BRAGA; CARVALHO; LUDERMIR, 2007). A Figura 10, mostra uma arquitetura de uma rede neural.

Figura 10 – Arquitetura de uma rede neural com uma camada de entrada, duas camadas escondidas e uma camada de saída



Fonte: Elaborada pelo autor

O processamento básico de informação da rede ocorre nos neurônios. Através de um processo de treinamento, as redes neurais passam a ser capazes de reconhecer padrões, mesmo que os dados utilizados nesse treinamento sejam não-lineares, incompletos ou até mesmo contraditórios.

A habilidade de manipular esses dados imprecisos faz com que as redes neurais sejam extremamente eficazes em tarefas onde especialistas não estão à disposição ou um conjunto de regras não pode ser facilmente formulado. No caso da visão por computador, tanto a dificuldade de estabelecer algoritmos de reconhecimento, quanto à natureza dos dados com os quais se trabalha, indicam as

redes neurais como uma solução promissora.

A implementação de uma ferramenta para diagnosticar possíveis problemas de saúde usando visão artificial baseado em redes neurais faz uso de suas características intrínsecas, isto é, da sua capacidade de extrair padrões de conjuntos de dados complexos.

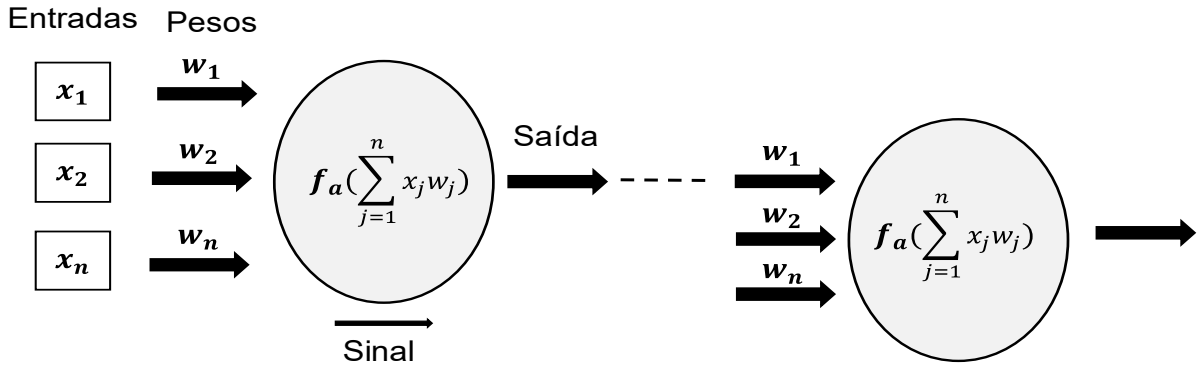
Pode-se dizer então que a visão artificial procura “compreender” a informação contida num sinal com o propósito de classificação, caracterização e/ou reconstrução dele.

Uma RNA é caracterizada principalmente por cinco fatores (HAYKIN, 1999):

- ***Arquitetura da RNA –*** determinar os parâmetros para construção da RNA.
- ***Topologias*** - organização dos neurônios nas camadas, o número de neurônios, o grau de conectividade e os tipos de conexões permitidas entre eles.
- ***Tipo de aprendizado*** – utilizado para extrair informações relevantes de padrões de informação apresentados pela RNA.
- ***Função de ativação*** - função que excita ou inibe o neurônio.
- ***Conjunto de treinamento e testes*** - métodos utilizados para estimar os parâmetros da RNA e desempenhar uma determinada tarefa.

O neurônio é a unidade de processamento fundamental na arquitetura de uma RNA. Na Figura 11 é apresentado um modelo simples de neurônio artificial. Cada terminal de entrada do neurônio recebe um valor, que é ponderado e processado por uma função matemática f_a. O resultado desta função é a resposta do neurônio para a entrada recebida.

Figura 11 – Neurônio artificial



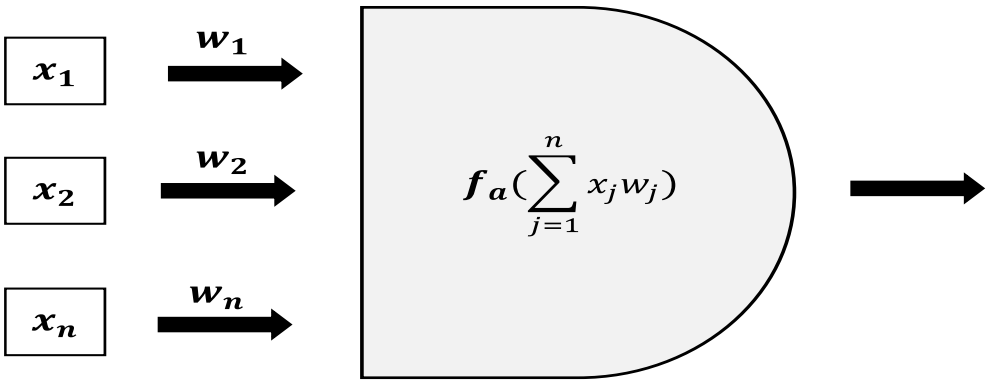
Fonte: Elaborada pelo autor

Supondo que um objeto **X** com *n* atributos, representado vetorialmente como $\boldsymbol{X} = [x_1, x_2, \dots, x_n]^T$, e um neurônio com *n* terminais de entrada cujos pesos podem ser representados na forma vetorial como $\boldsymbol{W} = [w_1, w_2, \dots, w_n]$. A entrada total recebida pelo neurônio u é definida pela equação:

$$u = \sum_{j=1}^{n} x_j w_j \tag{2.35}$$

A saída de um neurônio é definida por meio de uma função de ativação à entrada total, conforme ilustrado na Figura 12.

Figura 12 – Entrada total de um neurônio artificial



Fonte: Elaborada pelo autor

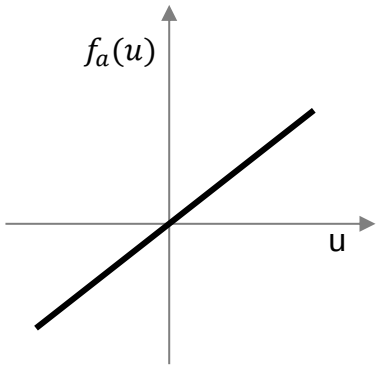
Existem várias funções de ativação propostas na literatura, destacando-se entre elas: linear, limiar, sigmoide e linear retificada.

- **Linear**: a saída do neurônio é igual à entrada, sendo normalmente utilizada na camada de saída de redes neurais, onde é necessário manter a proporcionalidade dos valores.

$$f_a(u) = u \tag{2.36}$$

Esta função pode ser visualizada na Figura 13.

Figura 13 - Função de ativação linear



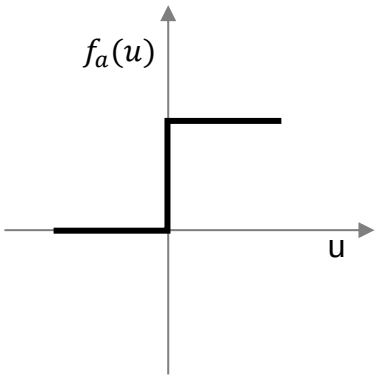
Fonte: FACELI *et al.* (2021)

- **Limiar**: é uma das mais simples e intuitivas. Ela funciona de maneira binária, onde a saída do neurônio é igual a 0 (zero) quando seu valor de entrada é negativo, e 1 (um) quando o valor de entrada é positivo.

$$f_a(u) = \begin{cases} 1, u \geq 0 \\ 0, u < 0 \end{cases} \tag{2.37}$$

Esta função pode ser visualizada na Figura 14.

Figura 14 - Função de ativação limiar



Fonte: FACELI *et al.* (2021)

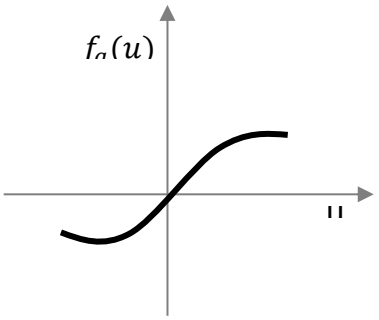
- **Sigmoide:** é definida como uma função crescente, que apresenta um balanço entre o comportamento linear e não-linear. Um exemplo de função sigmoide é a função tangente hiperbólica. Ela mapeia os valores de entrada para um intervalo entre -1 e 1, o que a torna especialmente útil em contextos em que é necessário normalizar os dados, definida pela

equação:

$$f_a(u) = \frac{1 - e^{-\alpha u}}{1 + e^{+\alpha u}} \tag{2.38}$$

Onde α é o parâmetro de inclinação da função. Esta função pode ser visualizada na Figura 15.

Figura 15 - Função de ativação sigmoide



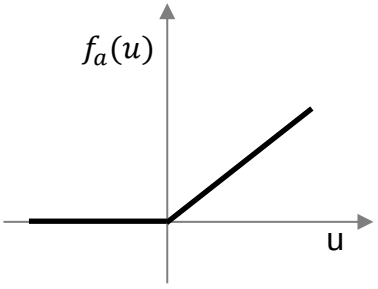
Fonte: FACELI *et al.* (2021)

- **Unidade linear retificada - Rectified Linear Unit (ReLU)**, é uma das mais populares e amplamente utilizadas em algoritmos de ML. Ele executa uma operação de limite não linear, onde qualquer valor de entrada menor ou igual a zero é definido como zero, enquanto valores positivos são mantidos inalterados. A operação é equivalente a

$$f_a(u) = \begin{cases} u, u > 0 \\ 0, u \leq 0 \end{cases} \tag{2.39}$$

Esta função pode ser visualizada na Figura 16.

Figura 16 – Função de ativação ReLU



Fonte: Elaborada pelo autor

Em uma RNA, os neurônios podem estar dispostos em uma ou mais camadas, conforme ilustrado anteriormente na Figura 10. A camada de entrada refere-se aos valores dos atributos que são utilizados como entrada pela RNA. A camada intermediária, também chamada de oculta ou escondida, recebe em seus terminais de entrada valores de saída dos neurônios da camada anterior e enviam seu valor de saída para a camada seguinte. A camada intermediária pode ser composta tanto de uma única camada quanto pode possuir mais de uma camada. A camada de Saída é composta pelos valores de saída gerados pela RNA.

Em uma RNA multicamadas, as conexões entre os neurônios podem apresentar diferentes padrões de conexão, e podem ser classificadas em (HAYKIN, 1999):

- **Completamente conectada** - quando os neurônios estão conectados a todos os neurônios da camada anterior e posterior.
- **Parcialmente conectada** - quando os neurônios estão conectados a apenas alguns dos neurônios da camada anterior e/ou posterior.
- **Localmente conectada**, são RNAs parcialmente conectadas, em que os neurônios se encontram em uma região bem definida.

Com relação a propagação das informações, as RNAs são classificadas em (HAYKIN, 1999):

- ***Feedback*** - permite que os neurônios recebam em seus terminais de entrada a saída de neurônios de uma camada posterior ou até da mesma camada.
- ***Feedforward*** - são RNAs que não permitem que os neurônios recebam retroalimentação em seus terminais de entrada.

Em relação ao aprendizado, que se dá na fase de treinamento da RNA, onde é realizado o ajuste dos parâmetros, principalmente em relação aos pesos $[w_1, w_2, \dots, w_n]$ associados às conexões, que fazem com que o modelo obtenha melhor desempenho, geralmente medido pela acurácia preditiva, alguns algoritmos foram propostos, tais como: correção de erro, Hebbiano, competitivo e termodinâmico (Boltzman). Dentre estes, se destaca o algoritmo de correção de erro, para o aprendizado supervisionado.

O processo de treinamento de várias RNAs geralmente é baseado na regra

delta, que é constituída da iteração de duas fases, uma fase *forward* e uma fase *backward*. Na fase *forward*, cada objeto de entrada é apresentado à RNA. O objeto é primeiramente recebido por cada um dos neurônios da primeira camada intermediária, assim é ponderado pelo peso associado às suas conexões de entrada correspondentes. Cada neurônio nessa camada aplica a função de ativação à sua entrada total e produz um valor de saída, que é utilizado como valor de entrada pelos neurônios da camada seguinte. Esse processo continua até que os neurônios da camada de saída cada um seu valor de saída, que é então comparado ao valor desejado para a saída desse neurônio. A diferença entre os valores de saída produzidos e desejados indica o erro cometido pela RNA para o objeto apresentado.

O valor do erro de cada neurônio da camada de saída é então utilizado na fase *backward* para ajustar seus pesos de entrada. O ajuste prossegue da camada de saída até a primeira ou única camada intermediária. A Equação (2.40) mostra como é feito o ajuste dos pesos de uma RNA pelo algoritmo *back-propagation*.

$$w_{jl}(t+1) = w_{jl}(t) + \eta x^j \delta_l \tag{2.40}$$

onde w_{jl} representa o peso entre um neurônio l e o j-ésimo atributo de entrada ou a saída do j-ésimo neurônio da camada anterior, durante a iteração t, δ_l indica o erro associado ao l-ésimo neurônio, e x^j indica a entrada recebida por esse neurônio. η é uma taxa de aprendizado da RNA. O valor da taxa de aprendizado tem uma forte influência no tempo necessário à convergência da RNA. Se a taxa for muito pequena, muitos ciclos poderão ser necessários para induzir um bom modelo. Por outro lado, se a taxa for muito grande, pode provocar oscilações que dificultam a convergência da RNA.

Como os valores dos erros são conhecidos apenas para os neurônios da camada de saída, o erro dos neurônios das camadas intermediárias precisa ser estimado. Uma maneira de fazer isso é utilizando os erros observados na camada subsequente. Esse processo envolve somar os erros da camada seguinte, ponderados pelos pesos das conexões associadas a esses neurônios. Esse método está representado na Equação (2.41).

$$\delta_l = \begin{cases} f_a' e_l, se\ n_l \in C_{sai} \\ f_a' \sum w_{lk}\, \delta_k, se\ n_l \in C_{int} \end{cases} \quad (2.41)$$

Onde n_l é o l-ésimo neurônio, C_{sai} representa a camada de saída, C_{int} representa uma camada intermediária, f_a' é a derivada parcial da função de ativação do neurônio, e e_l é o erro quadrático cometido pelo neurônio de saída quando sua resposta é comparada à desejada, e é definido pela equação:

$$e_l = \frac{1}{2}\sum_{q=1}^{k}(y_q - \hat{f}_q)^2 \quad (2.42)$$

Onde y_q é a saída desejada, $\hat{f}_q$ é a saída produzida.

A derivada parcial define o ajuste dos pesos, utilizando o gradiente descendente da função de ativação. Essa derivada mede a contribuição de cada peso no erro da RNA para a classificação de um dado objeto. Se essa derivada for positiva, o peso está provocando um aumento da diferença entre a saída desejada e a saída produzida. Assim sua magnitude deve ser reduzida para baixar o erro. Se essa derivada for negativa, o peso está contribuindo para que a saída produzida seja mais próxima da saída desejada. Desta forma, seu valor deve ser aumentado.

Ao treinar uma RNA, a inicialização dos pesos e *bias* da camada pode ter um grande impacto no desempenho do treinamento da RNA. Dependendo do tipo de camada, você pode alterar a inicialização dos pesos e *bias*. Existem alguns inicializadores dos pesos ao treinar uma RNA:

- **Glorot** – Os pesos de entrada são inicializados com o inicializador Glorot (GLOROT; BENGIO, 2010), mostrada na Equação (2.43).

$$\left[-\sqrt{\frac{6}{N_o + N_i}}, \sqrt{\frac{6}{N_o + N_i}}\right] \quad (2.43)$$

  Onde o valor de N_o, para RNA totalmente conectada, é o número de classes de saída e N_i o número de atributos de entrada.

- **He** – Inicialização dos pesos de entrada com o inicializador He (HE et al.,

2015). Amostra pesos da distribuição normal com média zero e variância $\frac{2}{N_i}$, onde N_i corresponde ao número de atributos de entrada.

- **Narrow-Normal** – Inicialização dos pesos de entrada por meio de uma distribuição normal com média zero e desvio padrão 0,01.

Bias ou viés em RNAs é um valor que é adicionado a cada neurônio, juntamente com as entradas, antes de ser aplicada uma função de ativação. Esse valor de bias permite que a RNA faça ajustes na função de ativação, deslocando-a para cima ou para baixo. Em outras palavras, o valor do *bias* permite que a RNA aprenda a melhor forma de mapear as entradas para as saídas desejadas. Sem o *bias*, a função de ativação seria sempre centrada em zero, o que limitaria a capacidade da RNA de aprender padrões complexos e não-linearidades nos dados. Existem algumas funções para inicializar o *bias*, especificada como um destes valores:

- **Zeros** — o *bias* é inicializado com zeros.
- **Ones** — o *bias* é inicializado com uns.
- **Narrow-normal** — o *bias* é inicializado por meio de uma distribuição normal com média zero e desvio padrão de 0,01.

Outros parâmetros são configurados durante a fase de treinamento de uma RNA, dentre eles destacam-se:

- ***Lambda*** — também chamado de taxa de regularização, é especificado como um número escalar não negativo. Compõe a função objetivo para minimização a partir da função de perda do erro quadrático médio e do termo de penalidade, que é incluído na função com o objetivo de forçar o conjunto de restrições a encontrar um ponto mínimo local.
- ***Standardize*** — *flag* utilizada para padronizar os dados do preditor. Especificado como numérico ou lógico 0 (falso) ou 1 (verdadeiro). Se for definido como verdadeiro (*true*), o *software* centralizará e dimensionará cada variável preditora numérica pela média e desvio padrão do atributo (variável) correspondente.

Para finalizar o treinamento diferentes critérios de parada podem ser utilizados, como, por exemplo, um número máximo de ciclos (épocas), ou uma taxa máxima de erros. Para que haja convergência, uma superfície de erro minimizada apresenta regiões de mínimos locais e de mínimo global. O objetivo do treinamento é atingir o mínimo global, para que haja uma alta acurácia, se ele ficar preso em um mínimo local, a RNA possuirá uma baixa acurácia preditiva.

2.2.2 Seleção de modelos preditivos

O objetivo de um método de seleção de modelos preditivos é selecionar uma boa função g. Nesse caso, pode ser utilizado o critério do risco quadrático para avaliar a qualidade da função. Logo, escolhe-se uma função g dentro de uma classe de candidatas $\boldsymbol{G}$ que possua bom poder preditivo (risco quadrático pequeno). Dessa forma, se quer evitar modelos que sofram de subajuste (*underfitting*) ou sobreajuste (*overfitting*). Uma vez que a função de risco $R(g)$ é desconhecida, é necessário estimá-la para avaliar a função $g \in G$.

O primeiro passo para construir boas funções de predição é criar um critério para medir o desempenho de uma dada função de predição $g: R^d \rightarrow R$. Isto é feito através de uma função de risco. Assumindo que *Y* possui valores em um conjunto *C* (por exemplo, *C* pode ser o conjunto {câncer, normal}. A função de risco é dada pela equação

$$R(g) \coloneqq \boldsymbol{E}\left[\boldsymbol{I}\left(Y, g(\boldsymbol{X})\right)\right] = \boldsymbol{P}(Y, g(\boldsymbol{X})) \tag{2.44}$$

ou seja, o risco de g é a probabilidade de erro em uma nova observação $(\boldsymbol{X}, Y)$ que não foi usada para estimar g. Onde $\boldsymbol{I}$ é a matriz identidade. A esperança $\boldsymbol{E}$ é tomada em relação à observação $(\boldsymbol{X}, Y)$ e à amostra $\{(\boldsymbol{X_1}, Y_1), \ldots, (\boldsymbol{X_n}, Y_n)\}$, o risco é chamado de *risco esperado*.

Em outras palavras, o risco esperado é a esperança do risco condicional sob todas as possíveis amostras usadas para criar g. Assim, ele pode ser entendido como uma garantia sobre o processo de criação de g ao invés de uma garantia sobre a particular função g criada para um dado problema. Quanto menor o risco, melhor é a

função de predição g. Logo, uma função de perda mais adequada para o contexto de classificação é

$$L(g(\boldsymbol{x}), Y) = \boldsymbol{I}(Y, g(\boldsymbol{X})) \tag{2.45}$$

A Equação (2.45) é chamada função de perda 0-1.

No contexto de classificação, a função que minimiza $\boldsymbol{E}\left[\boldsymbol{I}\left(Y, g(\boldsymbol{X})\right)\right]$, ou seja, a melhor função de classificação g, segundo a função de risco da Equação 2.46, é dada por

$$g(\boldsymbol{x}) = arg \max_{d \in C} P(Y = d|\boldsymbol{x}) \tag{2.46}$$

isto é, deve-se classificar $\boldsymbol{x}$ como sendo daquela classe com maior probabilidade *a posteriori*. Este classificador é conhecido como classificador de *Bayes*, já comentado na Seção 2.2.1.2. Nota-se que o risco é desconhecido, pois a função $P(Y = d|\boldsymbol{x})$ é desconhecida.

Pode-se estimar o risco de um método de classificação utilizando-se *data splitting* e/ou validação cruzada. *Data splitting* é a separação (particionamento) dos dados em diferentes subconjuntos, tais como: treinamento e validação. Considerando *data splitting,* para o treinamento pode, por exemplo, ser utilizado 80% dos dados: $\{(X_1, Y_1), \dots, (X_s, Y_s)\}$, já para a validação, 20% dos dados restantes: $\{(X_{s+1}, Y_{s+1}), \dots, (X_n, Y_n)\}$. Utiliza-se o conjunto de treinamento para estimar g e o conjunto de validação apenas para estimar $R(g)$ via

$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} I\left(Y_i \neq g(X_i)\right) \coloneqq \hat{R}(g), \tag{2.47}$$

isto é, avalia-se a proporção de erros no conjunto de validação.

Assim, uma forma de selecionar um modelo g dentro de um conjunto de modelos $\boldsymbol{G}$ consiste em utilizar validação cruzada para estimar $R(g)$ para cada $g \in \boldsymbol{G}$ e, então, escolher g com o menor risco estimado.

