A APLICAÇÃO DE SVM NA PREVISÃO DA RECORRÊNCIA DO CARCINOMA DIFERENCIADO DA TIREOIDE (CDT)

Caio Eduardo Monteiro de Souza[1]

Marcelo José Martins Pereira[2]

Maria Luisa Cervi Uzun[3]

RESUMO

O avanço tecnológico tem transformado diversas áreas sociais, incluindo a saúde e a prevenção de doenças. Este estudo explora o papel das Máquinas de Suporte de Vetor (SVM) na estratificação de risco e personalização terapêutica em neoplasias tireoidianas, destacando sua relevância no cenário clínico atual. O carcinoma diferenciado da tireoide (CDT) apresenta prognóstico favorável na maioria dos casos, mas abordagens preditivas são essenciais para casos complexos, como o carcinoma anaplásico da tireoide (ATC), de comportamento agressivo e alta mortalidade. Máquinas de Suporte de Vetor emergem como ferramentas robustas na medicina, oferecendo capacidade de análise precisa através da construção de hiperplanos em espaços multidimensionais e do uso de *kernels* para dados não linearmente separáveis. Seu uso, aliado a *frameworks* como o *Scikit-Learn*, permite ajustes de hiperparâmetros que equilibram precisão e generalização, tornando-as adequadas a dados clínicos heterogêneos. A crescente disponibilidade de bases de dados clínicos de longo prazo, como o *Differentiated Thyroid Cancer Recurrence*, potencializa a aplicação de SVM na identificação de padrões ocultos e redução de subjetividades inerentes às abordagens tradicionais. Este trabalho contribui para o debate sobre a incorporação da inteligência artificial na oncologia, integrando avanços computacionais às necessidades práticas da medicina contemporânea.

Palavras-chave: Inteligência Artificial. Estratificação de Risco. Máquinas de Suporte de Vetor. Medicina Personalizada. Neoplasias Tireoidianas.

ABSTRACT

Technological advancements have transformed various social areas, including health and disease prevention. This study explores the role of Support Vector Machines (SVMs) in risk stratification and therapeutic personalization for thyroid neoplasms, highlighting their relevance in current clinical scenario. Differentiated thyroid carcinoma (DTC) generally has a favorable prognosis; however, predictive approaches are essential for complex cases, such as anaplastic thyroid carcinoma (ATC), characterized by aggressive behavior and high mortality rates. Support Vector Machines emerge as robust tools in medicine, offering precise analysis through hyperplane construction in multidimensional spaces and the use of kernels for nonlinearly separable data. Their integration with frameworks like Scikit-Learn allows for hyperparameter tuning that balances accuracy and generalization, making

[1] Graduando em Análise e Desenvolvimento de Sistemas pela Fatec Dr Thomaz Novelino - Franca/SP. Endereço eletrônico: caioeduardo1403@gmail.com

[2] Graduando em Análise e Desenvolvimento de Sistemas pela Fatec Dr Thomaz Novelino - Franca/SP. Endereço eletrônico: marjopereira@gmail.com

[3] Professora Associada da Fatec Dr Thomaz Novelino - Franca/SP. Endereço eletrônico: maria.uzun@fatec.sp.gov.br.

them suitable for heterogeneous clinical datasets. The increasing availability of long-term clinical databases, such as Differentiated Thyroid Cancer Recurrence, enhances the application of SVMs by identifying hidden patterns and reducing the subjectivity of traditional approaches. This study contributes to the discussion on the incorporation of artificial intelligence in oncology, aligning computational advances with the practical needs of contemporary medicine.

Keywords: *Artificial Intelligence. Personalized Medicine. Risk Stratification. Support Vector Machines. Thyroid Neoplasms.*

1 INTRODUÇÃO

O uso crescente da tecnologia permeia diversos aspectos sociais, este artigo abordará nas áreas como saúde e prevenção de doenças. Sendo assim, o estudo tem como objetivo de como Máquinas de Suporte de Vetor (SVM) podem contribuir para a estratificação de risco e personalização terapêutica em neoplasias tireoidianas ganha relevância frente ao cenário clínico atual.

Embora o carcinoma diferenciado da tireoide (CDT) represente a maioria dos casos com prognóstico favorável, às diretrizes de 2021 da *American Thyroid Association* destacam a importância de abordagens preditivas para casos complexos, como o carcinoma anaplásico da tireoide (ATC), que apresenta comportamento biológico agressivo e mortalidade elevada (BIBLE *et al.*, 2021). Tais avanços reforçam a necessidade de modelos computacionais precisos para orientar vigilância e intervenções em subtipos tumorais heterogêneos, alinhando-se às recomendações contemporâneas de medicina personalizada (BIBLE *et al.*, 2021).

