Universidade de Brasília
Departamento de Estatística

Aplicação de Máquinas de Vetores de Suporte na Detecção de Câncer de Mama

Lucas Menezes e Silva

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Lucas Menezes e Silva

Aplicação de Máquinas de Vetores de Suporte na Detecção de Câncer de Mama

Orientador: Prof. Felipe Sousa Quintino

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Sumário

1 Introdução 10

2 Metodologia 12

2.1 Introdução ao Princípio do SVM 12

2.2 Maximização da Margem (*Hard-Margin* SVM) 14

2.2.1 A Formulação do Problema de Otimização 15

2.2.2 Multiplicadores de Lagrange 15

2.3 Relativizando as Margens (*Soft-Margin* SVM) 17

2.3.1 Introdução das Variáveis de Relaxação (ξ_i) 17

2.3.2 Função Objetivo Ajustada 18

2.3.3 Condições de Otimalidade KKT 18

2.3.4 Resolvendo a Lagrangiana para Obtenção da Forma Dual 19

2.3.5 Significado das Condições de Otimalidade 20

2.4 Funções *Kernel* para SVM 21

2.4.1 Tipos de *Kernels* 21

3 Resultados 23

3.1 Conjunto de Dados 23

3.1.1 Composição dos Dados 23

3.1.2 Variáveis Presentes 24

3.2 Preparação dos Dados 25

3.3 Construção e Avaliação Inicial dos Modelos SVM 25

3.3.1 Treinamento e Avaliação dos Modelos 25

3.3.2 Avaliação da Importância das Variáveis 32

3.3.3 Visualização do SVM - Variáveis Selecionadas 33

3.4 Aplicação do PCA e Avaliação dos Modelos 34

3.4.1 Visualização dos Dados sob Ótica do PCA 34

3.4.2 Aplicação de Modelo PCA e Análise dos Resultados 35

4 Conclusão 39

5 Apêndice 43

5.1 Banco de Dados 43

5.2 Códigos R 43

Lista de Tabelas

1 Matriz de confusão SVM com Kernel Linear 27
2 Matriz de confusão SVM com Kernel Polinomial 27
3 Indicadores de Desempenho dos Modelos 28
4 Resumo da Análise de Componentes Principais (PCA) - Parte 1 34
5 Resumo da Análise de Componentes Principais (PCA) - Parte 2 35
6 Matriz de Confusão – Modelo Linear SVM com PC1 e PC2 37
7 Matriz de Confusão – Modelo Polinomial SVM com PC1 e PC2 37
8 Indicadores de Desempenho dos Modelos de CPA 38

Lista de Figuras

1 Construção do Hiperplano SVM. 14
2 Malha de Valores - Modelo SVM com *Kernel* Linear 26
3 Malha de Valores - Modelo SVM com *Kernel* Radial 26
4 Malha de Valores - Modelo SVM com *Kernel* Sigmoidal 26
5 Malha de Valores - Modelo SVM com *Kernel* Polinomial 27
6 Curva ROC para o Modelo SVM com Kernel Linear. 29
7 Curva ROC para o Modelo SVM com Kernel Polinomial. 29
8 Curvas de Distribuição Acumulada Empírica para Cálculo do KS - Linear. 30
9 Curvas de Distribuição Acumulada Empírica para Cálculo do KS - Polinomial. 31
10 Importância Relativa das Variáveis no Modelo SVM com Kernel Polinomial. 32
11 Fronteiras de Decisão Geradas pelo SVM com Duas das Combinaçõs de Pares das 4 Variáveis Mais Explicativas. - Parte 1 33
12 Fronteiras de Decisão Geradas pelo SVM com Duas das Combinaçõs de Pares das 4 Variáveis Mais Explicativas. - Parte 2 33
13 Fronteiras de Decisão Geradas pelo SVM com Duas das Combinaçõs de Pares das 4 Variáveis Mais Explicativas. - Parte 3 34
14 Distribuição dos Dados no Espaço das Duas Primeiras Componentes Principais (PC1 e PC2). 35
15 Fronteira de Decisão do Modelo SVM com Kernel Linear no Espaço Bidimensional (PC1 e PC2). 36
16 Fronteira de Decisão do Modelo SVM com *Kernel* Polinomial no Espaço Bidimensional (PC1 e PC2). 36
17 Fronteira de Decisão do Modelo SVM com *Kernel* Polinomial e Linear no Espaço Bidimensional (PC1 e PC2). 37