Seja $\boldsymbol{G} = \{g_1, \dots, g_n\}$ um conjunto de classificadores estimados com base em um conjunto de treinamento, e seja $\widehat{\mathrm{R}}(\mathrm{g})$ o erro estimado de g com base no conjunto de validação da Equação (2.47). Seja g^* o modelo que minimiza o risco real $R(g)$

dentre $g \in \boldsymbol{G}$ e seja $\hat{\mathrm{g}}$ o modelo que minimiza o risco estimado $\hat{\mathrm{R}}(\mathrm{g})$ dentre $g \in \boldsymbol{G}$. Então, com probabilidade de no máximo ϵ ($\epsilon > 0$),

$$|R(\hat{g}) - R(g^*)| > 2\sqrt{\frac{1}{2(n-s)}\log\frac{2N}{\epsilon}} \tag{2.48}$$

A Equação (2.48) mostra que, quanto maior o número de classificadores em $\boldsymbol{G}$, menor é a probabilidade do melhor modelo ser recuperado. Esse é um dos fatores que fazem com que idealmente apenas um conjunto pequeno de modelos seja comparado no conjunto de teste e o ajuste (*tuning)* de cada modelo é feito utilizando validação cruzada dentro do treinamento ou particionando o treinamento em dois conjuntos: treinamento e validação. Assim, deixa-se o teste para comparar apenas o melhor modelo de cada algoritmo.

2.2.3 Otimização por hiperparâmetros

Os parâmetros são essenciais para algoritmos de aprendizado de máquina, sendo a parte do modelo que é aprendida com os dados de treinamento. Um modelo treinado é uma função matemática particular, pertencente a um certo tipo de algoritmo de aprendizado de máquina, que foi determinado por uma tupla particular de parâmetros. Os parâmetros que permitem a customização da função são os parâmetros do modelo, e são exatamente o que a máquina vai aprender com os dados.

No decorrer do treinamento, os parâmetros do modelo são ajustados automaticamente. Porém, no processo de construção de um modelo treinado, mais parâmetros são necessários para definir como o algoritmo fará isso. Em aprendizado de máquina, utiliza-se hiperparâmetros para denotar esse tipo específico de parâmetros. Os hiperparâmetros não podem ser aprendidos usando o algoritmo que precisa deles, mas devem ser ajustados antes da etapa de treinamento, manual ou automaticamente. Os hiperparâmetros são utilizados para configurar diversos aspectos do algoritmo de aprendizagem e podem ter efeitos variados no modelo resultante e na sua performance (CLAESEN; MOOR, 2015).

Em resumo, os parâmetros do modelo são estimados a partir dos dados

automaticamente. Já os hiperparâmetros do modelo são ajustados antes do treinamento e são usados para ajudar a estimar os parâmetros do modelo.

Encontrar os parâmetros mais eficazes para o processo de aprendizagem do modelo é geralmente referido como otimização de hiperparâmetros. Existem, principalmente, três métodos de otimização de hiperparâmetros:

- **Otimização Bayesiana** — A otimização bayesiana é uma estratégia poderosa para encontrar extremos de funções objetivas que são difíceis de avaliar. Ela funciona construindo uma distribuição posterior de funções (processo gaussiano) que melhor descreve a função que se deseja otimizar. À medida que o número de observações aumenta, a distribuição posterior melhora — Teorema do Limite Central — e o algoritmo torna-se mais certo de quais regiões no espaço de parâmetros valem a pena explorar e quais não.
- **Pesquisa em grade** — A pesquisa em grade (*grid search*) é um processo que pesquisa e testa exaustivamente todas as combinações possíveis do espaço de hiperparâmetros do algoritmo.
- **Busca Aleatória** – A busca aleatória (*random search*) é um método que seleciona aleatoriamente um conjunto de valores para cada hiperparâmetro do modelo em cada iteração.

Para cada algoritmo de aprendizado de máquina, há hiperparâmetros específicos que são testados para avaliar o modelo.

Os parâmetros principais para a **SVM** são:

- **Restrição** (*BoxConstraint*) — Ajusta os limites dos vetores de suporte. testa entre valores positivos, por padrão com escala de log no intervalo [$1e^{-3}$,$1e^{3}$].
- **Função de Kernel** (*KernelFunction*) — utilizada para calcular a matriz de Gram. Testa entre os tipos: gaussiano, linear e polinomial.
- **Escala do kernel** (*KernelScale*) —O *software* divide todos os elementos da matriz preditora X pelo valor de *KernelScale*. Em seguida, o *software* aplica a norma de *kernel* apropriada para calcular a matriz Gram. Testa entre valores positivos, por padrão com escala de log no intervalo [$1e^{-3}$,$1e^{3}$].

- **Grau ou ordem do polinômio** (*PolynomialOrder*) — testa entre inteiros no intervalo [2, 4].
- **Padronizar** (*Standardize*) - testa entre verdadeiro (‘*true*’) e falso (‘*false*’).

Os parâmetros principais para o ***Naive Bayes*** são:

- **Distribuição dos dados** (*DistributionNames*) — testa entre ‘normal’ (distribuição gaussiana) e ‘*kernel*’ (Estimativa de densidade de suavização de kernel).
- **Núcleo** (*Kernel*) — testa entre ‘*normal’* (gaussiana), ‘*box’* (uniforme), ‘*epanechnikov’* e ‘*triangle’* (triangular).
- **Largura** (*Width*) - largura da janela de suavização do kernel. Testa entre valores reais, por padrão com escala logarítmica no intervalo [$1e^{-3}$,$1e^{3}$].
- **Padronizar** (*Standardize*) – testa entre verdadeiro (‘*true*’) e falso (‘*false*’).

Os parâmetros principais para o **KNN** são:

- **Distância** (*Distance*) – testa entre *'cityblock’* (distância de Manhattan), *'chebychev'* (Distância de Chebyshev), *'correlation'* (Um menos a correlação linear amostral entre observações, *'cosine'* (Um menos o cosseno do ângulo incluído entre as observações, *'euclidean'* (distância euclidiana), *'hamming'* (Distância de Hamming, porcentagem de coordenadas diferentes), *'jaccard'* (Um menos o coeficiente de *Jaccard*, a porcentagem de coordenadas diferentes de zero que diferem), *'mahalanobis'* (Distância de Mahalanobis, calculada usando uma matriz de covariância definida positiva), *'minkowski'* (Distância de Minkowski), *'seuclidean'* (Distância euclidiana padronizada) e *'spearman'* (Um menos a correlação de classificação de *Spearman* da amostra entre as observações).
- **Peso da Distância** (*DistanceWeight*) — Função de ponderação da distância. Testa entre ‘*equal’* (sem ponderação), ‘*inverse’* (o peso = 1/distância) e ‘squaredinverse’ (o peso = 1/distância^2).
- **Expoente** (*Expoent*) — Expoente da distância de Minkowski, testa entre valores reais positivos, por padrão no intervalo [0.5,3].
- **Número de vizinhos** (*NumNeighbors*) — testa entre valores inteiros

positivos, por padrão com escala logarítmica no intervalo [1, max(2,*round*(Número de observação/2))].

- **Padronizar** (*Standardize*) – testa entre verdadeiro (‘*true*’) e falso (‘*false*’).

Os parâmetros principais para a **Árvore de decisão** são:

- **Número máximo de divisões** (*MaxNumSplits*) — corresponde ao número de ramificações. Testa entre números inteiros, por padrão com escala logarítmica no intervalo [1, *max*(2, *NumObservations*-1)].
- **Número mínimo de nós folha** (*MinLeafSize*) — testa entre números inteiros, por padrão com escala logarítmica no intervalo [1, *max*(2, *floor*(*NumObservations*/2))].
- **Número mínimo de nós de ramificação** (*MinParentSize*) — consiste em um valor inteiro positivo. Cada nó de ramificação na árvore possui um número mínimo de nós de ramificação definido por este parâmetro. Se ‘*MinParentSize’* e ‘*MinLeafSize’* são ambos fornecidos, o algoritmo usará a configuração que fornece mais nós folha: *MinParentSize* = *max(MinParentSize, 2*MinLeafSize*).
- **Critério de divisão** (*SplitCriterion*) — Para duas classes, testa entre *'gdi'* e *'deviance'* (desvio, também conhecido como entropia cruzada). Para três ou mais classes, o algoritmo testa a opção *'twoing'* (regra de dois).
- **Número de variáveis da amostra** (*NumVariablesToSample*) — número de preditores a serem selecionados aleatoriamente para cada divisão.

Os parâmetros principais para a **RNA** são:

- **Função de ativação** (*activation*) - Testa os valores *'relu'* (Linear retificada), *'tanh'* (Tangente hiperbólica), *'sigmoid'* (Sigmoide), *'none'* (Identidade).
- ***Lambda*** — fitcnet otimiza Lambda em valores contínuos no intervalo [1e-5, 1e5]/*NumObservations*, onde o valor é escolhido uniformemente no intervalo transformado em log.
- **Inicialização do bias da camada** (*LayerBiasesInitializer*) — testa entre *'zeros'* (cada camada totalmente conectada terá um *bias* inicial de 0), *'ones'* (cada camada totalmente conectada terá um *bias* inicial de 1).
- **Inicialização dos pesos da camada** (*LayerWeightsInitializer*) — testa

entre *'glorot'* (Inicialize os pesos com a inicialização *Glorot*), *'he'* (Inicialize os pesos com o inicializador *He*).

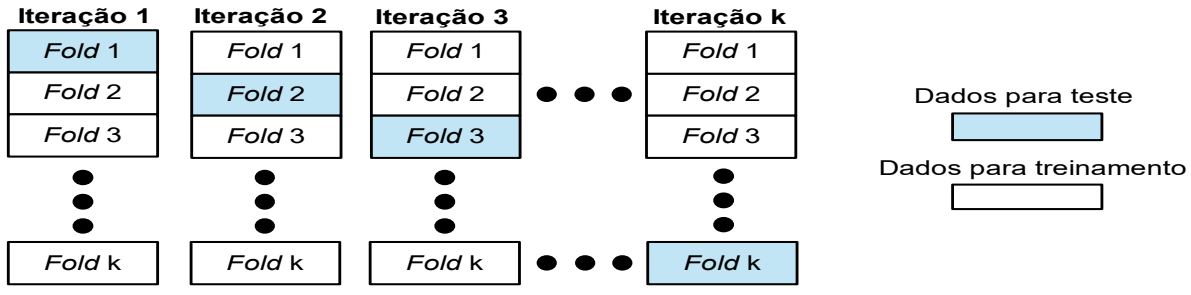
- **Tamanho das camadas** (*LayerSizes*) – quantidade de neurônios em cada camada intermediária. O algoritmo otimiza cada camada totalmente conectada separadamente entre 1 e 300 neurônios por camada, amostrados em uma escala logarítmica.

2.2.4 Medidas de desempenho

Na aplicação de algoritmos de ML em problemas reais, em geral, o conhecimento que se tem do domínio sendo investigado é provido unicamente pelo conjunto de exemplos, a partir do qual a indução de um modelo preditivo é então realizada. Os algoritmos de ML apresentados, anteriormente, na Seção 2.2.1 podem ser utilizados na indução de modelos de classificação a partir de um conjunto de exemplos rotulados. Não é possível estabelecer *a priori* que um algoritmo de ML em particular se sairá melhor na resolução de qualquer tipo de problema. Mesmo com o uso de certas heurísticas, diversos algoritmos de ML podem ser considerados candidatos à solução de um dado problema.

No método de validação cruzada *k-fold cross-validation*, o conjunto de exemplos é dividido em k subconjuntos de tamanho aproximadamente igual. Os objetos de ***k-1*** partições são utilizados no treinamento de um algoritmo de predição, o qual é então testado na partição restante. Esse processo é repetido ***k*** vezes, utilizando em cada ciclo uma partição diferente para teste. O desempenho final do preditor é dado, normalmente, pela média dos desempenhos observados sobre cada subconjunto de teste. A Figura 17 ilustra como é realizada a validação cruzada.

Figura 17 – Validação cruzada *k-fold*



Fonte: Elaborada pelo autor

Em processamento de sinais biomédicos e reconhecimento de padrões, a metodologia de desempenho usual é medida calculando algumas medidas estatísticas sobre o resultado dos testes (BUSHBERG *et. al.,* 2001). Os resultados da classificação dos testes podem ser divididos em: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN). Sendo VP e VN o número de amostras que são corretamente classificadas, respectivamente, como positiva ou negativa pelo classificador, FP e FN representam o número de amostras correspondentes aos casos que são erroneamente classificados como positivo ou negativo, respectivamente. Tais números são utilizados para gerar medidas capazes de quantificar o desempenho da metodologia, para avaliar o quão este é eficiente e se os objetivos foram alcançados. Na prática, é comum avaliar o desempenho de um classificador com base em uma matriz de confusão como a apresentada na Tabela 1.

Tabela 1 – Matriz de confusão

Valor verdadeiro	**Valor predito**	
	Y = 0	Y = 1
Y = 0	VN	FN
Y = 1	FP	VP

Com base na matriz de confusão, pode-se definir várias medidas de desempenho, tais como:

A Acurácia, que é a taxa de acerto do classificador durante a fase de teste, e é definida por:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.49}$$

A Sensibilidade é a proporção de verdadeiros positivos que são corretamente classificados, ou seja, dos pacientes doentes, quantos foram corretamente identificados como doentes, e é definida por:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2.50}$$

A Especificidade é a proporção de verdadeiros negativos que são

corretamente classificados, ou seja, dos pacientes não-doentes, quantos foram corretamente identificados como não-doentes, e é definida por:

$$Especificidade = \frac{VN}{VN + FP} \tag{2.51}$$

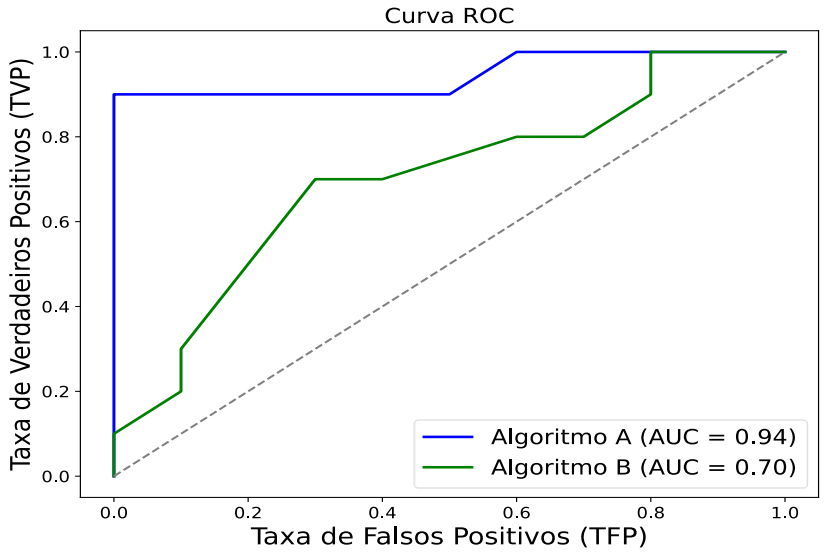
A Área sob a Curva ROC - *Area Under the Curve* (AUC) é uma forma de representar graficamente a relação entre a Taxa de Falsos Positivos (TFP) e a Taxa de Verdadeiros Positivos (TVP) ou sensibilidade, definidas pelas equações a seguir.

$$TFP = \frac{FP}{FP + VN} \tag{2.52}$$

$$TVP = \frac{VP}{VP + FN} \tag{2.53}$$

O gráfico ROC é bidimensional. Os valores da TVP são representados no eixo y e os valores da TFP no eixo x, no plano cartesiano. O desempenho do classificador é então plotado nessa curva. A medida da AUC produz valores entre 0 e 1. Valores mais próximos de 1 são considerados melhores. A Figura 18 ilustra a curva ROC gerada por dois algoritmos de ML. O algoritmo A tem um desempenho melhor, já que a área embaixo da sua curva é maior que a área gerada pelo algoritmo B.

Figura 18 – Exemplo de curva ROC para dois algoritmos



Fonte: Elaborada pelo autor

A análise pela curva ROC possibilita a avaliação do desempenho de modelos de classificação de maneira independente de fatores como limiar de classificação, os custos associados às classificações incorretas e à distribuição das classes (PRATI, 2006). De fato, o uso de diferentes limiares de classificação representa uma maior ou menor ênfase à classe positiva, permitindo lidar com questões de desbalanceamento das classes e de diferentes custos de classificação (FACELI et al., 2021).

A curva ROC é especialmente útil porque permite visualizar o *trade-off* entre sensibilidade e especificidade em diferentes pontos de corte. Um modelo ideal teria uma curva que passa pelo canto superior esquerdo do gráfico, indicando alta sensibilidade e baixa taxa de falsos positivos. Essa análise é amplamente utilizada em diversas áreas, incluindo a saúde, para avaliar a acurácia de testes diagnósticos e em aprendizado de máquina para comparar a eficácia de diferentes modelos de classificação.

3 TRABALHOS RELACIONADOS

Neste Capítulo serão apresentados alguns trabalhos que utilizaram técnicas de aprendizado de máquina para o diagnóstico de câncer de próstata. Além de pesquisas originais utilizando métodos científicos tradicionais, estudos nesta área têm discutido prevenção, tratamento e qualidade de vida de quem convive com câncer de próstata, por meio da utilização do aprendizado de máquina - *Machine Learning* (ML).

Foi realizada uma busca abrangente em bases de dados relevantes, como *PubMed*, *Scopus*, *Web of Science*, *IEEE Xplore*, entre outras. Títulos e resumos foram triados para identificar estudos potencialmente relevantes e correlatos a esta Tese, focando principalmente nas palavras-chave: câncer de próstata, aprendizado de máquina, triagem e diagnóstico. Posteriormente, os textos completos dos estudos selecionados foram avaliados para confirmar sua relevância. As informações sobre os autores, ano de publicação, métodos utilizados, resultados e conclusões, bem como quaisquer dados pertinentes à pesquisa, foram extraídas. Finalmente, a busca por periódicos foi revisada e atualizada periodicamente para garantir que a revisão se mantivesse atualizada com novos estudos e informações.

Os avanços recentes em diferentes campos da Inteligência Artificial (IA), especialmente na ML, são a grande razão pela qual a IA se prepara para assumir um papel central na vida das pessoas. A utilização de IA estar apenas iniciando a resolução de problemas importantes em áreas como comércio eletrônico, indústria aérea, guerra, diagnósticos médicos e quase todos os outros aspectos da vida humana. A IA fez avanços notáveis na última década, em grande parte devido ao financiamento sem precedentes.

Nos últimos anos, a IA foi impulsionada por aumentos exponenciais no poder da computação e pela disponibilidade de grandes quantidades de dados. Com o grande volume de dados que está sendo gerado e a crescente proliferação de dispositivos médicos e sistemas de registro digital, a área da medicina e os cuidados de saúde têm sido alvo constante de aplicação de sistemas que utilizam inteligência artificial. Como será visto nos trabalhos a seguir, já há algum tempo o câncer de próstata vem sendo objeto de estudo, seja por métodos automáticos de detecção, seja por análise de imagens ou pela análise de variáveis clínicas.

Wang et al. (2017) compararam a eficácia de métodos de aprendizado

profundo (*deep learning*) e não profundo (*non-deep learning*) na classificação automatizada de imagens por ressonância magnética para a detecção de câncer de próstata. Eles tinham disponíveis 2.602 imagens morfológicas (imagem axial 2D ponderada em T2) da próstata de 172 pacientes. Para cada imagem, um conjunto de transformação de características invariante à escala - *Scale-Invariant Feature Transform* (SIFT), foi extraído e codificado usando um dicionário pré-treinado. Fizeram dois experimentos, no primeiro, baseado em algoritmos que não são de *deep learning*, usaram o modelo *Bag-of-Word* (BoW) para agregar as características SIFT codificadas em uma representação vetorial para cada imagem. A classificação das imagens foi feita com um classificador SVM linear, e obtiveram uma AUC de 0,70. Já no segundo aplicaram redes neurais convolucionais profundas - *Deep Convolutional Neural Networks* (DCNN), e obtiveram uma AUC de 0,84.

Liu et al. (2019) desenvolveram dois novos modelos preditivos para determinar a necessidade de biópsia prostática em pacientes com níveis de PSA na zona cinzenta diagnóstica (4–10 ng/ml). Os dois modelos foram baseados em variáveis clínicas e demográficas, incluindo idade, PSA livre/total, volume da próstata e PSAD. Para o primeiro modelo foram utilizados todos os casos de câncer de próstata e para o segundo somente os casos de câncer de próstata clinicamente significativos - *Clinically Significant Prostate Cancer* (CSPCa), que considera a pontuação de *Gleason* igual ou superior a 7. Realizaram uma análise retrospectiva em 197 pacientes submetidos à biópsia de próstata com PSA entre 4 e 10 ng/ml. Destes, 47 pacientes tiveram câncer confirmado, enquanto os 150 pacientes restantes foram diagnosticados sem câncer após exame da patologia da biópsia. Testaram o método com regressão logística. Obtiveram como melhor desempenho para o primeiro modelo uma sensibilidade de 75,4%, especificidade de 75,8% e AUC de 0,775 e para o segundo modelo uma sensibilidade de 76,7%, especificidade de 80,1% e AUC de 0,819.