As Máquinas de Vetores de Suporte (SVM) destacam-se na medicina como ferramentas eficazes para estratificação de risco e apoio a decisões clínicas, graças à sua capacidade de construir hiperplanos ótimos em espaços multidimensionais, maximizando a margem entre classes (GERON, 2022). A aplicação de *kernels* — abordada em *Hands-On Machine Learning* — permite tratar dados não linearmente separáveis, expandindo seu uso em diagnósticos complexos e predição de desfechos (GERON, 2022). Combinadas com *frameworks* como *Scikit-Learn*, as SVM permitem ajustes precisos de hiperparâmetros (como *C* e *gamma*), equilibrando precisão e generalização, o que as torna adaptáveis a conjuntos de dados clínicos heterogêneos e dinâmicos, aliando robustez técnica à praticidade em ambientes médicos (GERON

2022).

Além disso, a crescente disponibilidade de bases de dados clínicos de longo prazo, como o *Differentiated Thyroid Cancer Recurrence* (BORZOOEI; TAROKHIAN, 2023), viabiliza a construção de modelos preditivos baseados em evidências. A integração de variáveis como idade, estágio tumoral e marcadores histológicos permite não apenas identificar padrões ocultos, mas também reduzir a subjetividade inerente a abordagens tradicionais. Este trabalho, portanto, busca não apenas explorar o potencial técnico das SVM, mas também contribuir para a discussão sobre a incorporação de inteligência artificial em oncologia, alinhando avanços computacionais às demandas práticas da medicina contemporânea (BORZOOEI; TAROKHIAN, 2023).

2 REFERENCIAL TEÓRICO E TRABALHO CORRELATOS

O carcinoma diferenciado da tireoide (CDT) representa a forma mais comum de câncer tireoidiano, englobando os carcinomas papilífero e folicular. Estes tumores, que derivam das células foliculares da glândula, correspondem à maioria das neoplasias malignas da tireoide, com comportamento clínico heterogêneo. As diretrizes de 2021 da American Thyroid Association destacam que, embora a maioria dos casos de CDT tenha prognóstico favorável, uma subpopulação exibe características de alto risco, como invasão vascular ou metastática, demandando estratificação precisa para orientar condutas terapêuticas (BIBLE *et al.*, 2021).

De acordo com as recomendações atualizadas, a incidência global de CDT continua a aumentar, impulsionada por avanços diagnósticos e vigilância aprimorada. A ênfase contemporânea recai sobre a integração de marcadores moleculares (como mutações em *BRAF* ou rearranjos de *RET/PTC*) para personalizar abordagens, reduzindo intervenções excessivas em tumores indolentes e identificando casos que exigem terapia adjuvante agressiva (BIBLE *et al.*, 2021).

Os carcinomas papilífero e folicular mantêm características histológicas distintas, mas o manejo atual prioriza a avaliação de risco dinâmico, alinhando-se a critérios como tamanho tumoral, invasão extratireoidiana e status de margens cirúrgicas. Em contraste, o carcinoma anaplásico da tireoide (ATC), descrito nas diretrizes como uma entidade clinicamente devastadora, surge frequentemente de transformação de CDT pré-existente, exibindo perda de diferenciação celular,

crescimento acelerado e resistência terapêutica. Sua abordagem requer intervenção multimodal imediata, incluindo cirurgia, radioterapia e terapias-alvo, conforme protocolos estabelecidos (BIBLE *et al.*, 2021).

O estudo atualizado dos carcinomas tireoidianos reforça a importância de modelos preditivos e tecnologias emergentes para otimizar a vigilância pós-tratamento. As diretrizes de 2021 enfatizam a necessidade de colaboração multidisciplinar, integrando patologia molecular, imagem avançada e inteligência artificial, a fim de enfrentar desafios como recorrência local, progressão metastática e resistência a terapêuticas convencionais (BIBLE *et al.*, 2021).

A complexidade do comportamento dos diversos tipos de CDT, com sua variabilidade entre casos de baixo risco e evoluções agressivas, mostra a necessidade de métodos para prever recorrências e orientar decisões médicas. Nesse contexto, as Máquinas de suporte de vetor (MSV) são uma opção para auxiliar em decisões clínicas.

As SVM são de acordo com IBM (2023) são em tradução livre: Máquinas de suporte de vetor (SVM) é um algoritmo de *machine learning* supervisionado que classifica dados encontrando a linha ou hiperplano mais otimizado que maximiza a distância entre cada classe em um espaço N-dimensional.

Para realizar o estudo do comportamento do modelo *SVM* na previsão de recorrência do câncer diferenciado da tireoide (CDT), é fundamental utilizar uma base de dados robusta e confiável, composta por informações clínicas de pacientes em acompanhamento médico prolongado. Nesse contexto, optou-se pelo emprego do conjunto de dados *Differentiated Thyroid Cancer Recurrence*, disponibilizado pelo *UC Irvine Machine Learning Repository* (UCI Machine Learning Repository). De acordo com os autores do repositório, em tradução livre: “Este conjunto de dados contém 13 características clinicopatológicas com o objetivo de prever a recorrência do câncer de tireoide bem diferenciado. Os dados foram coletados ao longo de 15 anos, com cada paciente acompanhado por um período mínimo de 10 anos” (BORZOOEI; TAROKHIAN, 2023). O *dataset* inclui informações detalhadas sobre parâmetros como idade, estágio tumoral, marcadores histológicos e tratamento pós-operatório, tornando-o relevante para análises preditivas em oncologia.