Agradecimentos

Agradeço sinceramente a todos aqueles que contribuíram para a realização deste trabalho e que estiveram ao meu lado durante toda essa jornada acadêmica. Em especial, gostaria de expressar minha profunda gratidão à minha família, em especial ao meu pai Elias, minha mãe Ângela, minha irmã Laura e meus avós Gualter e Irene, que foram fundamentais em cada etapa desse processo. Estendo esse agradecimento também a todos os meus primos, tios e demais familiares, cujo carinho e apoio foram essenciais ao longo da caminhada.

Aos meus colegas de faculdade, que marcaram diferentes fases da minha graduação, meu muito obrigado. Da primeira parte da jornada, agradeço a João Pedro Moreira Pupe, João Victor Melo, Guilherme Silva e Caio Cavalcante. Vocês foram companheiros de estudo e amizade, sempre dispostos a ajudar e compartilhar experiências que tanto contribuíram para meu desenvolvimento acadêmico e pessoal.

Da segunda parte da graduação, sou grato a Daneiel Paranaguá, Davi Esmeraldo, Arthur Rodrigues e Gabriela Lobo, que estiveram presentes em momentos decisivos, tornando os desafios mais leves e o aprendizado mais enriquecedor.

Também agradeço aos amigos Caio Campanatti, Matheus Moussa, Arthur Dantas, Gabriel Raya, Arthur Tavares, Eduardo Schneider, Caio Granja e Lucas Caldeira, cuja presença foi indispensável para que a jornada se tornasse ainda mais especial, repleta de boas conversas, incentivo e companheirismo.

Aos meus amigos de vida, que, mesmo não mencionados individualmente aqui, saibam que cada gesto de amizade e apoio foi um verdadeiro alicerce para que eu pudesse seguir em frente.

Gostaria também de expressar minha gratidão a todos os professores que passaram pela minha formação, pelo ensino de qualidade e pela inspiração constante. Em especial, agradeço aos professores George von Borries e Maria Teresa Leão, cujo exemplo e dedicação marcaram profundamente minha trajetória acadêmica. Um agradecimento especial ao meu orientador, professor Felipe Quintino, pela orientação, paciência e confiança no desenvolvimento deste trabalho.

A todos os amigos, familiares e colegas que de alguma forma contribuíram, saibam que cada palavra de incentivo, apoio nos momentos difíceis, conversas inspiradoras e a confiança no meu potencial foram fundamentais para a conclusão desta etapa.

A todos vocês, meu mais sincero obrigado. Sem o apoio, a amizade e a colaboração de cada um, este trabalho não seria possível.

Resumo

Este trabalho tem como objetivo aplicar Máquinas de Vetores de Suporte (SVM) para a detecção de câncer de mama, utilizando o conjunto de dados *Breast Cancer Wisconsin (Original)*. A pesquisa integra conceitos matemáticos da técnica de SVM com a implementação computacional em R, explorando duas diferentes configurações de *kernel*. As análises foram realizadas considerando critérios técnicos para a escolha do modelo mais adequado. Além da avaliação do desempenho dos classificadores, foram aplicadas técnicas de redução de dimensionalidade e análise de importância das variáveis, a fim de compreender a separabilidade dos dados e os fatores mais relevantes para a classificação. Os resultados obtidos demonstram o potencial da abordagem proposta como ferramenta de apoio ao diagnóstico médico.

Palavras-chave: Análise de Dados, Câncer de Mama, Classificação, Diagnóstico Médico Assistido, Máquinas de Vetores de Suporte.