Liu et al. (2020) desenvolveram um modelo multivariáveis para prever o câncer de próstata entre pacientes na zona cinzenta (4 a 10 ng/ml) do PSA. Foram identificados retrospectivamente 235 pacientes com níveis de PSA na zona cinzenta antes da biópsia prostática, entre 2014 e 2018. Havia 179 pacientes com câncer e 56 sem câncer. Utilizaram como variáveis clínicas: idade, PSA total, PSA livre, volume da próstata, PSA livre/total e PSAD. Todos os pacientes foram submetidos à biópsia

de próstata guiada por TRUS. A análise univariada mostrou que o volume prostático, PSAD e IRMmp foram preditores significativos de PCa e CSPCa. Modelos multivariados foram desenvolvidos para melhorar a precisão diagnóstica. Testaram o método com regressão logística, e obtiveram como resultado utilizando IRMmp: AUC de 0,69 para PCa e 0,79 para CSPCa, superando o volume prostático e PSAD na detecção de CSPCa. Já para modelos multivariados obtiveram acurácia de 78,9%, sensibilidade de 79,1%, especificidade de 78,4% e AUC de 0,79 para PCa, e acurácia de 83,5%, sensibilidade de 84,3%, especificidade de 82,7% e AUC de 0,84 para CSPCa, mostrando desempenho significativamente melhor do que a IRMmp isolada (P = 0,003 para PCa e P = 0,036 para CSPCa).

Park et al. (2017) desenvolveram a Calculadora Coreana de Risco de Câncer de Próstata para Câncer de Próstata de Alto Grau - *Korean Prostate Cancer Risk Calculator for High-Grade Prostate Cancer* (KPCRC-HG) que prevê a probabilidade de câncer de próstata com pontuação de *Gleason* 7 ou superior na biópsia inicial da próstata em um coorte coreana. Além disso, a KPCRC-HG foi validado e comparado com calculadoras de risco ocidentais baseadas na Internet em uma coorte de validação. Utilizando um modelo de regressão logística, a KPCRC-HG foi desenvolvida com base nos dados de 602 homens coreanos anteriormente não rastreados que foram submetidos a biópsias iniciais da próstata. Usaram 2.313 casos para validação, a KPCRC-HG foi comparada com o Estudo Europeu Randomizado de Triagem para Calculadora de Risco de PCa para câncer de alto grau - *European Randomized Study of screening for PC Risk Calculator for High-Grade cancer* (ERSPCRC-HG) e a Calculadora de risco de ensaio de prevenção do câncer de próstata 2.0 para câncer de alto grau - *Prostate Cancer Prevention Trial Risk Calculator 2.0 for High-Grade cancer* (PCPTRC-HG). Os preditores independentes incluíram a idade do paciente, resultados de toque retal, nível de PSA total, resultados do TRUS, volume da próstata e volume da zona de transição - *Transition Zone Volume* (TZV) foram avaliados por análises de regressão logística para detectar PCa de alto grau (pontuação de *Gleason* maior ou igual a 7). A precisão preditiva foi avaliada usando AUC e gráficos de calibração. A AUC da KPCRC-HG foi de 0,84 foi superior à da PCPTRC-HG (0,79, p<0,001), mas não diferente do ERSPCRC-HG (0,83) na validação externa. Os gráficos de calibração também revelaram melhor desempenho da KPCRC-HG e ERSPCRC-HG em comparação com a PCPTRC-HG na validação

externa. Com um ponto de corte de 5% para a KPCRC-HG, 253 dos 2.313 homens (11%) não teriam sido biopsiados, e 14 dos 614 casos de PCa com escore de *Gleason* 7 ou superior (2%) não teriam sido diagnosticados.

Yoo et at. (2019) desenvolveram um *pipeline* automatizado utilizando DCNN para a detectar CSPCa com base em imagens por ressonância magnética ponderadas por difusão - *Diffusion-Weighted Imaging* (DWI). O conjunto de dados consistiu em imagens de 427 pacientes, dos quais 175 apresentavam câncer de próstata clinicamente significativo e 252 eram pacientes saudáveis. O *pipeline* desenvolvido utilizou DCNN para a classificação das imagens. O modelo foi treinado com um conjunto de dados e validado com um conjunto de teste separado, composto por 108 pacientes que não foram utilizados na fase de treinamento. A performance do pipeline foi medida utilizando a AUC tanto a nível de fatia quanto a nível de paciente. O melhor desempenho obtido foi, a nível de fatia, uma acurácia de 85,3% e uma AUC de 0,87. Já a nível de paciente, a melhor acurácia foi de 83,7% e a AUC foi de 0,84.

Chen et al. (2021) desenvolveram um método para a identificação de câncer de próstata utilizando espectroscopia fotoacústica combinada com técnicas de aprendizado de máquina. O estudo incluiu amostras de tecido prostático de pacientes diagnosticados com câncer de próstata e de indivíduos saudáveis. A técnica de espectroscopia fotoacústica foi utilizada para obter assinaturas espectrais das amostras de tecido. Esta técnica combina a alta sensibilidade da espectroscopia óptica com a alta resolução espacial da ultrassonografia. Utilizaram os algoritmos Análise Discriminante Linear - *Linear Discriminant Analysis* (LDA) e Análise Discriminante Quadrática - *Quadratic Discriminant Analysis* (QDA). A performance dos modelos foi avaliada utilizando métricas como sensibilidade, especificidade, acurácia e AUC. Utilizaram 101 amostras (90 para treinamento e 11 para teste). Obtiveram uma AUC de 0,851 e 0,862 para a LDA e para QDA, respectivamente.

Lu et al. (2023) desenvolveram uma rede de aprendizado auto-supervisionado com dupla cabeça de atenção (*dual-head attentional bootstrap)* para a triagem de câncer de próstata em imagens de ultrassom transretal. O estudo utilizou um conjunto de dados de imagens adquiridas pelo TRUS de pacientes com suspeita de câncer de próstata. A rede proposta é uma arquitetura de aprendizado profundo que utiliza um mecanismo de *bootstrap* com dupla cabeça de atenção para melhorar a precisão da triagem. O modelo é treinado de forma auto-supervisionada, o que significa que ele

não requer rótulos manuais extensivos para o treinamento. Duas cabeças de atenção são usadas para focar em diferentes características das imagens, melhorando a capacidade do modelo de identificar regiões suspeitas. *Bootstrap* utiliza um processo iterativo para refinar as previsões do modelo, aumentando a robustez e a precisão. O desempenho do modelo foi uma acurácia de 80,46%.

Zheng et al. (2019) realizaram um estudo de coorte retrospectivo no Laboratório do Departamento de Urologia do Hospital da União da Universidade Médica de Fujian (Fuzhou, China) de janeiro de 2012 a março de 2018. O estudo teve como objetivo investigar o papel do volume prostático baseado em imagens de ressonância magnética e das concentrações de PSA ajustadas por zona na predição de câncer de próstata e câncer de próstata de alto risco. Os dados foram coletados de 422 pacientes que foram submetidos a IRMmp antes da biópsia da próstata inicial de 13 núcleos guiada por TRUS. Análises multivariadas utilizadas, como: volume da próstata, PSAD, PSA, volume da zona periférica, TZV, zona de transição de densidade do PSA e ressonância magnética foram preditores independentes. A AUC do melhor modelo preditivo incluindo PSA + volume da próstata + PSAD + IRM + TZV ou PSA + volume da próstata + PSAD + IRM + volume da zona periférica foi de 0,906, sensibilidade de 85,3% e especificidade de 80,5%. A Tabela 2 mostra a síntese dos trabalhos mais recentes sobre câncer de próstata.

Tabela 2 – Síntese dos trabalhos mais recentes sobre câncer de próstata

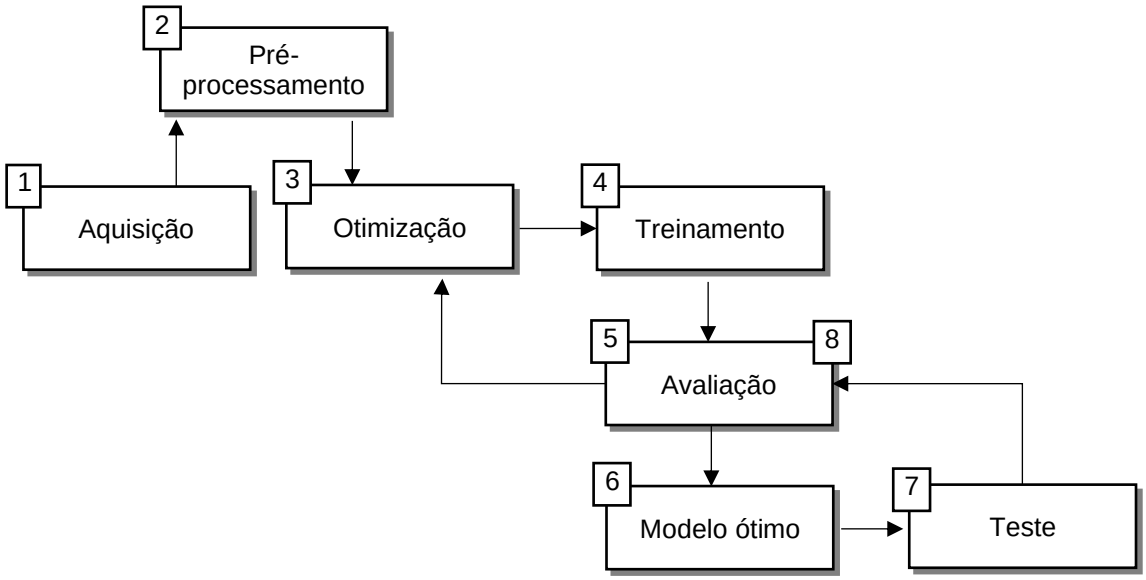
Trabalho	**Algoritmo**	**Método**	**Resultado**	
			AUC	**Acurácia (%)**
Wang et al. (2017)	SVM	IRM	0,700	-
Liu et al. (2019)	Regressão Logística	Multivariáveis	0,775	-
Liu et al. (2020)	Regressão Logística	Multivariáveis	0,790	-
Liu et al. (2019)	Regressão Logística	Multivariáveis com *Gleason* ≥ 7	0,819	
Liu et al. (2020)	Regressão Logística	Multivariáveis com *Gleason* ≥ 7	0,840	-
Wang et al. (2017)	DCNN	IRM	0,840	-
Park et al. (2017)	Regressão Logística	Multivariáveis	0.840	-
Yoo et al. (2019)	*Random Forest*	IRM	0,840	-
Chen et al. (2021)	LDA	Fotoacústica	0,851	76,30
Lu et al. (2023)	Bootstrap	IRM	-	80,46
Chen et al. (2021)	QDA	Fotoacústica	0,862	81,70
Yoo et al. (2019)	DCNN	IRM	0,870	-
Zheng et al. (2019)	Regressão Logística	Multivariáveis	0,906	82,70

Abreviações: AUC = *Area Under the Curve*; SVM = *Support Vector Machines*; IRM = Imagem por Ressonância Magnética; DCNN = *Deep Learning Convolutional Neural Network*.

4 METODOLOGIA PROPOSTA

Esta Tese propõe o desenvolvimento de um método de auxílio ao diagnóstico de câncer de próstata através da aplicação das técnicas de aprendizado de máquina e dados clínicos, para que sirva de triagem para a biópsia de câncer de próstata. Para isto foi desenvolvida uma metodologia composta de oito etapas, que serão descritas nas seções subsequentes. A primeira etapa é a aquisição dos dados, a segunda etapa é o pré-processamento, a terceira etapa é a otimização dos parâmetros, a quarta etapa é o treinamento de cada modelo, a quinta etapa é a avaliação dos modelos treinados, a sexta etapa é a geração do modelo ótimo para cada algoritmo de aprendizado de máquina, a sétima etapa compõe a fase de testes com novos dados não utilizados no treinamento, e a última etapa é a avaliação dos modelos após os testes. O diagrama em blocos, do método proposto, é mostrado na Figura 19.

Figura 19 - Diagrama esquemático do método proposto



Fonte: Elaborada pelo autor

4.1 Declaração de ética

O presente estudo foi realizado de acordo com a declaração de Helsinque e foi aprovado um projeto de pesquisa pela Comissão Científica (COMIC) e pelo Comitê de Ética em Pesquisa (CEP), conforme parecer CAAE: 45444621.6.0000.5086, e parecer técnico consubstanciado ID: 4.679.671, ambos pertencentes ao Hospital

Universitário da Universidade Federal do Maranhão (HU-UFMA), localizado na cidade de São Luís, capital do Estado do Maranhão, Brasil. Para proteger a privacidade desses dados clínicos, todos os princípios éticos dos direitos do paciente foram atendidos e os nomes dos participantes não foram utilizados. Todos os métodos foram realizados de acordo com diretrizes e regulamentos relevantes. O HU-UFMA autorizou o estudo e dispensou o termo de consentimento livre e esclarecido, pois os dados utilizados eram para um estudo retrospectivo sem afetar o atendimento ao paciente, e aprovou todos os experimentos.

4.2 Aquisição dos dados

Para a aquisição dos dados, foi realizada uma pesquisa observacional retrospectiva. Foram obtidos 274 prontuários médicos de pacientes atendidos no setor de urologia do HU-UFMA, sendo 137 com diagnóstico de câncer de próstata e 137 sem câncer de próstata, obtidos através do sistema do próprio hospital. Os prontuários abrangem o período de janeiro de 2017 a outubro de 2023.

Os critérios de inclusão na pesquisa foram: ter realizado biópsia da próstata, possuir prontuário completo (com todos os dados necessários para a pesquisa) e ter idade superior a 40 anos. Os dados adquiridos incluem informações sociodemográficas e variáveis clínicas, inclusive o resultado da biópsia da próstata, que é geralmente solicitada pelos médicos urologistas em casos de suspeita de câncer. É importante destacar que amostras com valores ausentes não foram utilizadas.

As variáveis clínicas utilizadas neste estudo para cada paciente foram: idade, raça, Hipertensão Arterial Sistêmica (HAS), Diabetes *Mellitus* (DM), tabagismo, etilismo, toque retal e PSA total. Essas variáveis estão entre os fatores de risco associados ao câncer de próstata (NATIONAL CANCER INSTITUTE, 2023) e são amplamente reconhecidas na literatura médica como relacionadas ao câncer de próstata. A variável toque retal refere-se ao peso estimado da próstata no momento da avaliação clínica realizada pelo médico urologista.

A coleta de dados foi realizada de forma rigorosa, garantindo a confidencialidade e a integridade das informações dos pacientes. A Tabela 3 apresenta uma amostra da base de dados contendo informações dos prontuários

médicos dos pacientes, destacando a diversidade e a abrangência dos dados coletados, que são essenciais para a validação do método proposto nesta Tese.

Tabela 3 – Parte da base de dados criada com os dados dos prontuários

IDADE	RAÇA	HAS	DM	TABAGISMO	ETILISMO	TOQUE	PSA TOTAL	GLEASON
53	BRANCA	NÃO	NÃO	NÃO	SIM	18	4,50	6
61	PARDA	NÃO	NÃO	NÃO	NÃO	40	7,50	7
51	PARDA	SIM	NÃO	NÃO	SIM	24,1	4,19	7
57	PARDA	NÃO	NÃO	SIM	NÃO	30	9,29	7
63	PARDA	NÃO	NÃO	SIM	EX	48,5	8,11	7
65	BRANCA	SIM	NÃO	EX	EX	34	3,00	0
87	PARDA	NÃO	NÃO	EX	EX	95	4,80	0
57	PARDA	NÃO	NÃO	NÃO	EX	66	5,78	0
68	PARDA	SIM	NÃO	EX	EX	30	1,32	0
65	BRANCA	SIM	NÃO	EX	NÃO	80	2,19	0

Abreviações: HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen*.

De acordo com a Tabela 3, podem-se extrair algumas informações. “NÃO” significa que o paciente não possui a doença (HAS ou DM) ou não foi exposto ao risco (Etilismo e Tabagismo), “EX” significa que o paciente foi exposto ao risco anteriormente. “SIM” significa que o paciente tem a doença ou está exposto ao risco. Quando o valor de *Gleason* é maior que zero (0), isso significa que o paciente tem câncer de próstata, já um *Gleason* igual a zero significa que a biópsia foi negativa.

4.3 Pré-processamento

A maioria das vezes não é possível utilizar algoritmos de ML diretamente sobre os dados coletados. Desta maneira, é necessária uma etapa de pré-processamento, que constituiu na padronização ou parametrização nos valores das características utilizadas. Esta padronização é necessária, pois os algoritmos de classificação precisam que os dados estejam em um formato utilizável. Portanto, o pré-processamento foi feito da seguinte forma:

- **Idade**: foi utilizada a idade, em anos, do paciente no dia da consulta, ou seja, um valor numérico inteiro.
- **Raça**: os números 1, 2, 3 ou 4 foram utilizados para codificar as raças (1

significa branca, 2 significa parda, 3 significa preta e 4 significa indígena).

- **HAS** e **DM**: os números 1 ou 2 foram utilizados (1 significa sim e 2 significa não).

- **Tabagismo** e **Etilismo**: os números 1, 2 ou 3 foram utilizados (1 significa sim, 2 significa não, 3 significa que o paciente foi exposto ao risco anteriormente). Para o etilismo, bastou o paciente assumir o consumo de álcool, o grau de alcoolismo não foi considerado para esta variável. Da mesma forma para o tabagismo, bastou o paciente informar que está exposto ao risco, sem considerar a quantidade de maços consumidos.
  - **Toque**: foi utilizado o peso estimado da próstata em gramas (g).
  - **PSA total**: foi utilizado o valor, numérico real em ng/ml, contido no exame de sangue.
  - **Rótulo**: foi utilizado para identificar cada amostra em uma classe (normal ou câncer). No aprendizado supervisionado, é necessária uma variável de saída, também chamada de variável alvo, para que no momento do treinamento do algoritmo de ML ele saiba a qual classe pertence àquela amostra específica, sendo assim, foi utilizado o valor 0 (zero) ou 1 (um), zero significa que o resultado da biópsia foi negativo, e 1 quando o resultado foi positivo (*Gleason* > 0).

A Tabela 4 mostra a parametrização e o tipo das variáveis do *dataset* utilizado.

Tabela 4 – Parametrização inicial de cada variável do *dataset*

Variável	Parametrização	Tipo
Idade	Anos	Inteiro
Raça	Branca (1), Parda (2), Preta (3), Indígena (4)	Inteiro
HAS	Sim (1), Não (2)	Inteiro
DM	Sim (1), Não (2)	Inteiro
Tabagismo	Sim (1), Não (2), Ex (3)	Inteiro
Etilismo	Sim (1), Não (2), Ex (3)	Inteiro
Toque	Gramas (g)	Real
PSA total	Ng/ml	Real
Rótulo	Normal (0), Câncer (1)	Inteiro

Abreviações: HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen*; Ng/ml = Nanogramas por mililitro.

4.4 Otimização

Encontrar os parâmetros mais eficazes para o processo de aprendizagem do modelo é geralmente referido como otimização de hiperparâmetros (B.J. Erickson et al.; M.M. Sartias et al.; R. Gandhi). Conforme explicado na Seção 2.2.3, existem, principalmente, três métodos mais utilizados de otimização de hiperparâmetros:

- **Otimização Bayesiana** — constrói uma distribuição posterior de funções (processo gaussiano) que melhor descreve a função que se deseja otimizar.
- **Pesquisa em grade** — é um processo que pesquisa e testa exaustivamente todas as combinações possíveis do espaço de hiperparâmetros do algoritmo.
- **Busca Aleatória** – é um método que seleciona aleatoriamente um conjunto de valores para cada hiperparâmetro do modelo em cada iteração.

Foi realizada uma extensa exploração de hiperparâmetros usando o parâmetro *'OptimizeHyperparameters'* da função *'fit'* do *Matlab*®, garantindo uma configuração otimizada para cada modelo de aprendizado de máquina. Várias combinações, já descritas na Seção 2.2.3, para ajuste de hiperparâmetros foram aplicadas para cada algoritmo de aprendizado de máquina. O uso de hiperparâmetros diferentes para classificação dará resultados diferentes, e a otimização bayesiana assume uma relação funcional entre os hiperparâmetros e a função de perda que o modelo final deve otimizar.

Para alcançar a melhor configuração, foram exploradas, sistematicamente, várias opções de valores dos parâmetros para cada algoritmo utilizado neste trabalho, que estão sumarizados a seguir:

- **SVM**: função do kernel (linear, polinomial ou gaussiana), parâmetro C, escala do kernel e padronização dos dados.
- ***Naive Bayes***: nomes de distribuição (normal ou *kernel*), largura, *kernel* (normal, caixa, *epanechnikov* e triângulo) e padronização dos dados.
- **KNN**: número de vizinhos, distância, peso da distância e padronização dos dados.
- **Árvore de decisão**: tamanho mínimo da folha, número máximo de divisão,

critério de divisão e número de variáveis a serem amostradas.

- **RNA**: número de camadas, função de ativação, *lambda*, inicializador de pesos de camada, inicializador de polarização de camada e padronização dos dados.

4.5 Treinamento

Treinar um modelo significa aprender bons valores para todos os pesos e tendências a partir de exemplos rotulados. Para a etapa de treinamento, 80% das amostras foram utilizadas. Estes dados foram divididos aleatoriamente, formando o conjunto de dados de treinamento. No aprendizado de máquina supervisionado, um algoritmo de ML constrói um modelo examinando muitos exemplos (amostras) e tentando encontrar um modelo que minimize as perdas ou erros. Cada modelo foi treinado utilizando algumas funções do *software* Matlab® R2023b.

Além disso, para evitar o *overfitting*, foi utilizada a técnica de validação cruzada (*10-fold cross-validation*), onde o conjunto de dados foi dividido em dez subconjuntos: nove subconjuntos para treinamento e um subconjunto para validação. Isso foi repetido dez vezes até que todos os subconjuntos fossem utilizados para treinar o classificador. O resultado final obtido no treinamento foi a média das dez iterações.

Alguns algoritmos de ML foram utilizados nesta etapa para verificar qual obteve o melhor desempenho, e assim ser utilizado para o método proposto. Os algoritmos usados para esta etapa foram: SVM, *Naive Bayes*, KNN, Árvores de Decisão e RNA. Cada algoritmo foi descrito, em detalhes, na Seção 2.2.1, e terão um breve resumo, as seguir.