O dado tem origem do artigo “*Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study* “em tradução livre: Aprendizado de máquina para estratificação de risco de pacientes com câncer de tireoide: um estudo de corte

de 15 anos, o estudo também cita a eficiência do uso de SVM's no campo médico, em tradução livre: Os SVMs são eficazes na seleção de características e redução de dimensionalidade (BORZOOEI *et al.*, 2023).

3 METODOLOGIA DE EXECUÇÃO

A metodologia estruturada adotada neste trabalho, que inclui etapas como pré-processamento de dados, seleção criteriosa de algoritmos e validação robusta, é fundamental para garantir a confiabilidade e reprodutibilidade dos resultados em projetos de machine learning. Conforme destacado por James *et al*. (2023) em *An Introduction to Statistical Learning* (2ª edição, corrigida em 2023), uma abordagem sistemática prioriza aplicações práticas e integra laboratórios com implementação em R, o que não apenas minimizar vieses e *overfitting*, mas também facilita a comparação objetiva de técnicas. Os autores enfatizam a importância de atualizações metodológicas, como a inclusão de métodos para análise de sobrevivência e testes múltiplos, bem como a revisão de códigos para alinhamento com versões recentes do R, garantindo que as decisões técnicas estejam ancoradas em práticas reprodutíveis e adaptadas às particularidades dos dados (JAMES *et al*., 2023).

A primeira fase consistiu em uma pesquisa bibliográfica exploratória, que visa oferecer uma compreensão profunda sobre o tema. Esse levantamento bibliográfico é essencial para identificar os avanços mais recentes e as abordagens metodológicas utilizadas em estudos similares, proporcionando uma base teórica sólida para as decisões que serão tomadas ao longo do desenvolvimento do trabalho.

A seguir, será realizada uma análise das diferentes variações do algoritmo SVM, com o objetivo de entender as diferenças em sua aplicabilidade. A partir dessa análise, será construída uma tabela que destaca os principais focos de utilização de cada variação, permitindo uma comparação clara das opções disponíveis. Essa etapa proporcionará uma visão detalhada sobre as características e limitações de cada variação, o que facilitará a escolha das mais adequadas para o problema específico da previsão da recorrência do CDT.

Com as variações do SVM identificadas e analisadas, será possível selecionar aquelas que serão aplicadas no desenvolvimento do modelo. Essa escolha será feita com base nas características dos dados disponíveis e nos objetivos do projeto, buscando garantir que a variação escolhida ofereça a melhor performance para a

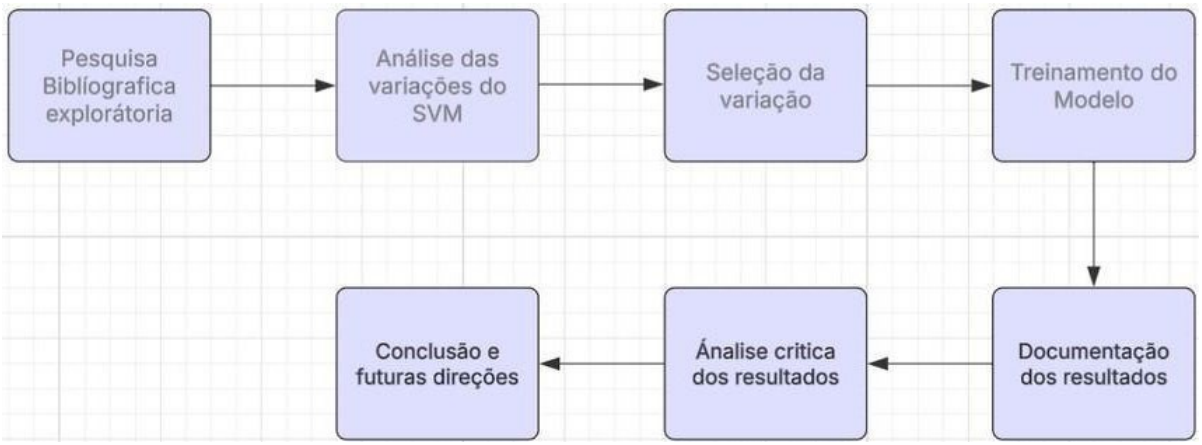
tarefa de previsão da recorrência. Após essa definição, o modelo será preparado e alimentado com os dados sobre o tema, que serão cuidadosamente extraídos de fontes relevantes e processados para garantir a qualidade e a adequação das informações utilizadas.