Abstract

This work aims to apply Support Vector Machines (SVM) for breast cancer detection using the *Breast Cancer Wisconsin (Original)* dataset. The research integrates mathematical concepts of the SVM technique with computational implementation in R, exploring two different *kernel* configurations. The analyses were conducted considering both technical criteria to select the most appropriate model. In addition to evaluating the classifiers' performance, dimensionality reduction techniques and variable importance analyses were applied to understand data separability and identify the most relevant factors for classification. The results demonstrate the potential of the proposed approach as a tool to support medical diagnosis.

Keywords: Breast Cancer, Classification, Computer-Aided Medical Diagnosis, Data Analysis, Support Vector Machines.

1 Introdução

O cǎncer de mama é uma das principais causas de morte entre mulheres em todo o mundo, sendo o segundo cǎncer mais comum e o quinto que mais mata pessoas no mundo de acordo com os dados da Organização Pan-Americana da Saúde, escritório regional da Organização Mundial da Saúde, (OPAS-OMS) em 2018[1], configurando-se como um problema de saúde pública que exige atenção contínua. Segundo o Instituto Nacional de Cǎncer (INCA), apenas no Brasil, em 2024, foram estimados mais de 73600 novos casos, tornando-se o tipo de cǎncer mais frequente entre as mulheres. A detecção precoce desempenha um papel essencial na redução da mortalidade associada a essa doença, permitindo tratamentos menos invasivos e maior probabilidade de remissão (TOMAZELLI et al., 2017).

Apesar da disponibilidade de métodos tradicionais como o autoexame e a mamografia, hǎ limitações significativas no diagnóstico, incluindo altas taxas de falsos-positivos que levam a biópsias desnecessárias, sobrecarregando o sistema de saúde e causando impacto emocional nos pacientes (OLIVEIRA, 2021). Nesse contexto, tecnologias emergentes, como os Sistemas de Diagnóstico Auxiliado por Computador (CAD), vêm sendo desenvolvidas para apoiar os especialistas no diagnóstico de nódulos mamários (AZEVEDO-MARQUES, 2001).

Os métodos de aprendizado de mǎquina vêm sendo estudados desde meados do século XX, onde foram impulsionados pelo avanço da matemática aliado ao avanço computacional. Entre os métodos mais estudados estão as árvores de decisão, redes neurais artificiais (RNA), mǎquinas de vetor de suporte (SVM), *K-Nearest Neighbors* (KNN) e algoritmos baseados em *ensemble*, como *Random Forests* e *Gradient Boosting*. Intrinsecamente existe a necessidade de falar sobre vantagens e desvantagens da utilização dos métodos. Condições como interpretabilidade das técnicas, nível de informação e grandes volumes de dados podem ser um problema. Por outro lado, a capacidade de previsibilidade, independemente do contexto, e a capacidade de lidar com grandes volumes de dados, são pontos muito válidos para a utilização desse tipo de método. Para um estudo mais aprofundado, contendo a abordagem teórica detalhada, assim como métodos mais complexos, recomenda-se o livro (BISHOP; NASRABADI, 2006) que possui desde conceitos básicos até os mais avançados em sua obra.

Entre as abordagens de inteligência artificial, destaca-se o uso de algoritmos de Aprendizado de Mǎquina, que têm mostrado resultados promissores na classificação de nódulos em malignos e benignos. Dentre as técnicas disponíveis, o SVM têm se destacado por sua eficácia na análise de dados multidimensionais, sendo amplamente aplicadas em

[1]disponível em: ⟨https://www.paho.org/pt/topicos/cancer⟩, acesso em 07/05/2025.

problemas de classificação médica (CRUZ; CRUZ; SANTOS, 2018). A versatilidade do SVM se expande por outros trabalhos como o de (MALATHI et al., 2022), que exploraram identificar emoções a partir de eletroencefalograma. Além disso, outro trabalho muito interessante sobre o uso de SVM é o de (DONG et al., 2015), em que utilizam a metodologia para detectar drogas no corpo humano a partir de compostos presentes na urina.