SVM é um método de aprendizagem supervisionada, capaz de classificar a partir de ***n*** indivíduos observados pertencentes a vários subgrupos, a qual classe um indivíduo pertence (CORTES; VAPNIK, 1995; LIANG; LIU; NIU, 2021; HAZARIKA; GUPTA, 2021; ESSAM et al., 2022; HAZARIKA; GUPTA, 2022). A ideia da SVM é construir um hiperplano como superfície de decisão, de forma que a margem de separação entre as classes seja a máxima possível. O objetivo do treinamento via SVM é obter hiperplanos que dividam as amostras de forma que os limites de generalização sejam otimizados. Mesmo quando as duas classes não sejam

totalmente separáveis, a SVM pode encontrar um hiperplano utilizando conceitos pertencentes à teoria de otimização (DING; PENG, 2005).

O *Naïve Bayes* é um classificador baseado em probabilidade que funciona com base no princípio do teorema de Bayes. Baseia-se na probabilidade condicional e na suposição de que os atributos são independentes uns dos outros. Embora esta suposição, às vezes, não seja válida para aplicações práticas, o desempenho deste classificador ainda está no mesmo nível de classificadores mais complexos. Os classificadores *Naïve Bayes* são modelos simples e com excelente desempenho. O desempenho dos modelos pode ser ajustado de acordo com as preferências individuais com base na aplicação.

O algoritmo KNN usa “similaridade de características” para prever os valores de quaisquer novos pontos de dados. Isso significa que ao novo ponto é atribuído um valor baseado em quão próximo ele se assemelha aos pontos do conjunto de treinamento (CUI et al., 2020). KNN é um método supervisionado não paramétrico simples e poderoso, que pode ser usado para classificação e regressão. *K* amostras mais próximas da amostra de teste são escolhidas do conjunto de dados de treinamento para classificar uma amostra de teste. Para tarefas de classificação, o rótulo dominante entre os rótulos alvo das *K* amostras de treinamento escolhidas é escolhido como o rótulo previsto para a amostra de teste (KHALILI et al., 2023).

A Árvore de Decisão é um modelo comum de aprendizagem supervisionada e uma ferramenta de apoio à decisão para classificação. Este modelo classifica os dados aprendendo regras de decisão simples derivadas das características dos dados. As profundidades máximas da árvore e a divisão mínima da amostra são alguns dos parâmetros que precisam ser determinados no processo de calibração (XING et al., 2020).

As RNAs são modelos matemáticos não lineares que mimetizam o cérebro humano nas características de aprendizagem e tomada de decisão, estimulando as habilidades cognitivas humanas. As RNAs são usadas para mapear e prever resultados em relacionamentos complexos entre determinadas entradas e saídas procuradas, e podem ser usadas para encontrar padrões em conjuntos de dados. As RNAs podem ser complexas, com camadas ocultas, e podem ser treinadas para representar e prever percepções multicamadas processando dados com aprendizado profundo (HOSSAIN et al., 2021).

Após o treinamento é feita a avaliação do mesmo através das medidas de desempenho mostradas anteriormente na Seção 2.2.4, para verificar se o modelo tem um desempenho satisfatório nesta etapa.

4.6 Modelo Ótimo, Testes e Avaliação dos modelos preditivos

Após a conclusão do treinamento com os hiperparâmetros otimizados, obtém-se cada modelo, a partir de cada algoritmo de ML, com os melhores parâmetros selecionados através das funções do *software* MatLab®. Este modelo está sendo chamado de “modelo ótimo”. Desta forma, pode-se aplicar os dados de teste a cada modelo ótimo construído a partir de cada algoritmo de ML, para avaliar a capacidade preditiva de cada modelo e seu poder de generalização para novos dados, não utilizados na etapa de treinamento.

Para testar a precisão preditiva de cada modelo de ML, foram utilizados 20% dos dados, este subconjunto é chamado de teste, correspondendo a 55 pacientes, divididos nas classes câncer ou normal. Este conjunto de dados é utilizado como entrada para cada modelo ótimo. Esta etapa é realizada da seguinte maneira: cada modelo gerado na etapa de treinamento é carregado (executado) no *software MatLab*. O modelo contém todos os parâmetros otimizados utilizados durante sua construção. O teste é realizado comparando a classe alvo (rótulo) já conhecida, com a variável de predição estimada pelo teste.

Logo após, é criada a matriz de confusão do teste, e a partir dela são calculadas as medidas estatísticas de desempenho, já descritas na Seção 2.2.4. Para avaliar os modelos preditivos, a confiabilidade do método e do classificador foram utilizadas a acurácia, sensibilidade, especificidade e área sob a curva ROC, em conjunto com a técnica estatística de validação cruzada *10-fold cross-validation (*KOHAVI, 1995).

4.7 Análise de correlação das variáveis

Uma das maneiras mais simples de entender como as variáveis do conjunto de dados estão relacionados é por meio de uma matriz de correlação. O coeficiente de correlação deve estar no intervalo [-1,1]. Um valor zero significa que não há correlação entre as variáveis, um valor maior que zero indica uma correlação positiva,

ou seja, à medida que o valor de uma variável aumenta, o mesmo acontece com o valor da outra variável que está sendo correlacionada. Enquanto um valor menor que zero indica uma correlação negativa, ou seja, os valores das variáveis são inversamente proporcionais, quando um cresce o outro diminui, e vice-versa. Quanto mais próximo o coeficiente de correlação estiver de 1 ou de -1 há uma maior correlação forte entre as variáveis.

Normalmente a correlação é classificada da seguinte maneira:

- ±0,9 a ±1 indica correlação muito forte.
- ±0,7 a ±0,9 indica correlação forte.
- ±0,5 a ±0,7 indica correlação moderada.
- ±0,3 a ±0,5 indica correlação fraca.
- 0 a ±0,3 indica correlação desprezível ou nula.

Em estatística, aborda-se a questão da significância de um resultado usando-se o conceito de hipótese nula. A hipótese nula simplesmente assume que um dado resultado estatístico foi obtido apenas por acaso, devido a flutuações probabilísticas dos eventos sendo medidos, e não devido a um efeito real que cause o resultado. Sempre que se trabalha com uma hipótese para explicar um dado fenômeno, é considerada a possibilidade de pelo menos uma hipótese concorrente a ela. No caso da estatística, a hipótese concorrente é chamada de hipótese alternativa.

Se este valor de probabilidade é suficientemente baixo (onde o critério de definição de “suficientemente baixo” tem que ser previamente definido), pode-se rejeitar a hipótese nula. A ideia é a de que, se a probabilidade de o resultado ser obtido por mero acaso for muito baixa, deve-se considerar que a hipótese do acaso não é suficientemente forte para explicar o ocorrido e, portanto, que a hipótese alternativa tem mais chances de oferecer uma explicação melhor.

Normalmente, o limiar do valor de probabilidade abaixo do qual a hipótese nula é rejeitada é 5% (p = 0,05). Se a probabilidade do evento caso a hipótese nula esteja certa for menor que 5%, rejeita-se a hipótese nula; caso a probabilidade for maior que 5%, a hipótese nula é aceita.

A correlação de *Spearman*, também conhecida como coeficiente de correlação de postos de *Spearman*, é uma medida não paramétrica que avalia a dependência entre duas variáveis classificadas em postos. Esse coeficiente é indicado

pela letra grega ρ (rho) e é utilizado quando as relações entre as variáveis não são necessariamente lineares, mas monotônicas. Ela apresenta as seguintes características principais:

- **Natureza Monotônica**: A correlação de *Spearman* mede a força e a direção de uma relação monotônica entre duas variáveis. Isso significa que, à medida que uma variável aumenta, a outra tende a aumentar ou diminuir de forma consistente, mas não necessariamente de maneira linear.
- **Transformação em Postos**: As variáveis são transformadas em postos antes do cálculo. Por exemplo, se temos uma lista de valores, cada valor é substituído por sua posição (posto) na lista ordenada.
- **Coeficiente de Correlação**: O coeficiente de *Spearman* varia de -1 a 1.

A correlação de *Spearman* é especialmente útil quando os dados não atendem aos pressupostos de normalidade exigidos pela correlação de Pearson, oferecendo uma alternativa robusta para a análise de relações monotônicas.

5 RESULTADOS E DISCUSSÃO

Neste capítulo, os resultados obtidos serão apresentados. O método proposto foi implementado usando os *softwares MatLab® v. R2023b* e *IBM SPSS Statistics v.25.*

Foi realizada uma pesquisa observacional retrospectiva com base nos prontuários de 274 pacientes do setor de Urologia do HU-UFMA, dos quais 137 não apresentavam câncer de próstata e 137 apresentavam. As variáveis utilizadas foram: idade, raça, Diabetes *Mellitus* (DM), etilismo, tabagismo, Hipertensão Arterial Sistêmica (HAS), Toque retal e PSA total), que serviram como dados de entrada para os algoritmos de aprendizado de máquina. Também foi criada uma variável chamada *"Rótulo"*, que identifica se o resultado da biópsia foi positivo (*Gleason* > 0), indicando a presença de câncer de próstata, ou negativo (normal), indicando a ausência da doença.

A média de idade dos pacientes no momento do diagnóstico foi de 67,23 anos (desvio padrão: ± 7,846), variando entre 46 e 92 anos. A maioria dos pacientes era parda (70,1%). A média do peso da próstata, medida pelo exame de toque retal, foi de 57,81 g (desvio padrão: ± 28,78), variando de 13 a 200 g. A média do PSA Total foi de 8,25 ng/ml (desvio padrão: ± 11), variando de 0,24 a 79,41 ng/ml. Quanto à frequência, 58,4% dos pacientes apresentavam HAS positiva e 41,6% negativa; 22,6% tinham DM positiva e 77,4% negativa; 8% eram tabagistas, 46% não fumantes e 46% ex-tabagistas; 31,4% apresentavam etilismo positivo, 27,7% negativo e 40,9% eram ex-etilistas. A Tabela 5 exibe estas informações.

Tabela 5 – Média ou percentual, desvio padrão, mediana, intervalo e frequências das variáveis

Variável	Média ou Percentual	Desvio padrão	Mediana	Intervalo	Frequência (%)		
					Sim	Não	Ex
Idade	67,23	7,846	68	46-92	-	-	-
Raça	-	-	-	-	-	-	-
Branca	24,8%	-	-	-	-	-	-
Parda	70,1%	-	-	-	-	-	-
Preta	4,7%	-	-	-	-	-	-
Indígena	0,4%	-	-	-	-	-	-
HAS	-	-	-	-	58,4	41,6	-
DM	-	-	-	-	22,6	77,4	-
Tabagismo	-	-	-	-	8	46	46
Etilismo	-	-	-	-	31,4	27,7	40,9
Toque retal	57,81	28,78	50	13-200	-	-	-
PSA total	8,25	11	5,22	0,24–79,41	-	-	-
Rótulo	-	-	-	-	50%	50%	-

Abreviações: HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*.

Considerando a distribuição das variáveis por classe (câncer ou normal) obteve-se algumas informações relevantes. Em relação aos pacientes com diagnóstico negativo da biópsia (normal), as médias foram: idade de 69,06 anos (intervalo: 51-92; desvio padrão: ± 7,86), peso da próstata, medido pelo toque retal, de 66,34 g (intervalo: 19-200; desvio padrão: ± 30,92) e PSA total de 5,61 ng/ml (intervalo: 0,24-79,41; desvio padrão: ± 9,73). Para os pacientes com câncer, as médias foram: idade de 65,39 anos (intervalo: 46-79; erro desvio: ± 7,41), peso da próstata, medido pelo toque retal, de 49,29 g (intervalo: 13-126; desvio padrão: ± 23,66) e PSA de 10,89 ng/ml (intervalo: 2,47-78; desvio padrão: ± 11,59). Estas informações estão detalhadas na Tabela 6.

Tabela 6 – Distribuição por classe das médias, intervalo e desvio padrão das variáveis idade, toque e PSA total

Variável	Normal			Câncer		
	Média	**Intervalo**	**Desvio padrão**	**Média**	**Intervalo**	**Desvio padrão**
Idade	69,06	51-92	7,86	65,39	46-79	7,41
Toque	66,34	19-200	30,92	49,29	13-126	23,66
PSA total	5,61	0,24-79,41	9,73	10,89	2,47-78	11,59

Abreviação: PSA = *Prostate-Specific Antigen*.

Ao analisar os resultados da Tabela 6, percebe-se que a média de idade entre os pacientes com e sem câncer de próstata é muito similar, não havendo diferença significativa. Isso sugere que a idade, por si só, pode não ser um fator distintivo forte entre os dois grupos neste estudo específico. No entanto, pacientes sem câncer de próstata apresentam, em média, um peso estimado da próstata maior do que aqueles sem câncer. Esse aumento pode ser um fator que levou à realização de biópsias nesses pacientes. Em relação ao PSA total, a média é consideravelmente maior nos pacientes com câncer em comparação aos sem câncer (normais). Essa diferença significativa no nível de PSA reforça seu papel como um marcador importante na detecção do câncer de próstata.

Quanto a frequência das variáveis, os pacientes com diagnóstico normal apresentaram as seguintes distribuições: raça branca (28,5%), parda (67,2%), preta (3,6%) e indígena (0,7%); HAS positiva em 56,2% dos casos e negativa em 43,8%; DM positiva em 19% e negativa em 81%; tabagismo positivo em 7,3%, negativo em 46% e ex-tabagistas em 46,7%; etilismo positivo em 21,9%, negativo em 31,4% e ex-etilistas em 46,7%. Para os pacientes com diagnóstico positivo (câncer), as

distribuições foram: raça branca (21,2%), parda (73%) e preta (5,8%); HAS positiva em 60,6% dos casos e negativa em 39,4%; DM positiva em 26,3% e negativa em 73,7%; tabagismo positivo em 8,8%, negativo em 46% e ex-tabagistas em 45,3%; etilismo positivo em 40,9%, negativo em 24,1% e ex-etilistas em 35%. Estas informações podem ser visualizadas na Tabela 7.

Tabela 7 – Distribuição das frequências das variáveis raça, HAS, DM, tabagismo e etilismo, por classe (normal ou câncer)

Variável	Normal (%)	Câncer (%)
Raça (Branca)	28,5	21,2
Raça (Parda)	67,2	73
Raça (Preta)	3,6	5,8
Raça (Indígena)	0,7	0
HAS (Sim)	56,2	60,6
HAS (Não)	43,8	39,4
DM (Sim)	19	26,3
DM (Não)	81	73,7
Tabagismo (Sim)	7,3	8,8
Tabagismo (Não)	46	46
Tabagismo (Ex)	46,7	45,3
Etilismo (Sim)	21,9	40,9
Etilismo (Não)	31,4	24,1
Etilismo (Ex)	46,7	35

Abreviações: HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*.

Ao analisar os resultados apresentados na Tabela 7, observa-se que, em relação à raça, A maioria dos pacientes é parda, com uma proporção maior no grupo com câncer (73%) em comparação ao grupo com diagnóstico normal (67,2%). A proporção de pacientes brancos é maior no grupo com diagnóstico normal (28,5%) do que no grupo com câncer (21,2%). A proporção de pacientes negros é ligeiramente maior no grupo com câncer (5,8%) em comparação ao grupo com diagnóstico normal (3,6%). Apenas pacientes do grupo com diagnóstico normal são indígenas (0,7%), sem representação no grupo com câncer.

Quanto à HAS, é mais prevalente em pacientes que têm pressão alta, independentemente da presença de câncer. A proporção de pacientes com hipertensão é ligeiramente maior no grupo com câncer (60,6%) em comparação ao grupo com diagnóstico normal (56,2%). Consequentemente, a proporção de pacientes sem hipertensão é maior no grupo com diagnóstico normal (43,8%) do que no grupo com câncer (39,4%).

No caso da Diabetes *Mellitus* (DM), observa-se que a proporção de pacientes com DM é maior no grupo com câncer (26,3%) em relação ao grupo com diagnóstico normal (19%). e a maioria dos pacientes, tanto com diagnóstico normal (81%) quanto com câncer (73,7%), não apresenta diabetes.

Em relação ao tabagismo, a proporção de pacientes que fumam atualmente é ligeiramente maior no grupo com câncer (8,8%) em comparação ao grupo com diagnóstico normal (7,3%). A proporção de pacientes que nunca fumaram é a mesma em ambos os grupos (46%). A proporção de ex-fumantes é similar entre os dois grupos, com 46,7% no grupo com diagnóstico normal e 45,3% no grupo com câncer.

Por fim, no que diz respeito ao etilismo, a frequência de consumo de bebidas alcoólicas é significativamente maior entre os pacientes com câncer (40,9%) em comparação ao grupo com diagnóstico normal (21,9%). A proporção de pacientes que não consomem bebidas alcoólicas é maior no grupo com diagnóstico normal (31,4%) em relação ao grupo com câncer (24,1%). A proporção de ex-consumidores de bebidas alcoólicas é maior no grupo com diagnóstico normal (46,7%) em comparação ao grupo com câncer (35%).

A hipertensão arterial, a diabetes *mellitus* e o consumo de álcool são mais prevalentes entre os pacientes com câncer de próstata, sugerindo uma possível associação entre essas condições e a presença de câncer. A proporção de pacientes pardos é maior em ambos os grupos, destacando a importância de considerar fatores étnico-raciais nas análises. O tabagismo atual não mostra uma variação significativa entre os grupos, mas a história de tabagismo (ex-fumantes) é comparável. Essas informações podem ser úteis para direcionar futuras pesquisas e intervenções no diagnóstico e tratamento do câncer de próstata.

Foi realizada a análise de correlação de *Spearman* das variáveis, permitindo identificar e quantificar a força e a direção das associações monotônicas entre elas. Este método não paramétrico é particularmente útil para avaliar relações não lineares e é robusto contra outliers, proporcionando uma visão mais detalhada das interações entre as variáveis estudadas. A análise de *Spearman* é especialmente valiosa em contextos em que as suposições de normalidade não são atendidas, oferecendo uma alternativa eficaz para a correlação de Pearson. Além disso, ao considerar a ordem dos dados em vez dos valores absolutos, a correlação de *Spearman* pode revelar padrões de associação que poderiam ser obscurecidos por métodos paramétricos

tradicionais. A Tabela 8 apresenta a correlação de Spearman para o conjunto de dados.

Tabela 8 – Correlação de *Spearman* do dataset

Variável	Tipo	Idade	Raça	HAS	DM	Tabagismo	Etilismo	Toque	PSA total
IDADE	Coef.	1,000	-0,001	**-,157****	-0,098	**,155***	**,136***	**,293****	-0,014
	Sig.		0,993	0,009	0,105	0,010	0,024	0,000	0,821
RAÇA	Coef.	-0,001	1,000	**-,121***	0,057	-0,023	-0,013	0,087	**,140***
	Sig.	0,993		0,045	0,347	0,704	0,830	0,149	0,020
HAS	Coef.	**-,157****	**-,121***	1,000	**,226****	-0,062	0,016	-0,059	-0,065
	Sig.	0,009	0,045		0,000	0,310	0,797	0,329	0,283
DM	Coef.	-0,098	0,057	**,226****	1,000	-0,016	-0,042	-0,067	-0,022
	Sig.	0,105	0,347	0,000		0,786	0,491	0,266	0,716
TABAGISMO	Coef.	**,155***	-0,023	-0,062	-0,016	1,000	**,187****	0,073	-0,036
	Sig.	0,010	0,704	0,310	0,786		0,002	0,227	0,553
ETILISMO	Coef.	**,136***	-0,013	0,016	-0,042	**,187****	1,000	**,155***	-0,095
	Sig.	0,024	0,830	0,797	0,491	0,002		0,010	0,117
TOQUE	Coef.	**,293****	0,087	-0,059	-0,067	0,073	**,155***	1,000	0,058
	Sig.	0,000	0,149	0,329	0,266	0,227	0,010		0,338
PSA total	Coef.	-0,014	**,140***	-0,065	-0,022	-0,036	-0,095	0,058	1,000
	Sig.	0,821	0,020	0,283	0,716	0,553	0,117	0,338	

Abreviações: Coef. = Coeficiente; Sig. = Significância; HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen*.
**A correlação é significativa no nível 0,01 (2 extremidades).
* A correlação é significativa no nível 0,05 (2 extremidades).

De acordo com os resultados mostrados na Tabela 8, a correlação de *Spearman* mostrou que há uma correlação **negativa e desprezível** entre idade e HAS ($\rho = -0,157$; $p < 0,01$), indicando que conforme a idade aumenta, a presença de hipertensão arterial sistêmica tende a diminuir. Entre raça e HAS ($\rho = -0,121$; $p < 0,05$), indica que determinadas raças podem ter menor prevalência de hipertensão arterial sistêmica. Há uma correlação **positiva e desprezível** entre idade e tabagismo ($\rho = 0,155$; $p < 0,05$), indica que conforme a idade aumenta, a tendência a fumar também aumenta. Entre idade e etilismo ($\rho = 0,136$; $p < 0,05$), indica que conforme a idade aumenta, a tendência ao consumo de álcool também aumenta. Entre idade e toque ($\rho = 0,293$; $p < 0,01$), indica que conforme a idade aumenta, o peso estimado da próstata tende a aumentar. Entre raça e PSA total ($\rho = 0,140$; $p < 0,05$), indica que certas raças podem apresentar níveis mais altos de PSA total. Entre HAS e DM ($\rho = 0,226$; $p < 0,01$), indica que pacientes com hipertensão arterial sistêmica também tendem a ter

diabetes *mellitus*. Entre tabagismo e etilismo ($\rho = 0{,}187$; $p < 0{,}01$), indica que pacientes que fumam tendem também a consumir álcool. Entre toque e etilismo ($\rho = 0{,}155$; $p < 0{,}05$), indica que pacientes que consomem álcool tendem a ter um peso estimado da próstata maior. As demais correlações não são significativas.

A presença de HAS está associada à idade e à DM, sugerindo que esses fatores de saúde podem estar interligados. As correlações entre tabagismo, etilismo e o peso estimado da próstata (toque retal) apontam para possíveis associações comportamentais e de saúde que merecem atenção. As relações observadas entre raça e PSA total, bem como raça e HAS, podem indicar diferenças étnico-raciais importantes na prevalência e nos níveis dessas condições. Esses *insights* podem ajudar a orientar futuros estudos e intervenções clínicas focadas em entender melhor e tratar o câncer de próstata e suas variáveis associadas.