Uma vez alimentado, o modelo será analisado e seus resultados serão documentados conforme os objetivos definidos. A documentação incluirá gráficos, tabelas e uma descrição detalhada das previsões feitas, permitindo uma interpretação clara dos resultados. A etapa seguinte será a análise crítica desses resultados, onde se verificará se eles correspondem com as expectativas iniciais e com os objetivos do projeto. Essa análise permitirá identificar o desempenho do modelo, possíveis áreas de melhoria e os pontos fortes da abordagem escolhida.

Por fim, será redigida a conclusão do trabalho, que abordará os principais achados e comparar os resultados obtidos com as previsões feitas no início do estudo. A conclusão também discutirá as limitações do trabalho e sugerirá possíveis direções para pesquisas futuras. Além disso, refletirá sobre a aplicabilidade do algoritmo SVM na área de saúde, especificamente no contexto da previsão de recorrência do CDT. Esse processo metodológico busca garantir a construção de um trabalho acadêmico rigoroso e bem fundamentado, com contribuições relevantes para a área de estudo.

Para facilitar o entendimento do processo, o fluxograma da Figura 1 foi desenvolvido, exemplificando o passo a passo a ser executado neste estudo.

Figura 1 - Fluxograma metodológico



Fonte: autores

3.1 Análise Comparativa das variações de SVM

Conforme detalhado na metodologia, propõe-se a realização de uma análise comparativa dos algoritmos de SVM, visando identificar em quais situações cada método pode contribuir para a previsão da recorrência do CDT. Para tanto, será elaborada uma tabela comparativa com base no conteúdo analisado, a fim de selecionar o método mais adequado aos objetivos deste projeto.

Os critérios estabelecidos para a comparação dos algoritmos incluem:

- Base matemática: fundamentação teórica e princípios matemáticos que sustentam cada algoritmo;
- *Kernel*: tipos de funções de *kernel* utilizadas e sua influência no desempenho do modelo;
- Hiperparâmetros e complexidade computacional: impacto dos hiperparâmetros na eficiência e no tempo de processamento;
- Aplicação ao CDT: relevância e eficácia de cada método para a previsão da recorrência do CDT.

Essa abordagem sistemática permitirá uma avaliação embasada e criteriosa, facilitando a escolha do algoritmo mais adequado ao contexto do projeto e contribuindo para a precisão na previsão da recorrência do CDT.

Tabela 1 - Comparação dos diferentes algoritmos de SVM (continuação)

Tipos de SVM	Base Matemática	Kernel	Hiperparâmetros e complexidade computacional	Aplicação ao CDT (Previsão de Recorrência)
C-SVM	Minimização de risco estrutural com função de perda hinge. Formulação primal/dual com parâmetro C para regularização (CHANG; LIN, 2022).	Linear, RBF, Polinomial, Sigmoid.	Hiperparâmetros:CC, γγ, grau (polinomial), coef0 (sigmoid). Complexidade:O(n2)O(n2) a O(n3)O(n3).	O modelo é eficaz em dados balanceados, onde o parâmetro C, GERON, 2022) controla overfitting ao regular a tolerância a erros. Para dados desbalanceados — comuns em contextos clínicos —, ajustes como class_weight e reamostragem (SMOTE/undersampling, são necessários para evitar viés e manter a precisão em classes minoritárias, aliando métricas como F1-score para avaliação robusta (GERON, 2022).

Tabela 1 - Comparação dos diferentes algoritmos de SVM (conclusão)

Tipos de SVM	Base Matemática	Kernel	Hiperparâmetros e complexidade computacional	Aplicação ao CDT (Previsão de Recorrência)
v-SVM	Formulação alternativa com parâmetro v(0,1), que controla a fração de vetores de suporte e margens de erro (MOHRI; ROSTAMIZADEH; TALWALKAR, 2022)	Linear, RBF, Polinomial.	Hiperparâmetros: vv, γγ, grau. Complexidade: Similar à C-SVM.	Útil para controlar a proporção de exemplos na margem. Adequado para dados com ruído ou sobreposição(JAMES *et al*., 2023)
LS-SVM	Substitui a perda *hinge* por mínimos quadrados, resolvendo um sistema linear (CHANG; LIN, 2022)	Linear, RBF, polinomial	Hiperparâmetros:γγ (regularização), γkernelγkernel. Complexidade:O(n3)O(n3), mas eficiente para pequenos conjuntos.	Indicado para dados com ruído. Menos sensível a outliers, mas pode ser menos robusto em alta dimensionalidade (CHANG; LIN, 2022)
One-Class-SVM	Aprende uma fronteira para encapsular os dados, minimizando a região de alta densidade (CHANG; LIN, 2022)	Principalmente RBF.	Hiperparâmetros:vv, γγ. Complexidade: Similar a C-SVM.	Relevante se a recorrência for rara (detecção de anomalias). Requer ajuste cuidadoso de vv(CHANG; LIN, 2022)