O SVM é uma técnica poderosa de aprendizado supervisionado, originalmente desenvolvida para resolver problemas de classificação binária. A ideia central do SVM é encontrar um *hiperplano de separação ótima*, que maximize a margem entre os exemplos de duas classes. Isso significa que o modelo busca a melhor divisão possível no espaço de características, minimizando o risco de erro de classificação em novos dados (MELLO; PONTI, 2018).

Diferentemente de outros estudos que combinam diferentes métodos, este trabalho se concentra exclusivamente na aplicação de SVM para a detecção de câncer de mama. O objetivo é avaliar o desempenho dessa técnica na análise de um conjunto de dados clínicos, verificando sua precisão e potencial para auxiliar no diagnóstico precoce. O estudo se fundamenta em um banco de dados amplamente utilizado na literatura. Dessa forma, espera-se contribuir para o desenvolvimento de ferramentas computacionais que possam ser integradas à prática médica, reduzindo os custos e melhorando a assertividade no diagnóstico do câncer de mama.

Por fim, o trabalho será dividido em 4 Seções. Na Seção 2, a metodologia do SVM é descrita em detalhes. Na Seção 3, os dados são descritos e analisados. Por fim, na Seção 4 temos a conclusão dos resultados obtidos.

2 Metodologia

2.1 Introdução ao Princípio do SVM

Nessa subseção e nas seguintes, vamos descrever os elementos matemáticos necessários para o método dos algoritmos SVM para classificação. As referências principais que foram utilizadas são (MELLO; PONTI, 2018) e (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Fixaremos os seguintes conceitos/termos:

- ***Domain set***: Um conjunto arbitrário $\mathcal{X}$.
- ***Label set***: conjunto de possíveis rótulos $\mathcal{Y} = \{-1, 1\}$.
- ***Training data***: $\mathrm{o}_n = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ é uma sequência finita de pares em $\mathbb{O} = \mathcal{X} \times \mathcal{Y}$.
- ***The learne's output***: Precisamos gerar uma regra de previsão $h : \mathcal{X} \to \mathcal{Y}$ capaz de prever o rótulo futuro do processo $y \in \mathcal{Y}$ baseado nos dados de entrada $\boldsymbol{x} \in \mathcal{X}$.
  - Esta função também é chamada *predictor*, *hypothesis* ou um *classifier*;
- **Classe dos preditores adimissíveis**:

$$\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y};\ h \text{ é Borel mensurável } ^2\ \}.$$

No contexto de aprendizado estatístico, é necessário que a função de perda seja borel mensurável para garantir a compatibilidade com a estrutura de medida.

Em problemas de aprendizado estatístico, utilizamos algoritmos de *Machine Learning* para, com base numa amostra de treinamento não vazia, determinar "o melhor" preditor $h : \mathcal{X} \to \mathcal{Y}$.

Para explicar melhor esse conceito do melhor preditor possível, na Teoria do Aprendizado Estatístico, considera-se que existe uma distribuição desconhecida $P(X, Y)$ sobre o espaço de entrada X e saída Y. O objetivo é encontrar uma função $h : X \to Y$ que minimize o *risco esperado*:

$$\mathcal{R}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[L(h(\boldsymbol{x}), y)\right]$$

[2] Uma função $f : X \to \mathbb{R}$ é dita *Borel mensurável* se, para todo conjunto de Borel $B \subseteq \mathbb{R}$, a pré-imagem $f^{-1}(B)$ pertence ă -ălgebra de Borel de X. Definição de (BILLINGSLEY, 1995)

onde L é uma função de perda que quantifica o erro entre a predição $h(X)$ e o valor real Y.

Como a distribuição P é desconhecida, utiliza-se uma amostra de treinamento $\{(x_i, y_i)\}_{i=1}^{n}$ para estimar o risco esperado por meio do *risco empírico*:

$$\mathcal{R}_{\text{emp}}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(\boldsymbol{x}_i), y_i)$$

O algoritmo de aprendizado busca então o preditor h que minimiza esse risco empírico dentro de um espaço de hipóteses $\mathcal{H}$:

$$h^* = \arg\min_{h\in\mathcal{H}} \mathcal{R}_{\text{emp}}(h)$$

Esse processo é conhecido como *Minimização do Risco Empírico* (*Empirical Risk Minimization* - ERM)[3] e é fundamental para o desenvolvimento de algoritmos de aprendizado de máquina.