A Escolha da análise de correlação de *Spearman* se deu por alguns motivos:

- **Distribuição dos Dados**: Os dados do estudo não seguem uma distribuição normal, logo a correlação de *Spearman* é mais apropriada do que a de Pearson.
- **Relação Monotônica**: A relação entre as variáveis é monotônica, mas não necessariamente linear.
- **Outliers**: Os dados contêm outliers que podem influenciar significativamente os resultados, logo *Spearman* oferece uma medida mais robusta.
- **Flexibilidade**: A capacidade de lidar com diferentes tipos de dados (ordinais, intervalares, racionais) torna *Spearman* uma escolha versátil.

Em resumo, a correlação de Spearman foi escolhida devido à sua robustez, flexibilidade e adequação para dados não normais e relações monotônicas. Essas características fazem dela uma ferramenta valiosa para analisar a associação entre variáveis clínicas em estudos médicos.

Outra abordagem para avaliar as variáveis é calcular a área sob a curva ROC em relação às classes (normal ou câncer). Esse cálculo ajuda a entender o quão representativa cada variável é para o trabalho em questão. A Tabela 9 apresenta os resultados dessa análise, fornecendo uma visão clara da eficácia das variáveis na distinção entre as classes.

Tabela 9 – Área sob a curva ROC AUC de cada variável

Variável	AUC	IC 95% Assintótico	
		Limite inferior	Limite superior
IDADE	0,369	0,303	0,435
RAÇA	0,540	0,472	0,608
HAS	0,478	0,410	0,547
DM	0,464	0,395	0,532
TABAGISMO	0,489	0,421	0,558
ETILISMO	0,404	0,337	0,471
TOQUE	0,322	0,259	0,385
PSA total	0,795	0,739	0,850

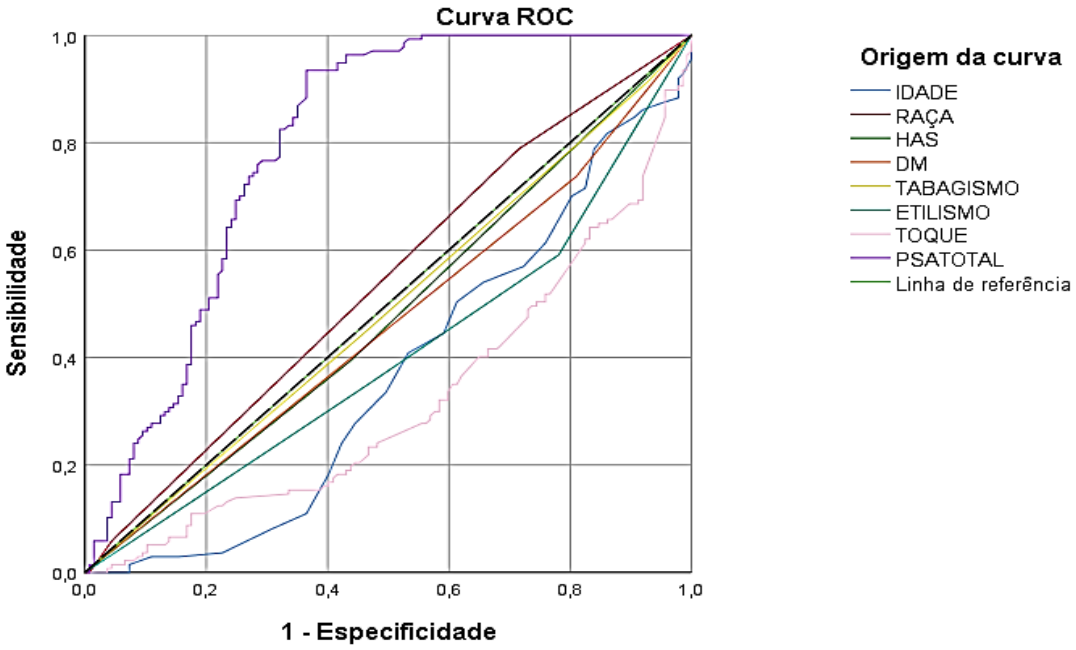
Abreviações: HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen;* AUC = *Area Under Curve*.

Ao analisar os dados da Tabela 9, algumas informações relevantes podem ser extraídas para o *dataset* utilizado. A idade não é um bom preditor de câncer de próstata, com uma AUC significativamente abaixo de 0,5. Isso indica que a idade, por si só, não discrimina bem entre pacientes com e sem câncer de próstata, e o intervalo de confiança também reforça a baixa capacidade preditiva da idade. A raça possui uma capacidade preditiva moderada, com a AUC ligeiramente acima de 0,5, indicando um leve poder discriminatório, e a margem de confiança relativamente ampla sugere variabilidade na predição baseada na raça. A HAS apresenta uma AUC abaixo de 0,5, indicando que não é um bom preditor de câncer de próstata, e o intervalo de confiança confirma a fraca capacidade discriminatória da HAS. Semelhante à HAS, a DM possui uma AUC abaixo de 0,5, demonstrando baixa eficácia preditiva, e o intervalo de confiança corrobora a baixa capacidade de discriminação da DM. O tabagismo também não se mostra como um bom preditor, com uma AUC perto de 0,5, e o intervalo de confiança sugere pouca variabilidade na capacidade preditiva do tabagismo. O etilismo possui uma AUC abaixo de 0,5, indicando fraca capacidade de predição, e o intervalo de confiança confirma a baixa eficácia preditiva. O peso da próstata, medido pelo exame de toque retal, indica uma discriminação fraca, e o intervalo de confiança confirma a baixa eficácia preditiva. No entanto o PSA total, indica boa discriminação. O intervalo de confiança é relativamente estreito, sugerindo alta precisão e consistência na capacidade preditiva.

Os intervalos de confiança fornecem uma medida da precisão das estimativas da AUC. Intervalos mais estreitos indicam maior precisão e menor variabilidade nos

dados. O PSA total se destaca como o melhor preditor, com uma AUC alta e um intervalo de confiança estreito. As demais variáveis apresentam AUCs baixas e intervalos de confiança que reforçam sua limitada capacidade preditiva. Os resultados indicam que a variável PSA total tem a melhor capacidade preditiva entre as variáveis analisadas, com alta precisão e consistência, mas utilizado isoladamente ainda não tem um excelente desempenho. As demais variáveis apresentam uma capacidade discriminatória limitada e podem não ser úteis como preditores isolados. A Figura 20 mostra a AUC gerada para cada variável.

Figura 20 – Área sob a curva ROC de cada variável em relação às classes



Fonte: Elaborada pelo autor

Os dados de cada variável foram parametrizados para garantir consistência e facilitar a análise subsequente. A Tabela 10, apresentada a seguir, exibe uma amostra parcial dos dados coletados, incluindo 10 dos 274 pacientes. Nesta tabela, é possível observar as parametrizações iniciais aplicadas a cada variável, o que permite uma visão preliminar da estrutura dos dados e das características dos pacientes. Essas parametrizações são essenciais para a padronização dos dados, assegurando que todas as variáveis estejam em um formato adequado para as análises estatísticas e de aprendizado de máquina que serão realizadas posteriormente.

Tabela 10 - Parte do dataset criado com 10 amostras das 274 coletadas

Idade	Raça	HAS	DM	Tabagismo	Etilismo	Toque	PSA total	Rótulo
66	2	1	2	3	1	39	5,61	1
68	2	2	1	3	2	70	5,96	1
73	1	2	2	1	1	40	6,14	1
63	3	1	2	2	2	44,7	10,71	1
70	2	1	2	2	1	60	5,91	1
65	2	2	2	2	2	80	63,68	0
74	2	1	1	2	2	80	4,20	0
74	3	2	2	2	2	80	9,79	0
61	2	1	2	3	3	65	10,59	0
68	2	2	2	2	2	60	11,49	0

Abreviações: HAS = Hipertensão Arterial Sistêmica; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen*.

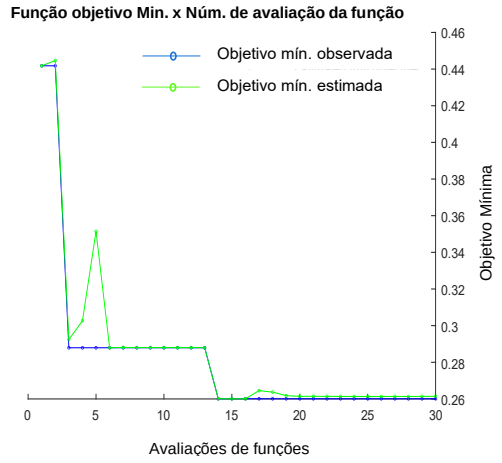
Após a parametrização inicial, os dados estavam prontos para serem utilizados no treinamento dos modelos de aprendizado de máquina. Para isso, foram aplicados diversos algoritmos, incluindo SVM, *Naive Bayes*, KNN, Árvores de Decisão - *Decision Trees* (DT) e RNA.

O conjunto de dados foi dividido aleatoriamente, com 80% destinado ao treinamento e 20% para teste. Além disso, foi utilizada a validação cruzada 10-fold para garantir a robustez dos resultados, resultando em 219 observações no conjunto de treinamento.

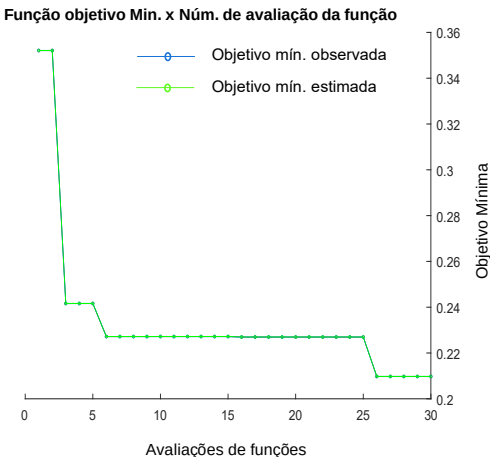
Durante a otimização dos parâmetros de cada modelo de aprendizado de máquina, foram testadas 30 diferentes parametrizações. O melhor resultado obtido em termos de desempenho foi considerado o modelo ótimo. A Figura 21 ilustra a função objetivo mínima observada e a função objetivo mínima estimada para cada iteração de cada modelo durante o estágio de treinamento, proporcionando uma visão clara do processo de otimização e da eficácia dos modelos testados.

Essa abordagem sistemática e rigorosa assegura que os modelos de aprendizado de máquina sejam treinados de maneira eficiente, maximizando a precisão e a generalização dos resultados para a aplicação prática no diagnóstico de câncer de próstata.

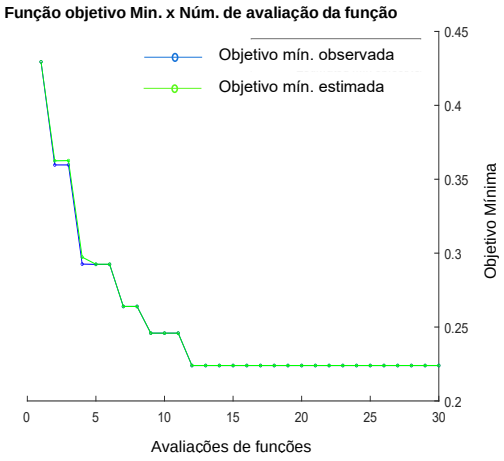
Figura 21 - Função objetivo mínima observada vs. Função objetivo mínima estimada por cada modelo de ML a) SVM b) *Naive Bayes* c) KNN d) DT e) RNA



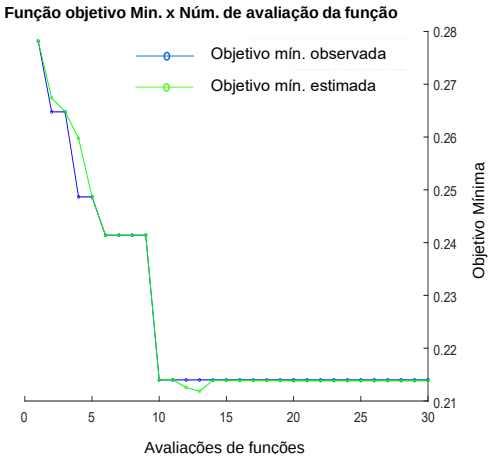
a) SVM



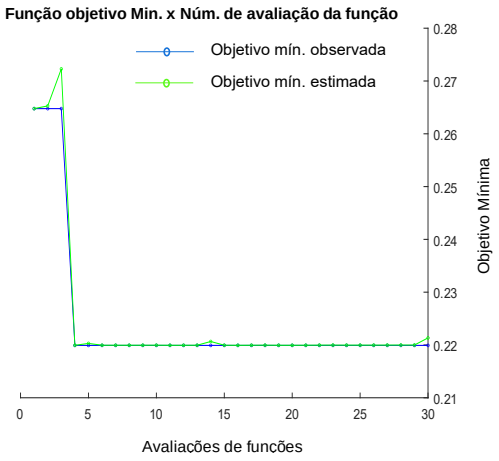
b) NAÏVE BAYES



c) KNN



d) DT



e) RNA

Fonte: Elaborada pelo autor

Diversas combinações de ajuste de hiperparâmetros foram aplicadas a cada modelo, conforme detalhado na Seção 4.4. Os melhores parâmetros identificados para cada modelo de aprendizado de máquina estão apresentados na Tabela 11.

Tabela 11 – Os melhores parâmetros obtidos para cada modelo de ML durante o treinamento

Modelo	**Parâmetros**
Naive Bayes	*DistributionNames* = ‘kernel’; *Width* = 0,1089; *Kernel* = ‘normal’; *standardize_data* = ‘true’.
RNA	*LayerSizes* = 11; *activation_function* = ‘relu’; *Lambda* = 1,269023252364904e-04; *LayerWeightsInitializer* = ‘glorot’; LayerBiasesInitializer = ‘zeros’; standardize_data = ‘false’.
KNN	*Number_of_neighbors* = 5; *distance* = ‘correlation’; *distance_weight* = ‘equal’; standardize data = ‘false’.
SVM	C = 0,0010029; *kernel_function* = ‘linear’; kernel_scale = 1, standardize_data = ‘false’.
DT	*Split_cliterion* = ‘gdi’; *MinParentSize* = 10; *MinLeafSize* = 1; *MaxSplits* = 144; *NumVariablestoSample* = 8.

Abreviações: RNA = Rede Neural Artificial; KNN = *K-Nearest Neighbor*; SVM = *Support Vector Machines*; DT = *Decision Tree*.

Ao analisar os hiperparâmetros obtidos para cada modelo de aprendizado de máquina, tem-se:

- ***Naive Bayes***
  - ***DistributionNames***: 'kernel' - Utiliza uma estimativa de densidade de kernel para modelar a distribuição dos dados.
  - ***Width***: 0,1089 - Largura do kernel, que controla a suavização da estimativa de densidade.
  - ***Kernel***: 'normal' - Tipo de kernel usado na estimativa de densidade.
  - ***Standardize_data***: 'true' - Os dados foram padronizados antes do treinamento.

- **Rede Neural Artificial (RNA)**
  - ***LayerSizes***: 11 - Número de neurônios na camada oculta.
  - ***Activation_function***: 'relu' - Função de ativação ReLU usada nas camadas ocultas.
  - ***Lambda***: 1,269023252364904e-04 - Parâmetro de regularização que controla a penalização dos pesos.
  - ***LayerWeightsInitializer***: 'glorot' - Inicializador de pesos Glorot (também

conhecido como Xavier).
  - ***LayerBiasesInitializer***: 'zeros' - Inicializador de bias com valores zero.
  - ***Standardize_data***: *'false'* - Os dados não foram padronizados antes do treinamento.

- ***K-Nearest Neighbors* (KNN)**
  - ***Number_of_neighbors***: 5 - Número de vizinhos considerados para a classificação.
  - ***Distance***: *'correlation'* - Métrica de distância baseada em correlação.
  - ***Distance_weight***: *'equal'* - Todos os vizinhos têm o mesmo peso na decisão.
  - ***Standardize_data***: *'false'* - Os dados não foram padronizados antes do treinamento.

- ***Support Vector Machine* (SVM)**
  - **C:** 0,0010029 - Parâmetro de regularização que controla o trade-off entre maximizar a margem e minimizar o erro de classificação.
  - ***Kernel_function***: 'linear' - Função kernel linear usada para transformar os dados.
  - ***Kernel_scale***: 1 - Escala do kernel.
  - ***Standardize_data***: *'false'* - Os dados não foram padronizados antes do treinamento.

- ***Decision Tree* (DT)**
  - ***Split_criterion***: 'gdi' - Critério de divisão baseado no índice de Gini.
  - ***MinParentSize***: 10 - Número mínimo de amostras necessárias para dividir um nó.
  - ***MinLeafSize***: 1 - Número mínimo de amostras que devem estar presentes em um nó folha.
  - ***MaxSplits***: 144 - Número máximo de divisões permitidas.
  - ***NumVariablestoSample***: 8 - Número de variáveis a serem consideradas para cada divisão.

Esses hiperparâmetros foram ajustados para otimizar o desempenho de cada modelo. A escolha e o ajuste correto dos hiperparâmetros são cruciais para garantir que os modelos de aprendizado de máquina tenham um bom desempenho e generalizem bem para novos dados.

Após o treinamento e otimização, foi obtido o modelo ótimo para cada algoritmo de ML. Em seguida, cada modelo foi testado utilizando o subconjunto de teste, composto por 20% dos dados restantes que não foram utilizados durante o treinamento. Posteriormente, foram extraídas as medidas de desempenho descritas na Seção 2.2.4, para avaliar o método proposto. As medidas de desempenho estão apresentadas na Tabela 12.

Tabela 12 – Análise de desempenho aplicado ao subconjunto de teste

Modelo	**VP**	**VN**	**FP**	**FN**	**Acurácia (%)**	**Sensibilidade (%)**	**Especificidade (%)**	**AUC (IC 95%)**
Naive Bayes	23	26	4	2	89,09	92	86,67	0,9187 (0,838 – 0,999)
RNA	20	24	6	5	80	80	80	0,9027 (0,811 – 0,995)
KNN	22	22	8	3	80	88	73,33	0,8947 (0,796 – 0,993)
SVM	20	22	8	5	76,36	80	73,33	0,8427 (0,722 – 0,963)
DT	20	25	5	5	81,82	80	83,33	0,8180 (0,695 – 0,941)

Abreviações: VP = Verdadeiro Positivo; VN = Verdadeiro Negativo; FP = Falso Positivo; FN = Falso Negativo; AUC = *Area Under the Curve*; IC = Intervalo de Confiança; RNA = Rede Neural Artificial; KNN = *K-Nearest Neighbor*; SVM = *Support Vector Machines*; DT = *Decision Trees*.

De acordo com os resultados, o modelo *Naïve Bayes* demonstrou o melhor desempenho geral, com uma acurácia de 89,09%, sensibilidade de 92%, especificidade de 86,67% e AUC de 0,9187 (IC 95%: 0,838-0,999). A AUC alta indica excelente discriminação, e o intervalo de confiança estreito sugere alta precisão. O modelo RNA apresentou bom desempenho com acurácia e sensibilidade equilibradas. A AUC alta indica boa discriminação, e o intervalo de confiança sugere consistência nos resultados. O modelo KNN apresentou boa sensibilidade, mas menor especificidade comparado aos outros modelos. A AUC alta indica boa discriminação, mas o intervalo de confiança é um pouco mais amplo, sugerindo variabilidade nos resultados. O modelo SVM apresentou desempenho moderado, com menor acurácia e AUC comparado aos outros modelos. O intervalo de confiança mais amplo indica

maior variabilidade nos resultados. O modelo DT apresentou bom desempenho com alta especificidade, mas menor AUC comparado aos melhores modelos. O intervalo de confiança mais amplo sugere variabilidade nos resultados.

O desempenho superior do modelo *Naive Bayes* pode ser explicado por algumas características intrínsecas do algoritmo e pela natureza dos dados:

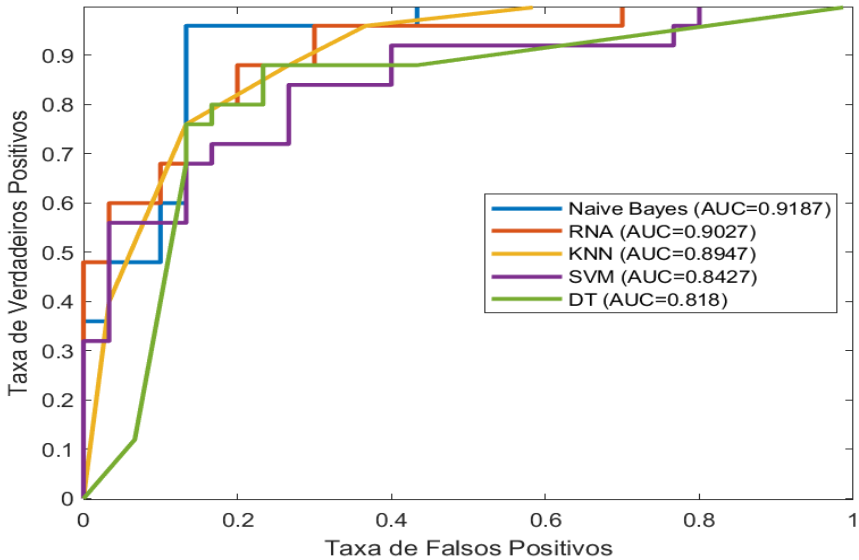
- ***Naive Bayes*** assume que todas as variáveis são independentes entre si, o que pode simplificar o modelo e reduzir o risco de *overfitting*, especialmente em conjuntos de dados menores ou com variáveis que realmente não são fortemente correlacionadas. A simplicidade do modelo também contribui para uma menor variância nos resultados, o que pode levar a um desempenho mais consistente. Se os dados se ajustam bem às distribuições assumidas pelo *Naive Bayes*, o modelo tende a performar muito bem. Ele pode lidar bem com dados balanceados e desbalanceados, desde que as probabilidades a priori sejam bem estimadas.
- **RNA**: Embora poderosas, as redes neurais articifiais podem sofrer de *overfitting* se não forem bem regularizadas e podem exigir mais dados para treinar efetivamente.
- **KNN**: pode ser sensível à escolha da métrica de distância e ao número de vizinhos, e pode não generalizar bem se os dados tiverem alta dimensionalidade.
- **SVM**: pode ser poderoso, mas a escolha do kernel e dos parâmetros de regularização é crucial. Pode não performar bem se os dados não forem linearmente separáveis ou se os parâmetros não forem bem ajustados.
- **DT**: As árvores de decisão podem ser propensas a *overfitting*, especialmente se não forem podadas adequadamente.