Fonte: autores

3.2 Escolha do modelo SVM

A seleção do algoritmo *C-SVM (C-Support Vector Classification)* para o estudo em questão demandou uma avaliação detalhada das variantes de SVM, visando identificar a mais adequada ao contexto clínico. Conforme abordado em *Hands-On Machine Learning* (GERON, 2022), o *C-SVM* destaca-se por sua eficácia em problemas de classificação binária, sendo amplamente utilizado devido ao parâmetro de regularização *C*. Esse hiperparâmetro permite ajustar o equilíbrio entre a maximização da margem de separação dos dados e a tolerância a classificações incorretas (GERON, 2022)

A exclusão de outras variantes, como o *One-Class SVM*, justifica-se por sua aplicação primária em detecção de anomalias em dados não rotulados (SCIKIT-LEARN, 2023), incompatível com cenários supervisionados. Já o *LS-SVM (Least Squares SVM)*, embora útil em regressão, apresenta limitações críticas em contextos

médicos: sua função de perda quadrática, conforme discutido por (GERON, 2022), reduz a robustez a *outliers,* comuns em dados clínicos, e impõe pressupostos restritivos sobre a distribuição dos resíduos, comprometendo a generalização do modelo.

Assim, a escolha do *C-SVM* alinha-se à necessidade de modelos interpretáveis e adaptáveis, capazes de lidar com a variabilidade inerente a dados médicos, enquanto mantém o controle sobre o *overfitting* por meio do ajuste de *C* (GERON, 2022). Essa abordagem assegura um equilíbrio técnico entre precisão e flexibilidade, essencial para aplicações críticas na área da saúde.

Por fim, o *v*-SVM (nu-SVM), apesar de permitir ajustes na proporção de vetores de suporte via parâmetro *v*, carece da intuitividade prática associada ao parâmetro C do C-SVM. Adicionalmente, o C-SVM consolida-se como referência consolidada em *frameworks* de *machine learning* (SCIKIT-LEARN, 2023), com ampla documentação e aplicações validadas em classificação binária.

3.3 Pré-processamento dos dados

Apesar da qualidade dos dados, etapas de pré-processamento como: normalização, codificação de variáveis categóricas e estratificação na divisão treino-teste são essenciais para assegurar a integridade e eficácia do modelo de *machine learning*. Conforme destacado em *Hands-On Machine Learning* (GERON, 2022), técnicas como padronização (*StandardScaler)* e normalização (*MinMaxScaler)* equilibram a escala dos atributos, facilitando a convergência do algoritmo e reduzindo a sensibilidade a *outliers*. A estratificação, por sua vez, preserva a distribuição proporcional das classes no conjunto de treino e teste, evitando viés na avaliação de desempenho (GERON, 2022).

Essas práticas são críticas para mitigar riscos como *overfitting* e vazamento de dados (*data leakage*), comuns quando transformações são aplicadas antes da divisão dos dados (GERON, 2022). Por exemplo, o uso de *pipelines* no *Scikit-Learn* garante que etapas como imputação de valores faltantes ou escalonamento sejam realizadas apenas no conjunto de treino, preservando a generalização do modelo para novos casos clínicos. Além disso, a codificação adequada de variáveis categóricas (como *OneHotEncoder)* evita distorções em algoritmos sensíveis a ordens arbitrárias,

assegurando resultados clinicamente confiáveis (GERON, 2022).

Primeiramente a variável alvo foi separada do resto do conjunto de dados para diminuir as chances de um vazamento de dados durante o treinamento do modelo. Em seguida, os dados foram divididos em categóricos e numéricos, para que assim fosse possível tratar cada grupo adequadamente.

Nas variáveis numéricas, o método *StandardScaler* foi aplicado, *scikit-learn,* e nas categóricas foi utilizado o *OneHotEncoder,* também com o *scikit-learn.*

A partir disso, o conjunto de dados foi dividido em subgrupos de treinamento (70%) e teste (30%). A divisão garantiu uma boa proporção das classes em ambas as amostras, mantendo a distribuição original. Tal abordagem garante que o modelo seja avaliado de forma controlada, sem exposição prévia aos dados de teste, com isso reduzindo o risco de *overfitting*. A implementação utilizou bibliotecas do *scikit-learn* com fixação de *random state* com reprodução dos resultados.

3.4 Treinamento do modelo

O treinamento do modelo, envolvendo a busca sistemática de hiperparâmetros (como *C* e *gamma* em SVMs) e validação cruzada, é fundamental para otimizar o desempenho preditivo e garantir a generalização do algoritmo. Conforme detalhado em *Hands-On Machine Learning* (GERON, 2022), técnicas como *GridSearchCV*, permitem explorar combinações de hiperparâmetros de forma estruturada, equilibrando viés e variância conforme a teoria do *bias-variance tradeoff* (GERON, 2022).