A construção do hiperplano de separação no algoritmo SVM necessita definir determinados vetores que desempenham papel fundamental na delimitação da fronteira de decisão.

Considere x_+ é a média das amostras cuja categoria é $y_i = +1$:

$$\boldsymbol{x}_+ = \frac{1}{m_+}\sum_{i\in\{1,\ldots,n\}\,;\,y_i=+1} \boldsymbol{x}_i$$

em que m_+ é o número total de amostras com $y_i = +1$. Note que o somatório considera apenas os vetores $\boldsymbol{x}_i$ cuja classe é $+1$.

De maneira análoga, x_- é definido como a média das amostras cuja classe é $y_i = -1$:

$$\boldsymbol{x}_- = \frac{1}{m_-}\sum_{i\in\{1,\ldots,n\}\,;\,y_i=-1} \boldsymbol{x}_i$$

em que m_- é o número total de amostras com $y_i = -1$. Nesse caso, o somatório considera apenas os vetores $\boldsymbol{x}_i$ cuja classe é -1.

O vetor $\boldsymbol{w}$ define a orientação do hiperplano no SVM, sendo calculado como a diferença entre os dois 'centros de massa' das características.

$$\boldsymbol{w} = \boldsymbol{x}_+ - \boldsymbol{x}_-$$

[3] O processo deve ser consultado em (VAPNIK, 1995)

O vetor $\boldsymbol{w}$ representa a direção que maximiza a separação geométrica entre as categorias $+1$ e -1. O hiperplano separa de forma ortogonal ao vetor $\boldsymbol{w}$ e é posicionado para maximizar a margem, ou seja, a distância entre as duas categorias, a partir dos vetores de suporte de acordo com a Figura 1.

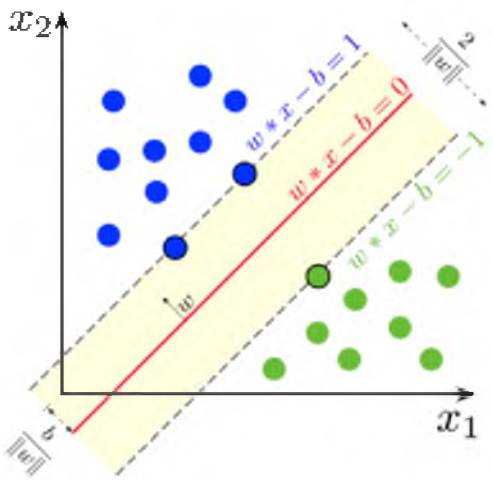


Figura 1: Construção do Hiperplano SVM.

2.2 Maximização da Margem (*Hard-Margin* SVM)

O SVM busca maximizar a margem geométrica $M = 1/\|\boldsymbol{w}\|$ visto que isso configura a maior área possível de separação entre os dados. Isso equivale a minimizar $\|\boldsymbol{w}\|$. Contudo, para simplificar o problema, o SVM utiliza a função objetivo $\frac{1}{2}\|\boldsymbol{w}\|^2$, pois ela é diferenciável e convexa. Aqui, $||\cdot||$ representa a norma (Euclidiana) do vetor.

Dessa forma, considere um conjunto de dados:

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}, \quad \boldsymbol{x}_i \in \mathbb{R}^d, \quad y_i \in \{-1, +1\}, \tag{2.2.1}$$

em que:

- $\boldsymbol{x}_i$ é o vetor de características do i-ésimo exemplo.
- y_i é o rótulo da categoria de $\boldsymbol{x}_i$ ($+1$ ou -1).

O objetivo do SVM é encontrar um hiperplano definido por:

$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0, \tag{2.2.2}$$

em que:

- $\boldsymbol{w}$ é o vetor que determina a orientação do hiperplano de divisão.

- $b \in \mathbb{R}$ é o termo que ajusta a posição do hiperplano.
- $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ representa o produto interno entre os vetores $\boldsymbol{w}$ e $\boldsymbol{x}$.