O *Naive Bayes* foi o melhor modelo neste caso específico, provavelmente devido à sua simplicidade, eficiência computacional, e adequação às características dos dados. Ele conseguiu capturar bem os padrões nos dados sem *overfitting*, resultando em um desempenho superior. Ele se destacou como o melhor modelo, com alta acurácia e excelente discriminação observada pela AUC. Os modelos RNA e KNN também apresentaram bom desempenho geral, enquanto os modelos SVM e DT tiveram desempenho moderado.

A AUC é uma métrica essencial para avaliar o desempenho dos classificadores, pois mede a capacidade do modelo em distinguir entre classes positivas e negativas. A Figura 22 ilustra a AUC gerada por cada modelo de aprendizado de máquina utilizado neste estudo. Nela, pode-se observar e comparar a eficácia de cada algoritmo (SVM, *Naive Bayes*, KNN, DT e RNA) em termos de sua precisão na classificação. A AUC varia de 0 a 1, onde valores mais próximos de 1 indicam um melhor desempenho do modelo. A análise detalhada da AUC permite identificar quais modelos apresentam maior capacidade discriminativa e são mais adequados para a aplicação prática no diagnóstico de câncer de próstata.

Além disso, a Figura 22 facilita a visualização das diferenças de desempenho entre os modelos, destacando aqueles que se sobressaem em termos de sensibilidade e especificidade. Essa comparação é fundamental para a seleção do modelo mais robusto e eficiente, garantindo que o método proposto nesta Tese seja baseado em uma abordagem de aprendizado de máquina rigorosamente avaliada e otimizada.

Figura 22 – AUC dos classificadores *Naive Bayes*, RNA, KNN, SVM e DT



Fonte: Elaborada pelo autor

Numerosos estudos já foram realizados sobre o diagnóstico de câncer de próstata utilizando aprendizado de máquina, conforme descrito no Capítulo 3. Esses estudos têm explorado diversas abordagens e algoritmos para melhorar a precisão e a eficiência do diagnóstico. No entanto, poucos estudos se concentraram

especificamente na análise de variáveis clínicas no contexto do rastreamento da biópsia do câncer de próstata. Ao incluir variáveis clínicas detalhadas, como idade, raça, hipertensão arterial sistêmica, diabetes *mellitus*, tabagismo, etilismo, toque retal e PSA total, o método proposto oferece uma perspectiva mais abrangente e personalizada para o diagnóstico. A Tabela 13 apresenta uma comparação entre o método proposto nesta Tese e outros trabalhos publicados sobre câncer de próstata. Esta comparação destaca as diferenças e semelhanças nas abordagens metodológicas, nos algoritmos utilizados e nos resultados obtidos.

Tabela 13 – Comparação com outros trabalhos

Trabalho	**Algoritmo**	**Método**	**Resultado**	
			AUC	**Acurácia (%)**
Wang et al. (2017)	SVM	IRM	0,700	-
Liu et al. (2019)	Regressão Logística	Multivariáveis	0,775	-
Liu et al. (2020)	Regressão Logística	Multivariáveis	0,790	-
Liu et al. (2019)	Regressão Logística	Multivariáveis com *Gleason* ≥ 7	0,819	
Liu et al. (2020)	Regressão Logística	Multivariáveis com *Gleason* ≥ 7	0,840	-
Wang et al. (2017)	DCNN	IRM	0,840	-
Park et al. (2017)	Regressão Logística	Multivariáveis	0.840	-
Yoo et al. (2019)	*Random Forest*	IRM	0,840	-
Chen et al. (2021)	LDA	Fotoacústica	0,851	76,30
Lu et al. (2023)	Bootstrap	IRM	-	80,46
Chen et al. (2021)	QDA	Fotoacústica	0,862	81,70
Yoo et al. (2019)	DCNN	IRM	0,870	-
Zheng et al. (2019)	Regressão Logística	Multivariáveis	0,906	82,70
Método proposto	***Naive Bayes***	**Multivariáveis**	**0,9187**	**89,09**

Abreviações: AUC = *Area Under the Curve*; SVM = *Support Vector Machines*; IRM = Imagem por Ressonância Magnética; DCNN = *Deep Learning Convolutional Neural Network*.

A comparação direta entre os resultados dos diferentes estudos é dificultada pela utilização de bases de dados distintas e métodos (Fotoacústica, IRM e Multivariáveis) diversos de aquisição dos dados. No entanto, ao realizar uma comparação indireta com base nos resultados obtidos, pode-se afirmar que, em síntese, o método proposto demonstrou-se competitivo na predição do câncer de próstata. Em particular, no que tange à AUC, que avalia a capacidade do modelo em distinguir entre pacientes com e sem a doença, o método proposto superou todos os estudos relacionados, alcançando um valor de 0,9187. Este valor indica uma excelente capacidade discriminatória, sugerindo que o modelo é altamente eficaz em diferenciar casos positivos e negativos de câncer de próstata. Assim, o método proposto nesta Tese revela-se competitivo na classificação do câncer de próstata, o

que é crucial para um rastreamento eficaz.

Com uma acurácia de 89,09%, o método proposto também superou os outros estudos em termos de precisão geral. Isso significa que o modelo tem uma alta taxa de acertos na classificação dos casos. Assim como alguns dos estudos comparados, o método proposto utilizou multivariáveis, o que pode ter contribuído para a robustez e precisão do modelo. A combinação de alta AUC e alta acurácia indica que o método proposto não só é preciso, mas também consistente em suas previsões.

Foi desenvolvido um protótipo de aplicativo para o método proposto, focado em facilitar a triagem e o diagnóstico de câncer de próstata utilizando aprendizado de máquina. Este aplicativo será uma ferramenta intuitiva, inovadora e de fácil acesso, destinada a facilitar a aplicação prática do método proposto. Ele permitirá aos médicos inserirem dados dos pacientes, como idade, raça, hipertensão arterial sistêmica, diabetes *mellitus*, hábitos de vida (tabagismo e etilismo) e resultados dos exames (toque retal e PSA total). Utilizando o melhor modelo de aprendizado de máquina obtido, o aplicativo analisará essas informações para fornecer uma avaliação rápida e precisa do risco de câncer de próstata, auxiliando na tomada de decisões clínicas. Este protótipo tem como objetivo não apenas melhorar a eficiência no diagnóstico, mas também proporcionar um suporte significativo na personalização do tratamento dos pacientes. O protótipo pode ser visualizado na Figura 23.

Figura 23 – Protótipo do aplicativo do método proposto

Fonte: Elaborada pelo autor

6 CONCLUSÃO

O câncer de próstata é altamente prevalente em todo o mundo e é uma das principais causas de mortalidade. A aplicação de técnicas de aprendizado de máquina no rastreamento de pacientes com risco de câncer de próstata, em conjunto com variáveis clínicas, pode proporcionar economia de custos para os pacientes e aliviar a carga sobre os sistemas de saúde. Além disso, essas técnicas auxiliam os médicos na tomada de decisões sobre a necessidade de exames adicionais e/ou biópsias de próstata.

No presente trabalho, é proposto um método para auxiliar na triagem de biópsia de câncer de próstata utilizando modelos de aprendizado de máquina com variáveis clínicas como entrada. Os dados utilizados incluem informações dos pacientes, como idade, raça, hipertensão arterial sistêmica, diabetes *mellitus*, tabagismo, etilismo, toque retal e PSA total. Foram incluídos no conjunto de dados 274 pacientes (137 com câncer e 137 sem câncer de próstata). O conjunto de dados foi dividido em 80% para treinamento e 20% para teste. Durante o treinamento, os parâmetros de cada modelo de aprendizado de máquina foram otimizados. Utilizou-se validação cruzada 10-*fold* para evitar *overfitting* e maximizar a robustez e a capacidade de generalização dos modelos. Os modelos foram então usados para prever e classificar os indivíduos em duas classes: câncer ou normal. Diversos critérios de avaliação foram aplicados para medir o desempenho dos modelos.

Os resultados deste estudo demonstram que o modelo *Naive Bayes* apresentou um desempenho excelente, mesmo utilizando apenas oito variáveis clínicas para classificar indivíduos com risco de câncer de próstata. Durante a fase de testes, o modelo alcançou uma acurácia de 89,09%, uma sensibilidade de 92%, uma especificidade de 86,67% e uma AUC de 0,9187. Entre os 55 pacientes do conjunto de teste, houve 4 (7,27%) falsos positivos e apenas 2 (3,64%) falsos negativos.

O método proposto mostrou-se viável para incorporação na prática clínica, permitindo que médicos e pacientes colham seus benefícios, como a redução de biópsias desnecessárias sem comprometer a capacidade de diagnóstico do câncer de próstata. A inovação deste trabalho reside na utilização de variáveis diretamente associadas ao aumento do risco de câncer de próstata e no alto desempenho alcançado com algoritmos de aprendizado de máquina. Este estudo inclui variáveis

clínicas como idade, raça, diabetes *mellitus*, alcoolismo, tabagismo, hipertensão arterial sistêmica, toque retal e PSA total, em vez de se limitar apenas ao PSA e/ou toque retal e/ou exames de imagem. O método proposto é de baixo custo, pois depende de um número reduzido de exames para auxiliar no diagnóstico do câncer.

No entanto, este estudo apresenta algumas limitações, como a ausência de certas variáveis nos prontuários de todos os pacientes, o que impediu sua utilização, como Índice de Massa Corporal (IMC), Hiperplasia Prostática Benigna (HPB), histórico familiar, vasectomia e hipercolesterolemia.

Para trabalhos futuros, pretende-se aumentar o número de variáveis clínicas, incluindo IMC, HPB, histórico familiar, vasectomia e hipercolesterolemia, para verificar seu impacto no diagnóstico de câncer de próstata, além de aumentar a quantidade de prontuários analisados. Também se planeja desenvolver um aplicativo para ser utilizado por médicos urologistas durante as consultas. Transformar o método em uma política pública no SUS, reforçando a contribuição do presente estudo.

Validando a pesquisa feita, foi publicado um artigo científico com classificação A4 em Engenharias IV, conforme extrato Qualis CAPES (2017-2020). Uma cópia do artigo se encontra no ANEXO I. A Tabela 14 lista o artigo científico relacionado diretamente a esta Tese.

Tabela 14 – Artigo produzido relacionado à Tese

Local	Título	Qualis Capes	DOI
REVISTA	Rede Neural Artificial aplicada ao diagnóstico de câncer de próstata	A4	https://doi.org/10.59681/2175-4411.v16.iEspecial.2024.1371

REFERÊNCIAS

ALJUAID, H; ALTURKI, N; ALSUBAIE, N; CAVALLARO, L; LIOTTA, A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Computer Methods and Programs in Biomedicine,* v. 223, 106951, 2022. Disponível em: https://doi.org/10.1016/j.cmpb.2022.106951.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. Redes neurais artificiais: teoria e aplicações. LTC, 2007.

BURGES, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, v. 2, p. 121–167, 1998. Disponível em: https://doi.org/10.1023/A:1009715923555.

BUSHBERG, J. T.; SEIBERT, A. J.; LEIDHOLDT, E. M.; BOONE, J. M. The Essential Physics of Medical Imaging, second ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2001.

CHEN, Y.; XU, C.; ZHANG, Z.; ZHU, A.; XU, X.; PAN, J.; LIU, Y.; WU, D.; HUANG, S.; CHENG, Q. Prostate cancer identification via photoacoustic spectroscopy and machine learning. *Photoacoustic,* v. 23, 100280, 2021. Disponível em: https://doi.org/10.1016/j.pacs.2021.100280.

CICHOSZ, S. L.; JENSEN, M. H.; HEJLESEN, O.; HENRIKSEN, S. D.; DREWES, A. M.; OLESEN, S. S. Prediction of pancreatic cancer risk in patients with new-onset diabetes using a machine learning approach based on routine biochemical parameters. Computer Methods and Programs in Biomedicine, v. 244, 107965, 2024, ISSN 0169-2607. Disponível em: https://doi.org/10.1016/j.cmpb.2023.107965.

CLAESEN, M.; MOOR, B. D. Hyperparameter search in machine learning. *CoRR*, v. 1,1502.02127, 2015. Disponível em: https://doi.org/10.48550/arXiv.1502.02127.

CORREAS, J. M.; HALPERN, E. J.; BARR, R. G.; GHAI, S.; WALZ, J.; BODARD, S.; DARIANE, C.; ROSETTE, J. Advanced ultrasound in the diagnosis of prostate cancer. *World J. Urol*., v. 39, p. 661-676, (2020). Disponível em: https://doi.org/10.1007/s00345-020-03193-0.

CORTES, C.; VAPNIK, V. Support-vector networks. *Mach Learn, v.* 20, p. 273–297, 1995. Disponível em: https://doi.org/10.1007/BF00994018.

COSMA, G.; MCARDLE, S. E.; FOULDS, G. A. et al. Prostate Cancer: Early Detection and Assessing Clinical Risk Using Deep Machine Learning of High Dimensional Peripheral Blood Flow Cytometric Phenotyping Data. *Front Immunol*. v. 12, 2021. Disponível em: https://doi.org/10.3389/fimmu.2021.786828.

CUI, L.; ZHANG, Y.; ZHANG, R.; LIU, Q. H. A modified efficient KNN method for antenna optimization and design, *IEEE Trans. Antennas Propag*, v. 68(10), p. 6858–6866, 2020. Disponível em: https://doi.org/10.1109/TAP.2020.3001743.

DHAGE, S. N.; RAINA, C. K. A review on machine learning techniques. *International Journal on Recent and Innovation Trends in Computing and Communication*, v. 4, p. 395-399, 2016.

DING, C.; PENG, H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *J. Bioinform Comput Biol.,* v. 3(2), p. 185-205, 2005. Disponível em: https://doi.org/10.1142/s0219720005001004.

ERICKSON, B. J.; KORFIATIS, P.; AKKUS, Z.; KLINE, T. L. Machine learning for medical imaging. *Radiographics*, v. 37, p. 505–515, 2017. Disponível em: https://doi.org/10.1148/rg.2017160130.

ESSAM, Y.; HUANG, Y. F.; NG, J. L. et al. Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms. *Sci Rep*, v. 12, 3883, 2022. Disponível em: https://doi.org/10.1038/s41598-022-07693-4.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina. 2ª edição. Editora LTC – Livros Técnicos e Científicos. Rio de Janeiro, 2021.

GANDHI, R. Naïve Bayes classifier, Towards data science, 2018. Disponível em: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, p. 249-256. 2010.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. *Mach Learn,* v. 3, p. 95–99, 1988.

HAYKIN, S. “Neural Networks: A comprehensive foundation”. 2. ed. [S.l.]: Prentice Hall, 1999.

HAZARIKA B. B.; GUPTA, D. Density weighted twin support vector machines for binary class imbalance learning. *Neural Process. Lett.*, v. 54, p. 1091–1130, 2022. Disponível em: https://doi.org/10.1007/s11063-021-10671-y.

HAZARIKA, B. B.; GUPTA, D. Density-weighted support vector machines for binary class imbalance learning. *Neural Comput & Applic., v.* 33, p. 4243–4261, 2021. Disponível em: https://doi.org/10.1007/s00521-020-05240-8.

HE, K.; ZHANG, X.; SHAOQING, R.; SUN, j. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *IEEE international conference on computer vision* (ICCV), p. 1026-1034, 2015. Disponível em: https://doi.org/10.1109/ICCV.2015.123.

HICKS, R. M.; SIMKO, J. P.; WESTPHALEN, A. C.; NGUYEN, H. G., GREENE K. L.; ZHANG, L.; CARROLL, P. R.; HOPE, T. A. Diagnostic accuracy of 68 Ga-PSMA-11 PET/MRI compared with multiparametric MRI in the detection of prostate Cancer.

Radiology, v. 289, p. 730–737, 2018. Disponível em: https://doi.org/10.1148/radiol.2018180788.

HOSSAIN, M. D.; KABIR, M. A.; ANWAR, A. et at. Detecting autism spectrum disorder using machine learning techniques. *Health Inf Sci Syst*, v. 9, 17, 2021. Disponível em: https://doi.org/10.1007/s13755-021-00145-9.

HUA, T. K. A short review on machine learning. *Authorea*, 2022. Disponível em: https://doi.org/10.22541/au.166490976.66390273/v1.

HUGOSSON, J.; ROOBOL, M. J.; MÅNSSON, M.; TAMMELA, T. L. J.; ZAPPA, M.; NELEN, V.; KWIATKOWSKI, M.; LUJAN, M.; CARLSSON, S. V.; TALALA et al. A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer. *European Urology, v.* 76(1), p. 43-51, 2019. Disponível em: https://doi.org/10.1016/j.eururo.2019.02.009.

INCA - Instituto Nacional do Câncer. Atlas de Mortalidade por câncer, 2021. Disponível em: https://www.inca.gov.br/app/mortalidade. Acesso em 14/12/2023.

INCA - Instituto Nacional do Câncer. Câncer de Próstata, 2023. Disponível em: https://www.inca.gov.br/tipos-de-cancer/cancer-de-prostata. Acesso em 14/12/2023.

INCA - Instituto Nacional do Câncer. Registros de Câncer de Base Populacional-RCBP, 2022. Disponível em: https://www.inca.gov.br/numeros-de-cancer/registro-de-cancer-de-base-populacional. Acesso em 23/11/2023.

JUTEL, M.; WIACEK, M. Z.; ORDAK, M.; PFAAR, O.; EIWEGGER, T.; RECHENMACHER, M.; AKDIS, C. A. The artificial intelligence (AI) revolution: how important for scientific work and its reliable sharing. *Allergy*, v.78(8), p. 2085-2088, 2023. Disponível em: https://doi.org/10.1111/all.15778.

KHALILI, H.; RISMANI, M.; NEMATOLLAHI, M. A. et at. Prognosis prediction in traumatic brain injury patients using machine learning algorithms. *Sci Rep*, v. 13, 960, 2023. Disponível em: https://doi.org/10.1038/s41598-023-28188-w.

KIM, M. H.; YOO, S.; CHOO, M. S. et al. The role of the serum 25-OH vitamin D level on detecting prostate cancer in men with elevated prostate-specific antigen levels. *Sci Rep*, v. 12, 14089, 2022. Disponível em: https://doi.org/10.1038/s41598-022-17563-8.

KOHAVI, R. A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International joint Conference on artificial intelligence*, v. 2, p. 1137-1145, 1995. Disponível em: https://dl.acm.org/doi/10.5555/1643031.1643047.

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J., v.* 13, p. 8–17, 2015.

KULKARNI, V.; KULKARNI, M.; PANT, A. Quantum computing methods for supervised learning. *Quant. Mach. Intell.*, v. 3(2), 23, 2021. Disponível em: https://doi.org/10.1007/s42484-021-00050-0.

LEE, C.; LIGHT, A.; ALAA A. et al. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. *The Lancet Digital Health,* v. 3, p. 158-165, 2021. Disponível em: https://doi.org/10.1016/S2589-7500(20)30314-9.

LIANG, B.; LIU, Z.; NIU, Y. B. Shearer cutting pattern recognition based on multi-scale fuzzy entropy and support vector machine. *Earth Environ, Sci*. v. 692, 042062, 2021. Disponível em: https://doi.org/10.1088/1755-1315/692/4/042062.

LIU, J.; DONG, B.; QU, W. et al. Using clinical parameters to predict prostate cancer and reduce the unnecessary biopsy among patients with PSA in the gray zone. *Sci Rep*, v. 10, 5157, 2020. Disponível em: https://doi.org/10.1038/s41598-020-62015-w.

LIU, J.; WANG, Z.; LI, M.; ZHOU, M.; YU, Y.; ZHAN, W. Establishment of two new predictive models for prostate cancer to determine whether to require prostate biopsy when the PSA level is in the diagnostic gray zone (4–10 ng ml−1). *Asian Journal of Andrology*, v. 22(2), p. 213-216, 2019. Disponível em: https://doi.org/10.4103/aja.aja_46_19.

LU, X.; LIU, X.; XIAO, Z.; ZHANG, S.; HUANG, J.; YANG, C.; LIU, S. Self-supervised dual-head attentional bootstrap learning network for prostate cancer screening in transrectal ultrasound images. *Computers in Biology and Medicine*, v. 165, 107337, 2023. Disponível em: https://doi.org/10.1016/j.compbiomed.2023.107337.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. Machine learning: An artificial intelligence approach. *Springer Science & Business Media*, 2013.

MICHIE, D.; SPIEGELHALTER, D. J; TAYLOR, C. C. Machine Learning, Neural and Stastitical Classification. Ellis Horwood, Upper Sadle River, NJ, USA, 1994.

MITCHELL, T. M. Machine Learning. Book News, Inc.®, ISBN: 0070428077. Portland, OR, 1997.

MYSONA, D. P.; PUROHIT, S.; RICHARDSON, K. P. et al. Ovarian recurrence risk assessment using machine learning, clinical information, and serum protein levels to predict survival in high grade ovarian cancer. *Sci Rep*, v. 13, 20933, 2023. Disponível em: https://doi.org/10.1038/s41598-023-47983-z.

NASRABADI, N. M. Pattern recognition and machine learning. Journal of electronic imaging 16, 049901, 2007.
NASTESKI, V. An overview of the supervised machine learning methods. *Horizons B*, v. 4 p. 51–62, 2017.