A metodologia, exemplificada no livro com o uso de *pipelines* do *Scikit-Learn,* não só maximiza métricas como *F1-score* (relevante em problemas com classes desbalanceadas), mas também assegura robustez em cenários clínicos reais. A validação cruzada *k-fold* previne o *overfitting* ao avaliar o modelo em múltiplos subconjuntos dos dados, garantindo que a performance não dependa de uma divisão específica treino-teste (GERON, 2022). Além disso, a inclusão de etapas como o *HalvingGridSearchCV* otimiza recursos computacionais ao descartar combinações pouco promissoras em estágios iniciais, uma vantagem crítica em conjuntos de dados médicos de grande escala.

Essa abordagem, aliada à estratificação rigorosa dos dados, assegura que o modelo mantenha confiabilidade na predição de recorrência de condições como o

CDT, evitando vieses induzidos por distribuições assimétricas ou *data leakage* (GERON, 2022).

Com os dados prontos para utilização, foi implementada uma abordagem de busca em grade (*grid search*) por meio da biblioteca scikit-learn, especificamente utilizando o método *GridSearchCV*. Foram testados os hiperparâmetros C (com valores 0.1, 1 e 10) e Gamma (com valores 0.001, 0.01 e 0.1), totalizando 9 combinações possíveis (3 valores para C * 3 valores para Gamma). A validação cruzada foi configurada com 5 *folds* divididos para garantir robustez na avaliação, e o critério de seleção do melhor modelo foi baseado no f1_score médio calculado entre as divisões de treino e validação (Figura 2).

Figura 2 - Treinamento do modelo

```
param_grid = {
    'classifier__C': [0.1, 1, 10],
    'classifier__gamma': [0.001, 0.01, 0.1]
}


grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='f1')
grid_search.fit(X_train, y_train)
```

Fonte: autores

3.5 Validação do modelo

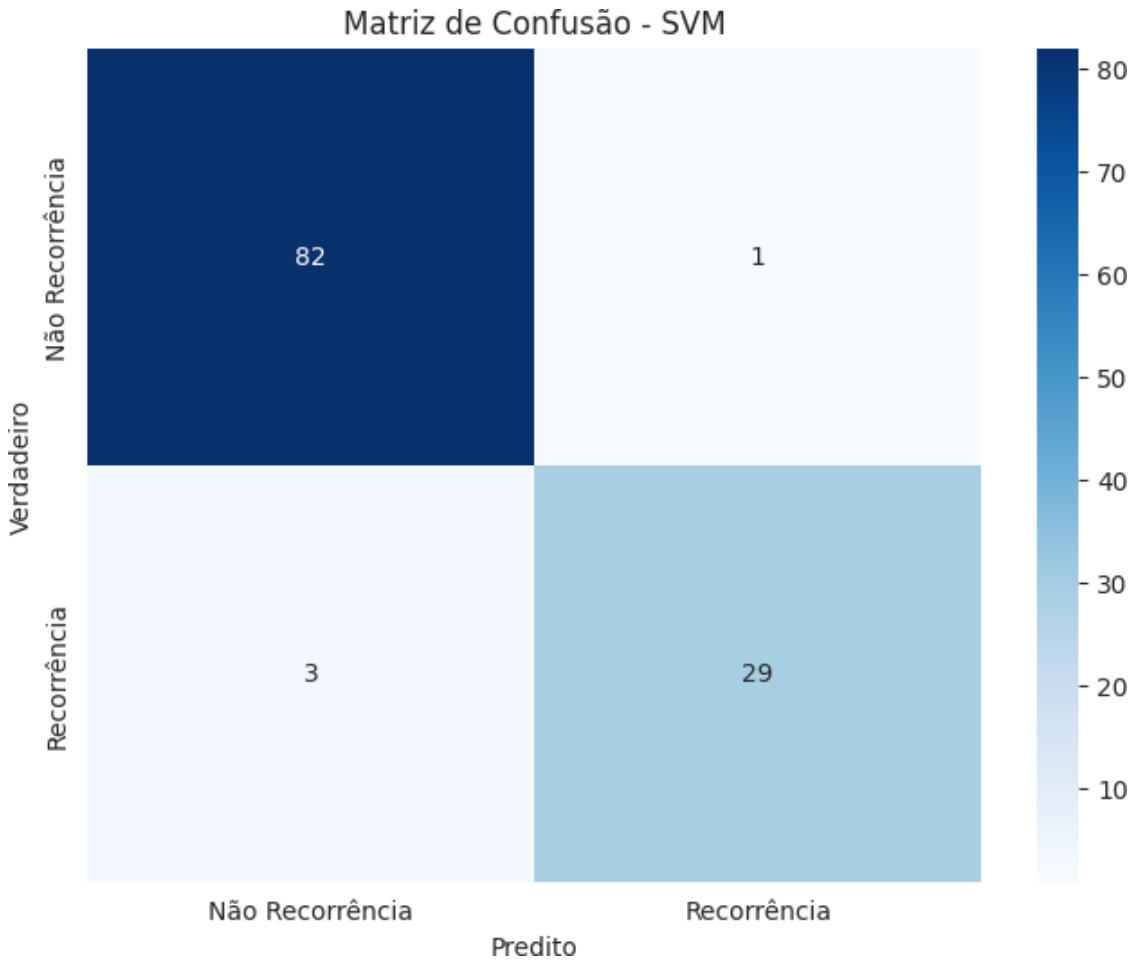
A validação do modelo, por meio de métricas como matriz de confusão, acurácia, *F1- score* e análise da curva ROC, é vital para assegurar a confiabilidade clínica das previsões e a generalização do algoritmo. Conforme destacado por James *et al*. (2023) em *An Introduction to Statistical Learning* (2ª edição corrigida), a análise rigorosa de desempenho, integrada a técnicas como validação cruzada, permite identificar *overfitting*, equilibrar sensibilidade e especificidade, e validar a capacidade preditiva em dados não vistos. A curva ROC, oferece uma visão clara do equilíbrio entre taxas de verdadeiros positivos e falsos positivos, sendo essencial para avaliar a discriminação do modelo.

4 RESULTADOS E DISCUSSÃO

Com o modelo treinado esta na hora de avaliar seus resultados e garantir que não ouve nenhum erro no procedimento, para isso foram utilizadas algumas métricas para a avaliação, a primeira foi a acurácia do modelo, que teve um ótimo resultado de 97%, podendo chegar a 81% nos piores casos. Já o *F1-score* em seu resultado macro teve um resultado de 0.98.

A figura 3 apresenta a matriz de confusão, onde as colunas representam as classes previstas e as linhas as classes verdadeiras, sendo uma matriz 2x2 com uma classificação binarias de recorrência ou não do CDT. A alta quantidade de valores na diagonal principal (82 e 29) mostram previsões corretas realizadas pelo modelo, e na diagonal secundaria (1 e 3) estão valores em que o modelo falhou em prever.

Figura 3 - Matriz de confusão

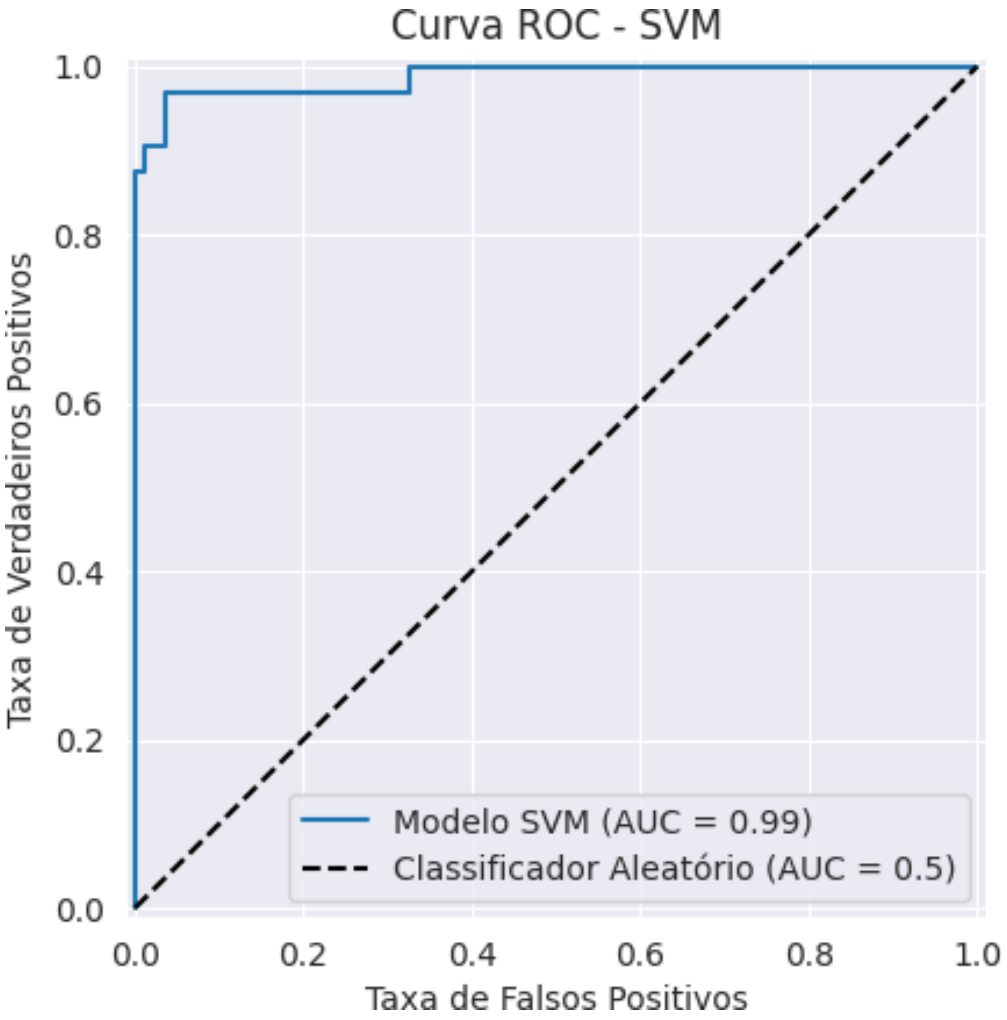


Fonte: autores

Já a Figura 4 ilustra a curva ROC, que avalia a capacidade do modelo de prever

os dados conforme o limiar de classificação. No eixo horizontal, a taxa de falsos positivos reflete casos erroneamente classificados como recorrência, enquanto no eixo vertical, a taxa de verdadeiros positivos mostra uma proporção de recorrências identificadas corretamente. A área sob a curva (AUC = 0,99), próxima ao valor máximo de 1,0, demonstra que o modelo consegue detectar casos positivos sem criar alarmes falsos. A proximidade da curva ao canto superior esquerdo, junto com a AUC elevada, indica um desempenho relevante do modelo nos dados apresentados.

Figura 4 - Curva ROC



Fonte: autores

Os resultados deste estudo mostram que os modelos de SVM, têm potencial para auxiliar na identificação da recorrência do CDT. Esses modelos demonstraram capacidade de analisar padrões complexos, conforme destacado no trabalho de (BORZOOEI; TAROKHIAN, 2023). No entanto, é importante lembrar que esses

algoritmos não substituem a avaliação médica, mas complementam o processo diagnóstico.

As SVM tiveram sua eficiência comprovada em estudos anteriores, sendo um deles o estudo de Cortes e Vapnik em 1995, onde demonstraram a capacidade desses modelos estatísticos de classificar os dados em diversos cenários. Em nosso estudo, o melhor modelo treinado apresentou uma taxa de acerto média de 97%, entretanto, nos piores cenários pode chegar até a 81%. Os resultados obtidos foram similares aos relatados por Borzooei e Tarokhian em seu estudo em 2023.Em resumo, as SVM são valiosas para triagem inicial e análise de dados complexos, mas sua aplicação deve ser combinada com avaliação profissional detalhada. Essa abordagem híbrida pode reduzir erros diagnósticos e acelerar o início de tratamentos, beneficiando tanto profissionais quanto pacientes.