NATIONAL CANCER INSTITUTE. Tests to Diagnose and Stage Prostate Cancer". American Cancer Society [Online], 2019. Disponível em: https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/how-diagnosed.html. Acesso em 14/01/2024.

NATIONAL CANCER INSTITUTE. About Prostate Cancer. American Cancer Society [Online], 2021. Disponível em: https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html. Acesso em 14/01/2024.

NATIONAL CANCER INSTITUTE. Key Statistics for Prostate Cancer. American Cancer Society [Online], 2024. Prostate Cancer. Disponível em: https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html. Acesso em 14/01/2024.

NATIONAL CANCER INSTITUTE. Prostate Cancer Risk Factors. American Cancer Society [Online], 2023. Prostate Cancer. Disponível em: https://www.cancer.org/cancer/types/prostate-cancer/causes-risks-prevention/risk-factors.html. acesso em 23/11/2023. Acesso em 14/01/2024.

NATIONAL CANCER INSTITUTE. Screening Tests for Prostate Cancer. American Cancer Society [Online], 2020. Disponível em: https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/tests.html. Acesso em 14/01/2024.

NATIONAL CANCER INSTITUTE. What is cancer?. American Cancer Society [Online], 2022. Disponível em: https://www.cancer.org/cancer/understanding-cancer/what-is-cancer.html. Acesso em 02/02/2024.

PARK, J. Y.; YOON, S.; PARK, M. S.; CHOI, H.; BAE, J. H. et al. Development and External Validation of the Korean Prostate Cancer Risk Calculator for High-Grade Prostate Cancer: Comparison with Two Western Risk Calculators in an Asian Cohort. *PLOS ONE,* v. 12(1), e0168917, 2017. Disponível em: https://doi.org/10.1371/journal.pone.0168917.

PINSKY, P. F.; MILLER, E.; PROROK, P.; GRUBB, R.; CRAWFORD, E. D.; ANDRIOLE, G. Extended follow-up for prostate cancer incidence and mortality among participants in the Prostate, Lung, Colorectal and Ovarian randomized cancer screening trial. *BJU Int.,* v. 123(5), p. 854-860, 2019. Disponível em: https://doi.org/10.1111/bju.14580.

PRATI, R. Novas abordagens em Aprendizado de Máquina para a geração de Regras, Classes Desbalanceadas e Ordenação de Casos. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP, 2006.

QUINLAN, R. Induction of decision tree. *Mach Learn*, v. 1, p. 81-106, 1986. Disponível em: https://doi.org/10.1007/BF00116251.

SARKER, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.*, v. 2, 420, 2021. Disponível em: https://doi.org/10.1007/s42979-021-00815-1.

SARTIAS, M. M.; YASAR, A. Performance analysis of ANN and Naïve Bayes classification algorithm for data classification. *Int. J. Intell. Syst. Appl. Eng.*, v. 7, p. 88–91, 2019. Disponível em: https://doi.org/10.18201/ijisae.2019252786.

SCHÖLKOPF, B.; SMOLA, A. Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond. *The MIT Press*, Cambridge, MA, 2018.

SHETTY, S.; SHETTY, S.; SINGH, C. V.; RAO, A. Supervised machine learning: algorithms and applications. *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Applications*, p. 1–16, 2022. Disponível em: https://doi.org/10.1002/9781119821908.ch1.

STABILE, A.; GIGANTI, F.; ROSENKRANTZ, A. B.; TANEJA, S. S.; VILLEIRS, G.; GILL, I. S.; ALLEN, C.; EMBERTON, M.; MOORE, C. M.; KASIVISVANATHAN V. Multiparametric MRI for prostate cancer diagnosis: current status and future directions. *Nat. Rev. Urol.*, v. 17, p. 41–61, 2020. Disponível em: https://doi.org/10.1038/s41585-019-0212-4.

WANG, X.; YANG, W.; WEINREB, J. et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Sci Rep*, v. 7, 15415, 2017. Disponível em: https://doi.org/10.1038/s41598-017-15720-y.

WANI, N. A.; KUMAR, R.; BEDI, J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*, v. 243, 107879, 2024. Disponível em: https://doi.org/10.1016/j.cmpb.2023.107879.

WICKRAMASINGHE, C. S.; AMARASINGHE, K.; MARINO, D. L.; RIEGER, C.; MANIC, M. Explainable unsupervised machine learning for cyber-physical systems. *IEEE Access*, v. 9, p. 131824–131843, 2021.

XING, L.; HE, J.; LI, Y. et al. Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza. *Accid. Anal. Prev.*, v. 135, 105343, 2020. Disponível em: https://doi.org/10.1016/j.aap.2019.105343.

YOO, S.; GUJRATHI, I.; HAIDER, M. A. et al. Prostate Cancer Detection using Deep Convolutional Neural Networks. *Sci Rep, v.* 9, 19518, 2019. Disponível em: https://doi.org/10.1038/s41598-019-55972-4.

ZHENG, S.; JIANG, S.; CHEN, Z.; HUANG, Z.; SHI, W. et al. The roles of MRI-based prostate volume and associated zone-adjusted prostate-specific antigen concentrations in predicting prostate cancer and high-risk prostate cancer. *PLOS ONE*, v. 14(11), e0218645, 2019. Disponível em: https://doi.org/10.1371/journal.pone.0218645.

ANEXO I – artigo: Rede neural artificial aplicada ao diagnóstico de câncer de próstata

CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Rede neural artificial aplicada ao diagnóstico de câncer de próstata

Artificial neural network applied to prostate cancer diagnosis

Red neuronal artificial aplicada al diagnóstico del cáncer de próstata

Wesley Batista Dominices de Araujo[1], Ewaldo Eder Carvalho Santana[2], Nilviane Pires Silva[1], Carlos Magno Sousa Junior[2], Giullianno Lopes Moura[3], José Arnon Linhares Moraes dos Santos[3], Paloma Larissa Arruda Lopes[4], Wesley do Nascimento Silva[4], João Pedro Pereira Gonçalves[4], Felipe Castelo Branco Rocha Silva[4]

1 Programa de Pós-graduação em Engenharia Elétrica, Universidade Federal do Maranhão - UFMA, São Luís (MA), Brasil.
2 Departamento de Engenharia da Computação, Universidade Estadual do Maranhão - UEMA, São Luís (MA), Brasil.
3 Hospital Universitário da Universidade Federal do Maranhão – HU-UFMA, São Luís (MA), Brasil.
4 Departamento de Medicina, Universidade Federal do Maranhão - UFMA, São Luís (MA), Brasil.

Autor correspondente: Prof. Me. Wesley Batista Dominices de Araujo
E-mail: wesleydominices@gmail.com

Resumo

Objetivo: Desenvolver um método para auxiliar no diagnóstico de câncer de próstata utilizando Rede Neural Artificial aplicada às variáveis clínicas. Método: Foi realizada uma pesquisa observacional retrospectiva em 274 prontuários médicos do Hospital Universitário da Universidade Federal do Maranhão. Foram utilizadas as variáveis clínicas: idade, raça, hipertensão arterial sistêmica, diabetes mellitus, tabagismo, etilismo, toque retal e PSA total. Foi criado um modelo de Rede Neural Artificial para classificação preditiva. Resultados: O modelo apresentou acurácia de 80%, sensibilidade de 80%, especificidade de 80% e área sob a curva ROC de 0,9027. Conclusão: Obteve-se um excelente desempenho na predição do câncer de próstata. Este método pode ser incorporado à prática clínica, pois médicos e pacientes podem colher os benefícios dele, reduzindo biópsias desnecessárias, sem comprometer a capacidade de diagnosticar o câncer de próstata.

Descritores: Câncer de próstata; Diagnóstico; Rede Neural Artificial.

CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Abstract

Objective: Develop a method to assist in the diagnosis of prostate cancer using Artificial Neural Network applied to clinical variables. Method: Retrospective observational research was carried out on 274 medical records from the University Hospital of the Federal University of Maranhão. The following clinical variables were used: age, race, systemic arterial hypertension, diabetes mellitus, smoking, alcohol consumption, digital rectal exam, and total PSA. An Artificial Neural Network model was created for predictive classification. Results: The model presented an accuracy of 80%, sensitivity of 80%, specificity of 80% and area under the ROC curve of 0.9027. Conclusion: Excellent performance was obtained in predicting prostate cancer. This method can be incorporated into clinical practice as doctors and patients can reap the benefits of it by reducing unnecessary biopsies without compromising the ability to diagnose prostate cancer.

Keywords: Prostate cancer; Diagnosis; Artificial Neural Network.

Resumen

Objetivo: Desarrollar un método para ayudar en el diagnóstico del cáncer de próstata utilizando Rede Neuronal Artificial aplicadas a variables clínicas. Método: Se realizó una investigación observacional retrospectiva en 274 prontuarios del Hospital Universitario de la Universidad Federal de Maranhão. Se utilizaron las variables clínicas: edad, raza, hipertensión arterial sistémica, diabetes mellitus, tabaquismo, consumo de alcohol, examen de tacto rectal y PSA total. Se creó un modelo de Red Neuronal Artificial para la clasificación predictiva. Resultados: El modelo presentó una precisión del 80%, sensibilidad del 80%, especificidad del 80% y área bajo la curva ROC de 0,9027. Conclusión: Se obtuvo excelente desempeño en la predicción del cáncer de próstata. Este método se puede incorporar a la práctica clínica, ya que los médicos y los pacientes pueden aprovechar sus beneficios al reducir las biopsias innecesarias sin comprometer la capacidad de diagnosticar el cáncer de próstata.

Descriptores: Cáncer de próstata; Diagnóstico; Red Neural Artificial.

Introdução

O câncer de próstata (PCa) é um dos tipos de câncer mais prevalentes entre os homens em nível mundial, com cerca de 1 milhão de casos por ano[1]. Adicionalmente, sua incidência em nível global está crescendo, sendo que mais de 80% dos homens

diagnosticados ainda não apresentam metástase[2]. A última estimativa nos Estados Unidos, para 2024, apontou o PCa como sendo o segundo tipo mais frequente em homens, com cerca de 299.010 novos casos e 35.250 mortes[3]. No Brasil, a situação não é diferente, somente no ano de 2021 houve 16.301 mortes por PCa, e a estimativa de novos casos no ano de 2024 é de 71.740, que corresponde a 30,0% dos tumores incidentes no gênero masculino[4].

Alguns fatores de risco podem influenciar para que uma pessoa possa contrair o PCa, tais como: idade, raça/etnia, histórico familiar, alterações genéticas herdadas, dieta, obesidade, tabagismo, exposição química, inflamação na próstata, infecções transmitidas sexualmente e vasectomia[5].

Para iniciar o diagnóstico de PCa são utilizados, basicamente, o exame clínico, chamado de exame de toque retal e o exame laboratorial, que é a dosagem do PSA (*Prostate-Specific Antigen*), iniciado em 1986[6]. Se houver alterações em ambos os exames, outros exames poderão ser solicitados, como endoscópios, radiológicos e ressonância magnética. Mas nenhum desses exames tem 100% de precisão, levando ao paciente a ter que fazê-los, às vezes, desnecessariamente, e com um alto custo financeiro.

Atualmente, a biópsia prostática é o único procedimento capaz de diagnosticar o PCa. Tradicionalmente é feita guiada por ultrassom transretal (TRUS), entretanto, tem uma taxa de detecção de câncer inferior a 30% em uma próstata benigna. Cerca de 55% das biópsias dão resultado negativo, mesmo que o paciente esteja com câncer de próstata. Além de ter um custo financeiro elevado, é muito invasiva, dolorida, e o paciente tem 5% de chance de desenvolver infecções como urosepse. Mais de 60% dos resultados da biópsia da próstata são negativos nos Estados Unidos a cada ano, esta é uma razão para filtrar a indicação de biópsia, diminuindo as biópsias negativas e aumentando a acurácia do exame[6]. Outros exames, além do TRUS, que têm sido utilizados no diagnóstico de PCa, são a ressonância magnética e a tomografia computadorizada. Contudo, não fornecem informações sobre a composição química, apresentam baixa resolução e incorrem em custos elevados[7].

Diante disso, é importante o uso de métodos alternativos, que servirão para auxiliar o médico na tomada de decisão em relação à indicação ou não da biópsia da próstata. Esses métodos são baseados no aprendizado de máquina, do inglês, *Machine Learning* (ML). O ML é um ramo da inteligência artificial (IA) que se baseia na ideia de o sistema aprender um padrão a partir de uma base de dados, usando

ferramentas probabilísticas e estatísticas, e tomando decisões ou previsões sobre os novos dados[8].

A Rede Neural Artificial (RNA) é um método de IA, que funciona como um modelo matemático não linear que mimetiza o cérebro humano nas características de aprendizagem e tomada de decisão, estimulando as habilidades cognitivas humanas. A RNA é usada para mapear e prever resultados em relacionamentos complexos entre determinadas entradas e saídas procuradas, e pode ser usada para encontrar padrões em conjuntos de dados. A RNA pode ser complexa, com camadas ocultas, e pode ser treinada para reconhecer padrões e classificá-los com alta precisão[9]. Algoritmos de Aprendizado de máquina têm sido bastante utilizados em pesquisas sobre diagnóstico de câncer[10,11].

A principal motivação desta pesquisa é utilizar uma técnica de IA aplicada aos dados clínicos dos pacientes para fazer a predição de forma a auxiliar no diagnóstico de câncer de próstata. Além disso a patologia leva há muitos óbitos, logo é necessário um método que tenha um baixo custo para auxílio ao diagnóstico precoce e aumento da sobrevida do paciente. O baixo custo é referente ao paciente não ter a necessidade de realizar alguns exames, tais como: hemograma, coagulograma, elementos anormais do sedimento (EAS), urocultura, ureia, creatinina, biópsia de próstata guiada por TRUS e anatomopatológico de biópsia prostática.

Esta pesquisa apresenta como principais contribuições:

- Um novo método para auxiliar na triagem do câncer de próstata utilizando variáveis clínicas, tais como: idade, raça, diabetes mellitus, etilismo, tabagismo, hipertensão arterial sistêmica, exame de toque retal e PSA total, juntamente com um modelo de Rede Neural Artificial;
- Um método que evite biópsias desnecessárias;
- Uma nova estratégia para viabilizar um método otimizado de auxílio ao diagnóstico precoce de baixo custo e com variáveis não invasivas.

Método

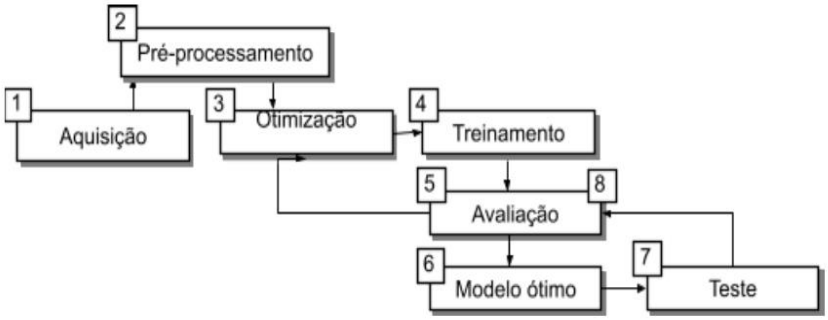
Declaração de ética

O presente estudo foi aprovado pela Comissão Científica (COMIC) e pelo Comitê de Ética em Pesquisa (CEP), conforme parecer CAAE: 45444621.6.0000.5086, e parecer técnico consubstanciado ID: 4.679.671, ambos pertencentes ao Hospital

Universitário da Universidade Federal do Maranhão (HU-UFMA), localizado na cidade de São Luís, capital do Estado do Maranhão, Brasil. Todos os princípios éticos dos direitos do paciente foram atendidos e os nomes dos participantes não foram utilizados. O HU-UFMA autorizou o estudo e dispensou o termo de consentimento livre e esclarecido, pois os dados utilizados eram para um estudo retrospectivo, e aprovou todos os experimentos.

O método proposto é composto de oito etapas. O diagrama em blocos dele é mostrado na Figura 1.

Figura 1 - Diagrama em blocos do método proposto



Aquisição dos dados

Para que os dados fossem adquiridos, uma pesquisa observacional retrospectiva foi realizada. Foram obtidos 274 prontuários médicos (137 com câncer de próstata e 137 sem câncer de próstata) dos pacientes que fazem acompanhamento no setor de urologia do HU-UFMA, no período de janeiro de 2017 a outubro de 2023. Os critérios de inclusão na pesquisa foram: ter realizado a biópsia da próstata e ter no prontuário todas as variáveis necessárias à pesquisa. Nos dados adquiridos há informações sociodemográficas e variáveis clínicas, incluindo o resultado da biópsia da próstata. As variáveis clínicas selecionadas de cada paciente utilizadas neste trabalho foram: idade, raça, hipertensão arterial sistêmica (HAS), diabetes mellitus (DM), tabagismo, etilismo, toque retal e PSA total, pois estão entre os fatores de risco associados ao câncer de próstata, e todas são amplamente aceitas na literatura médica como relacionadas ao câncer de próstata[12]. A variável toque retal se refere ao peso estimado da próstata no momento da avaliação clínica.

Pré-processamento

O pré-processamento foi feito da seguinte forma:

- **Idade**: foi utilizada a idade, em anos, do paciente no dia da consulta, ou seja, um valor numérico inteiro.
- **Raça**: os números 1, 2, 3 ou 4 foram utilizados para codificar as raças (1 significa branca, 2 significa parda, 3 significa preta e 4 significa indígena).
- **HAS** e **DM**: os números 1 ou 2 foram utilizados (1 significa sim e 2 significa não).
- **Tabagismo** e **Etilismo**: os números 1, 2 ou 3 foram utilizados (1 significa sim, 2 significa não, 3 significa que o paciente foi exposto ao risco anteriormente). Para o etilismo, bastou o paciente assumir o consumo de álcool, o grau de alcoolismo não foi considerado para esta variável. Da mesma forma para o tabagismo, bastou o paciente informar que está exposto ao risco, sem considerar a quantidade de maços consumidos.
- **Toque**: foi utilizado o peso estimado da próstata em gramas (g).
- **PSA total**: foi utilizado o valor, numérico real em ng/ml, contido no exame de sangue.
- **Rótulo**: foi utilizado para identificar cada amostra em uma classe (normal ou câncer). No aprendizado supervisionado, é necessária uma variável de saída, também chamada de variável alvo, para que no momento do treinamento do algoritmo de ML ele saiba a qual classe pertence àquela amostra específica, sendo assim, foi utilizado o valor 0 (zero) ou 1 (um), zero significa que o resultado da biópsia foi negativo, e 1 quando o resultado foi positivo (*Gleason* > 0).

Otimização

Encontrar os parâmetros mais eficazes para o processo de aprendizagem do modelo é geralmente referido como otimização de hiperparâmetros[13]. Foi realizada uma extensa exploração de hiperparâmetros usando o parâmetro *'OptimizeHyperparameters'* da função 'fit' do *Matlab*®, garantindo uma configuração otimizada para o modelo utilizado. Várias combinações para ajuste de hiperparâmetros foram aplicadas ao algoritmo de RNA. Para alcançar a melhor configuração, foram exploradas, sistematicamente, várias opções de valores dos parâmetros para o algoritmo de RNA utilizado neste trabalho, dentre as quais se destacam: quantidade de neurônios na camada oculta, função de ativação, *lambda*, inicialização dos pesos das camadas e inicialização dos pesos do *bias* das camadas.

Treinamento

Para a etapa de treinamento, foram utilizadas 80% das amostras. Estes dados foram divididos aleatoriamente, formando o conjunto de dados de treinamento. O modelo foi treinado utilizando algumas funções do *software Matlab*®.

Além disso, para evitar o *overfitting*, foi utilizada a técnica de validação cruzada (*10-fold cross-validation*), onde o conjunto de dados foi dividido em dez subconjuntos: nove subconjuntos para treinamento e um subconjunto para validação. Isso foi repetido dez vezes até que todos os subconjuntos fossem utilizados para treinar o classificador. O resultado final obtido no treinamento foi a média das dez iterações.

O algoritmo de RNA foi utilizado no treinamento do modelo. Após o treinamento foi feita a avaliação dele através das medidas de desempenho, para verificar se o modelo tem um desempenho satisfatório nesta etapa.

Modelo ótimo, Testes e Avaliação do modelo

Após a conclusão do treinamento com os hiperparâmetros otimizados, obteve-se o modelo otimizado, a partir do algoritmo de RNA, com os melhores parâmetros. Desta forma, pode-se aplicar os dados de teste ao modelo ótimo construído a partir do algoritmo de RNA, para avaliar a capacidade preditiva dele e seu poder de generalização para novos dados, não utilizados na etapa de treinamento. Foram utilizados 20% dos dados no conjunto de teste. Este conjunto de dados é utilizado como entrada para o modelo ótimo. O teste é realizado comparando a classe alvo (rótulo) já conhecida, com a variável de predição estimada pelo teste.

Para avaliar o modelo preditivo, a confiabilidade do método e do classificador foram utilizadas a acurácia, sensibilidade, especificidade e área sob a curva ROC, em conjunto com a técnica estatística de validação cruzada *10-fold cross-validation*[14]. Em modelos preditivos, a metodologia de desempenho usual é medida calculando algumas medidas estatísticas sobre o resultado dos testes. Os resultados da classificação dos testes podem ser divididos em: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN). Sendo VP e VN o número de amostras que são corretamente classificadas, respectivamente, como positiva ou negativa pelo classificador, FP e FN representam o número de amostras correspondentes aos casos que são erroneamente classificados como positivo ou negativo, respectivamente. Tais números são utilizados para gerar medidas capazes

de quantificar o desempenho da metodologia, para avaliar o quão este é eficiente e se os objetivos foram alcançados.