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos neste estudo destacam o potencial significativo das Máquinas de Suporte de Vetor (SVM) na identificação e análise de padrões relacionados à recorrência do carcinoma diferenciado da tireoide (CDT). Com uma acurácia média de 97% e um F1-score macro de 0,98, o modelo demonstrou alta eficiência, especialmente na correta classificação de casos conforme ilustrado pela matriz de confusão. A elevada AUC (0,99) da curva ROC reforça a capacidade do modelo de distinguir entre casos positivos e negativos de forma confiável, minimizando alarmes falsos e garantindo um desempenho robusto.

Embora os resultados validem a eficácia das SVM na estratificação de risco, é fundamental destacar que tais ferramentas devem ser utilizadas em conjunto com a análise profissional médica, complementando o processo diagnóstico. A combinação de técnicas avançadas de aprendizado de máquina com a expertise clínica pode otimizar a tomada de decisão e acelerar intervenções terapêuticas, alinhando-se às diretrizes da medicina personalizada. Dessa forma, este trabalho reafirma o papel indispensável da inteligência artificial na oncologia contemporânea, integrando precisão técnica às demandas práticas e proporcionando benefícios significativos para médicos e pacientes.

REFERÊNCIAS

BIBLE, K. C. *et al.* 2021 *American Thyroid Association Guidelines for Management of Patients with Anaplastic Thyroid Cancer.* ***thyroid****,* v. 31, n. 3, p. 337 – 386, 2021.

BORZOOEI, S. *et al. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study.* ***European Archives of Oto-Rhino-Laryngology***, v. 281, p. 2095 – 2104, outobro 2023. Acesso em: 6 mar. 2025.

BORZOOEI, S.; TAROKHIAN, A. ***Differentiated Thyroid Cancer Recurrence [dataset].*** 2023. Disponível em: https://doi.org/10.24432/C5632J. Acesso em: 4 de março de 2025.

CHANG, C.; LIN, C. LIBSVM: *A Library for Support Vector Machines.* ***ACM Transactions on Intelligent Systems and Technology***, v. 2, n. 3, p. 1 – 27, ago. 2022.

GERON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media*, v. 3, 2022.

IBM. ***What are support vector machines (SVMs)?*** 2023. Disponível em: https: //www:ibm:com/topics/support-vector-machine. Acesso em: 9 nov. 2024.

JAMES, G. *et al.* ***An Introduction to Statistical Learning****: with Applications in Python. 2. ed. [S.l.]: Springer*, 2023.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Kernel Methods in Modern Machine Learning: Theory and Practice.* ***Foundations and Trends® in Machine Learning***, v. 15, n. 1, p. 1 – 150, 2022.

SCIKIT-LEARN. ***Support Vector Machines.*** 2023. Disponível em: https://scikit-learn.org/stable/modules/svm.html. Acesso em: 20 fev. 2025.