A Acurácia (Acu), que é a taxa de acerto do classificador durante a fase de teste, e é definida por:

$$Acu = \frac{VP+VN}{VP+VN+FP+FN} \quad (1)$$

A Sensibilidade (Sen) é a proporção de verdadeiros positivos que são corretamente classificados, ou seja, dos pacientes doentes, quantos foram corretamente identificados como doentes, e é definida por:

$$Sen = \frac{VP}{VP+FN} \quad (2)$$

A Especificidade (Esp) é a proporção de verdadeiros negativos que são corretamente classificados, ou seja, dos pacientes não-doentes, quantos foram corretamente identificados como não-doentes, e é definida por:

$$Esp = \frac{VN}{VN+FP} \quad (3)$$

A Área sob a Curva ROC (AUC) é uma forma de representar graficamente a relação entre a taxa de falsos positivos (TFP) e a taxa de verdadeiros positivos (TVP) ou sensibilidade.

$$TFP = \frac{FP}{FP+VN} \quad (4)$$

$$TVP = \frac{VP}{VP+FN} \quad (5)$$

O gráfico ROC é bidimensional. Os valores da TVP são representados no eixo y e os valores da TFP no eixo x, no plano cartesiano. O desempenho do classificador é então plotado nessa curva. A medida da AUC produz valores entre 0 e 1. Valores mais próximos de 1 são considerados melhores.

Resultados e Discussão

Nesta Seção, os resultados obtidos serão apresentados. O método proposto foi implementado usando o *software MatLab*® v. R2023b e o *software IBM SPSS*®.

Foi realizada uma pesquisa observacional retrospectiva através dos prontuários de 274 pacientes (137 sem câncer de próstata e 137 com câncer de próstata) do setor de Urologia do HU-UFMA, e foi criado um *dataset* a partir dos mesmos. Foram

utilizadas oito variáveis clínicas (idade, raça, hipertensão arterial sistêmica (HAS), diabetes mellitus (DM), etilismo, tabagismo, toque e PSA total), que serviram de entrada pelo algoritmo de Rede Neural Artificial usado neste trabalho. Foi criada uma variável chamada de "Rótulo", a qual identifica se o resultado da biópsia foi positivo (*Gleason* > 0), ou seja, o paciente tem câncer de próstata, ou se o resultado foi negativo (normal), ou seja, o paciente não tem câncer de próstata.

A média de idade dos pacientes no momento do diagnóstico foi de 67,23 anos (desvio padrão ± 7,846), com intervalo entre 46 e 92 anos. A maioria dos pacientes era pardo (70,1%). A média do exame do toque (peso da próstata) foi de 57,81 g (desvio padrão ± 28,78), com intervalo entre 13 e 200, e a média do PSA Total foi de 8,25 ng/ml (desvio padrão ± 11), com intervalo entre 0,24 e 79,41 ng/ml. Com relação à frequência, têm-se HAS positiva: 58,4% e negativa: 41,6%; DM positiva: 22,6% e negativa: 77,4%; Tabagismo positivo: 8%, negativo: 46% e Ex: 46%; Etilismo positivo: 31,4%, negativo: 27,7% e Ex: 40,9%, conforme exibido na Tabela 1.

Tabela 1 - média ou percentual, desvio padrão, mediana, intervalo e frequência das variáveis

Variável	Média ou Percentual	Desvio padrão	Mediana	Intervalo	Frequência (%)		
					Sim	Não	Ex
Idade	67,23	7,846	68	46-92	-	-	-
Raça	-	-	-	-	-	-	-
Branca	24,8%	-	-	-	-	-	-
Parda	70,1%	-	-	-	-	-	-
Preta	4,7%	-	-	-	-	-	-
Indígena	0,4%	-	-	-	-	-	-
HAS	-	-	-	-	58,4	41,6	-
DM	-	-	-	-	22,6	77,4	-
Tabagismo	-	-	-	-	8	46	46
Etilismo	-	-	-	-	31,4	27,7	40,9
Toque retal	57,81	28,78	50	13-200	-	-	-
PSA total	8,25	11	5,22	0,24–79,41	-	-	-
Rótulo	-	-	-	-	50	50	-

Abreviações: HAS = *Hipertensão Arterial Sistêmica*; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen*.

Considerando a distribuição das variáveis por classe (câncer ou normal) obteve-se algumas informações relevantes. Com relação aos pacientes com diagnóstico negativo da biópsia (normal) têm-se, em relação à média: idade de 69,06 anos (intervalo: 51-92; desvio padrão: ± 7,86), toque retal de 66,34 g (intervalo: 19-200; desvio padrão: ± 30,92) e PSA de 5,61 ng/ml (intervalo: 0,24-79,41; desvio padrão: ± 9,73). Já em relação aos pacientes com câncer obteve-se as seguintes informações, em relação à média: idade de 65,39 anos (intervalo: 46-79; desvio

padrão: ± 7,41), toque retal de 49,29 g (intervalo: 13-126; desvio padrão: ± 23,66) e PSA de 10,89 ng/ml (intervalo: 2,47-78; desvio padrão: ± 11,59). Percebe-se que em relação à média de idade não há diferença tão significativa entre os pacientes com câncer ou sem câncer, pois ambas as médias são muito próximas. O média do toque retal (peso estimado da próstata) nos pacientes com câncer é maior do que os pacientes sem câncer. Provavelmente devido a esse aumento que os pacientes foram submetidos à biópsia da próstata. Em relação ao PSA total, observa-se que a média é bem maior nos pacientes com câncer do que os sem câncer (normal).

Em relação à frequência, para os pacientes com diagnóstico normal, obteve-se as seguintes informações: raça branca (28,5%), parda (67,2%), preta (3,6%) e indígena (0,7%); HAS: 56,2% (sim) e 43,8% (não); DM: 19% (sim) e 81% (não); Tabagismo: 7,3% (sim), 46% (não) e 46,7% (ex); Etilismo: 21,9% (sim), 31,4% (não) e 46,7% (ex). Já para os pacientes com diagnóstico positivo (câncer) da biópsia têm-se: raça branca (21,2%), parda (73%) e preta (5,8%); HAS: 60,6% (sim) e 39,4% (não); DM: 26,3% (sim) e 73,7% (não); Tabagismo: 8,8% (sim), 46% (não) e 45,3% (ex); Etilismo: 40,9% (sim), 24,1% (não) e 35% (ex). Estas informações podem ser visualizadas na Tabela 2.

Tabela 2 – Distribuição das frequências das variáveis raça, HAS, DM, tabagismo e etilismo, por classe (normal ou câncer)

Variável	Normal (%)	Câncer (%)
Raça (Branca)	28,5	21,2
Raça (Parda)	67,2	73
Raça (Preta)	3,6	5,8
Raça (Indígena)	0,7	0
HAS (Sim)	56,2	60,6
HAS (Não)	43,8	39,4
DM (Sim)	19	26,3
DM (Não)	81	73,7
Tabagismo (Sim)	7,3	8,8
Tabagismo (Não)	46	46
Tabagismo (Ex)	46,7	45,3
Etilismo (Sim)	21,9	40,9
Etilismo (Não)	31,4	24,1
Etilismo (Ex)	46,7	35

Abreviações: HAS = *Hipertensão Arterial Sistêmica*; DM = *Diabetes Mellitus*; PSA = *Prostate-Specific Antigen*.

Analisando os resultados obtidos na Tabela 2, percebe-se que em relação à frequência da raça, tanto nos pacientes com câncer ou sem câncer é maior nas pessoas pardas. Em relação à HAS, tanto nos pacientes com câncer ou sem câncer é

maior nos pacientes que tem pressão alta. Em relação à DM, tanto nos pacientes com câncer ou sem câncer, é maior nos pacientes que não têm diabetes. Em relação à frequência do tabagismo, percebe-se que tanto nos pacientes com câncer ou sem câncer, a frequência é maior naqueles que nunca fumaram ou são ex-fumantes. Em relação ao etilismo, a frequência de quem faz uso de bebidas alcoólicas é bem maior nos pacientes com câncer de próstata, praticamente o dobro da frequência dos que não tem câncer.

Após a parametrização inicial, os dados estão prontos para serem utilizados pelo algoritmo de RNA. Os dados foram divididos aleatoriamente. 80% deles foram usados para o treinamento do modelo. Foi aplicada a validação cruzada 10-*fold*. Durante a otimização dos parâmetros do modelo de RNA, foram verificadas 30 parametrizações diferentes. Os melhores parâmetros, que geraram o melhor resultado, são apresentados na Tabela 3.

Tabela 3 – Os melhores parâmetros obtidos para o modelo de RNA durante o treinamento

Modelo	Parâmetros
Rede Neural Artificial	Quantidade de neurônios na camada oculta = 11; Função de ativação = ReLU; *Lambda* = 1,269023252364904e-04; Inicialização dos pesos das camadas = Glorot; Inicialização do *bias* das camadas = zeros;

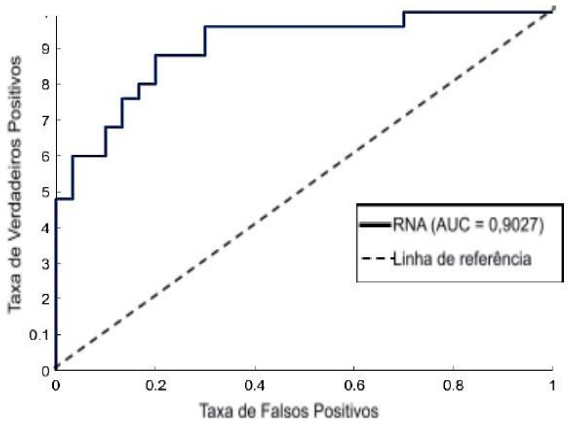
Após o treinamento, foi obtido o modelo ótimo para o algoritmo de RNA. Assim, ele foi testado utilizando o subconjunto de teste com 20% dos dados restantes que não foram utilizados durante o treinamento. Assim, extraiu-se as medidas de desempenho, para avaliar o método proposto, que estão exibidas na Tabela 4. A Figura 2 mostra a AUC gerada para o modelo proposto.

Tabela 4 – Análise de desempenho aplicado ao subconjunto de teste

Modelo	VP	VN	FP	FN	Acu (%)	Sen (%)	Esp (%)	AUC (IC 95%)
Rede Neural Artificial	20	24	6	5	80	80	80	0,9027 (0,811 – 0,995)

Abreviações: VP = Verdadeiro Positivo; VN = Verdadeiro Negativo; FP = Falso Positivo; FN = Falso Negativo; Acu = Acurácia; Sen = Sensibilidade; Esp = Especificidade; AUC = *Area under the Receiver Operating Characteristic*; IC = Intervalo de confiança.

Figura 2 – AUC do classificador RNA



Numerosos estudos já foram realizados sobre o diagnóstico de câncer de próstata usando aprendizado de máquina.

Wang *et al.*[15] tinham disponíveis 2.602 imagens da próstata de 172 pacientes. Fizeram dois experimentos, no primeiro, usaram o modelo BoW (*Bag-of-Word*) para agregar as características codificadas em uma representação vetorial para cada imagem. A classificação das imagens foi feita com um classificador SVM (*Support Vector Machine)* linear, e obtiveram sensibilidade de 49,4%, especificidade de 81,7% e AUC de 0,70. Já no segundo, aplicaram aprendizado profundo com rede neural convolucional (DCNN), e obtiveram sensibilidade de 69,3%, especificidade de 83,9% e AUC de 0,84.

Liu *et al.*[16] tinham 197 pacientes submetidos à biópsia de próstata com PSA entre 4 e 10 ng/ml. Utilizaram como variáveis clínicas: idade, volume da próstata (PV), PSA livre/total (f/tPSA) e densidade do PSA (PSAD). Testaram o método com regressão logística, e obtiveram sensibilidade de 75,4%, especificidade de 75,8% e AUC de 0,775.

Liu *et al.*[17] desenvolveram um modelo multivariáveis para prever o câncer de próstata entre 235 pacientes na zona cinzenta (4 a 10 ng/ml) do PSA total. Utilizaram como variáveis clínicas: idade, tPSA, fPSA, PV, f/tPSA e PSAD, e obtiveram AUC de 0,79.

Park *et al.*[18] desenvolveram a Calculadora Coreana de Risco de Câncer de Próstata para Câncer de Próstata de Alto Grau (KPCRC-HG) que prevê a probabilidade de câncer de próstata com *Gleason* maior ou igual a 7. Utilizaram um modelo de regressão logística com base nos dados de 602 pacientes. As variáveis utilizadas foram: idade, toque retal, PSA total, TRUS, PV e TZV (volume da zona de transição). A AUC foi de 0,84.

Yoo *et al.*[19] usaram um classificador *Random Forest* para classificar 427 pacientes (175 com câncer e 252 sem câncer de próstata) utilizando variáveis estatísticas de primeira ordem. Obtiveram como melhor desempenho uma AUC de 0,87.

Chen *et al.*[20] utilizaram espectroscopia fotoacústica e aprendizado de máquina para identificação do câncer de próstata em 101 pacientes (90 para treinamento e 11 para teste). Obtiveram uma AUC de 0,851 para a LDA (*Linear Discriminant Analysis*) e 0,862 para a QDA (*Quadratic Discriminant Analysis*). Sensibilidade de 78,2% e especificidade de 87,1% para a QDA.

A Tabela 5 mostra a comparação entre o método proposto com os trabalhos publicados anteriormente.

Tabela 5 – Comparação com outros trabalhos relacionados ao câncer de próstata

Trabalho	Algoritmo	Método	Sen (%)	Esp (%)	AUC
Wang *et al.*[15]	SVM	MRI	49,4	81,7	0,700
Liu *et al.*[16]	Regressão Logística	Multivariáveis	75,4	75,8	0,775
Liu *et al.*[17]	Regressão Logística	Multivariáveis	-	-	0,790
Wang *et al.*[15]	DCNN	MRI	69,3	83,9	0,840
Park *et al.*[18]	Regressão Logística	Multivariáveis	-	-	0,840
Chen *et al.*[20]	LDA	Fotoacústica	-	-	0,851
Chen *et al.*[20]	QDA	Fotoacústica	78,2	87,1	0,862
Yoo *et al.*[19]	Random Forest	MRI	-	-	0,870
Método proposto	**RNA**	**Multivariáveis**	**80**	**80**	**0,9027**

Abreviações: Sen = Sensibilidade; Esp = Especificidade; AUC = *Area under the Receiver Operating Characteristic*; SVM = *Support Vector Machine*; MRI = *Magnetic Resonance Image*; DCNN = Deep *Convolutional Neural Network;* LDA = *Linear Discriminant Analysis*; QDA = *Quadratic Discriminant Analysis*; RNA = Rede Neural Artificial.

Em síntese, o método proposto demonstrou-se competitivo na predição do câncer de próstata. No que tange à sensibilidade, que avalia a capacidade do modelo em identificar corretamente os casos positivos de câncer, o método proposto mostrou-se superior a todos os estudos correlatos, atingindo uma taxa de acerto de 80%. No que concerne à especificidade, que mede a capacidade do modelo em fornecer resultados negativos para pacientes não-doentes, o método proposto apresentou resultados ligeiramente inferiores ao trabalho de Wang et al. [15], que utilizaram múltiplas imagens de cada paciente, assumindo independência entre elas e classificando-as individualmente. Também apresentou resultados inferiores ao trabalho de Chen et al. [20], contudo, este último avaliou apenas 11 pacientes, uma amostra insuficiente para validar a eficácia do método na predição de casos negativos. Em relação à AUC (Área sob a Curva), que representa a habilidade do modelo em

distinguir entre pacientes com e sem a doença, o método proposto superou todos os estudos relacionados, alcançando um valor de 0,9027.

Conclusão

O câncer de próstata tem alta prevalência em todo o mundo e causa muitas mortes. A utilização de técnicas de inteligência artificial para auxiliar no rastreamento de pacientes com risco de câncer de próstata juntamente com variáveis clínicas pode ajudar o paciente a economizar dinheiro e reduzir a carga sobre os sistemas de saúde, além de auxiliar o médico na tomada de decisão sobre a solicitação de exames adicionais e biópsia da próstata. Os resultados deste trabalho mostraram que utilizando um modelo de RNA houve um excelente desempenho na predição do câncer de próstata. Também ficou demonstrado que pode ser incorporado à prática clínica, pois médicos e pacientes podem colher facilmente os benefícios dele, reduzindo biópsias desnecessárias sem comprometer a capacidade de diagnosticar o câncer de próstata.

Este trabalho obteve alto desempenho usando um modelo de Rede Neural Artificial aplicado às variáveis clínicas: idade, raça, hipertensão arterial sistêmica, diabetes mellitus, etilismo, tabagismo, toque retal e PSA total para triagem da biópsia do câncer de próstata. O método tem um baixo custo, já que depende de poucos exames, além de todas as variáveis não serem invasivas.

Agradecimentos

Ao Hospital Universitário da Universidade Federal do Maranhão pela aprovação do projeto de pesquisa e disponibilização dos prontuários médicos dos pacientes.

Referências

1. Kim MH, Yoo S, Choo MS, Cho MC, Son H, Jeong H. The role of the serum 25-OH vitamin D level on detecting prostate cancer in men with elevated prostate-specific antigen levels. Sci Rep. 2022 Aug;12:14089. Available from: https://doi.org/10.1038/s41598-022-17563-8.

2. Lee C, Light A, Alaa A, Thurtle D, Schaar M, Gnanapragasam VJ. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. The Lancet Digital Health. 2021 Mar;3:158-165. Available from: https://doi.org/10.1016/S2589-7500(20)30314-9.

3. American Cancer Society. Key Statistics for Prostate Cancer [Internet]. 2024 [cited 2024 Jan 19]. Available from: https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html.

4. INCA. Câncer de Próstata. Instituto Nacional de Câncer [Internet]. 2023 [cited 2023 Aug 16]. Available from: https://www.inca.gov.br/tipos-de-cancer/cancer-de-prostata.

5. American Cancer Society. Prostate Cancer Risk Factors. [Internet]. 2023 [cited 2023 Nov 22]. Available from: https://www.cancer.org/cancer/types/prostate-cancer/causes-risks-prevention/risk-factors.html.

6. Cosma G, McArdle SE, Foulds GA, Hood SP, Reeder S, Johnson C, et al. Prostate Cancer: Early Detection and Assessing Clinical Risk Using Deep Machine Learning of High Dimensional Peripheral Blood Flow Cytometric Phenotyping Data. Front Immunol. 2021 Dec;12:786828. Available from: https://doi.org/10.3389/fimmu.2021.786828.

7. Correas JM, Halpern EJ, Barr RG, Ghai S, Walz J, Bodard S, Dariane C, Rosette J, Advanced ultrasound in the diagnosis of prostate cancer, World J. Urol. 2020 Apr;39:661-676. Available from: https://doi.org/10.1007/s00345-020-03193-0.

8. Nasrabadi NM. Pattern Recognition and Machine Learning. Journal of Electronic Imaging. 2007 Oct;16(4):049901. Available from: https://doi.org/10.1117/1.2819119.

9. Faceli K, Lorena AC, Gama J, Carvalho ACPLF. Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina. 2ª edição. Editora LTC – Livros Técnicos e Científicos. Rio de Janeiro, 2021.

10. Fonseca AU, Felix JP, Vieira GS, Rocha BM, Nogueira EA, Araújo CEE, et al. Diagnosticando Tuberculose com Redes Neurais Artificiais e Recursos BPPC. J Health Inform [Internet]. 20º de julho de 2023 [citado 16º de maio de 2024];15(Especial). Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1106

11. Santos PD, Yahata E, Piheiro TS, Oliveira FS de, Simões PW. Algoritmos de Machine Learning para Predição da Sobrevida do Câncer de Mama. J Health Inform [Internet]. 20º de julho de 2023 [citado 16º de maio de 2024];15(Especial). Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1091

12. Nacional Cancer Institute. Prostate cancer risk factors. American Cancer Society [Online], 2023. Prostate Cancer. Available from: https://www.cancer.org/cancer/types/prostate-cancer/causes-risks-prevention/risk-factors.html.

13. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging, Radiographics. 2017 Feb;37(2):505–515. Available from: https://doi.org/10.1148/rg.2017160130.

14. Kohavi R. A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: International joint Conference on artificial intelligence. 1995 Aug;2:1137-1145. Available from: https://dl.acm.org/doi/10.5555/1643031.1643047.

15. Wang X, Yang W, Weinreb J, Han J, Li Q, Kong X, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci Rep. 2017 Nov;7:15415. Available from: https://doi.org/10.1038/s41598-017-15720-y.

16. Liu J, Wang ZQ, Li M, Zhou MY, Yu YF, Zhan WW. Establishment of two new predictive models for prostate cancer to determine whether to require prostate biopsy when the PSA level is in the diagnostic gray zone (4–10 ng ml^{-1}). Asian Journal of Andrology. 2019 Mar;22(2):213-216. Available from: https://doi.org/10.4103/aja.aja_46_19.

17. Liu J, Dong B, Qu W, Wang J, Xu Y, Yu S, et al. Using clinical parameters to predict prostate cancer and reduce the unnecessary biopsy among patients with PSA in the gray zone. Sci Rep. 2020 Mar;10:5157. Available from: https://doi.org/10.1038/s41598-020-62015-w.

18. Park JY, Yoon S, Park MS, Choi H, Bae JH, Moon DG, et al. Development and External Validation of the Korean Prostate Cancer Risk Calculator for High-Grade Prostate Cancer: Comparison with Two Western Risk Calculators in an Asian Cohort. PLOS ONE. 2017 Jan;12(1):0168917. Available from: https://doi.org/10.1371/journal.pone.0168917.

19. Yoo S, Gujrathi I, Haider MA, Khalvati F. Prostate Cancer Detection using Deep Convolutional Neural Networks. Sci Rep. 2019 Dec;9:19518. Available from: https://doi.org/10.1038/s41598-019-55972-4.

20. Chen Y, Xu C, Zhang Z, Zhu A, Xu X, Pan J, et al. Prostate cancer identification via photoacoustic spectroscopy and machine learning. Photoacoustics. 2021 Sep;23:100280. Available from: https://doi.org/10.1016/j.pacs.2021.100280.