A margem geométrica M é definida como a distância entre o hiperplano e os vetores de suporte mais próximos de cada classe. Formalmente:

$$M = \frac{1}{\|\boldsymbol{w}\|}. \tag{2.2.3}$$

Os vetores de suporte são os exemplos de treinamento que estão mais próximos do hiperplano, ou seja, aqueles que satisfazem a condição de igualdade:

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) = 1. \tag{2.2.4}$$

2.2.1 A Formulação do Problema de Otimização

A formulação do problema de otimização consiste na minimização de

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{2.2.5}$$

sujeito a:

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad \forall i = 1, \ldots, n\,, \tag{2.2.6}$$

em que:

- $\boldsymbol{x}_i$ é um vetor de características de treinamento,
- $y_i \in \{-1, +1\}$ é o rótulo da classe do vetor $\boldsymbol{x}_i$.

A restrição garante que os pontos de cada classe estejam corretamente posicionados em relação ao hiperplano.

2.2.2 Multiplicadores de Lagrange

Os multiplicadores de Lagrange são uma ferramenta matemática para achar os máximos e mínimos no problema multivariável. Em d=3, podemos resumir o método dos multiplicadores de Lagrange à:

Suponha que existem funções $f(x, y, z)$ e $g(x, y, z)$ diferenciáveis e $\nabla g \neq 0$, em que ∇ é o gradiente, quando $g(x, y, z) = 0$. Para encontrar os valores máximo e mínimo locais de f sujeitos à restrição $g(x, y, z) = 0$, basta encontrar os valores de x, y, z e λ que

satisfazem simultaneamente as equações:

$$\begin{cases} \nabla f(x,y,z) = \lambda \nabla g(x,y,z), \\ g(x,y,z) = 0. \end{cases}$$

Chamamos o escalar λ de multiplicador de Lagrange. O escalar é essencial para transformar problemas com restrições em um problema sem restrições, conforme detalhado no Capítulo 14.8 do livro (THOMAS; WEIR; HASS, 2012).

Dentro dessa equalização matemática do SVM, as restrições sobre os valores de entrada são utilizadas no problema primal, que define a função de Lagrange como:

$$\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1 \right],$$

em que $\alpha_i \geq 0$ são os multiplicadores de Lagrange, que controlam o impacto das restrições.

Solução da Lagrangiana e Formulação Dual

A solução da Lagrangiana é encontrada ao derivar $\mathcal{L}(\boldsymbol{w}, b, \alpha)$ em relação às variáveis livres $\boldsymbol{w}$ e b:

1. Derivada em relação a $\boldsymbol{w}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0 \quad \Longrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i.$$

   Isso mostra que $\boldsymbol{w}$ é uma combinação linear dos vetores $\boldsymbol{x}_i$, ponderados pelos multiplicadores α_i.

2. Derivada em relação a b:

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0.$$

   Essa equação garante que o hiperplano é equilibrado em relação às classes.

Após a formulação dual do problema de otimização, podemos concluir que a abordagem matemática do SVM se baseia em uma solução que equilibra o conceito de separação máxima entre classes e a tolerância a erros. O objetivo principal é encontrar os multiplicadores de Lagrange (α_i) que determinam os vetores de suporte, ou seja, os pontos dos dados mais relevantes para a definição do hiperplano de separação. Esses vetores de suporte são fundamentais porque, mesmo em grandes conjuntos de dados, apenas uma fração dos exemplos contribui para a solução final, o que torna o SVM eficiente. Além

da eficiência, o SVM é um modelo muito consistente pela capacidade dele não precisar de grandes volumes de dados para entender certos tipos de padrões. Dado isso, a condição

$$\sum_{i=1}^{n} \alpha_i y_i = 0,$$

garante o equilíbrio entre as classes, assegurando que a contribuição ponderada dos rótulos seja nula. Isso reflete a natureza do problema de classificação, em que o hiperplano precisa ser neutro em relação às classes. Já a restrição

$$0 \leq \alpha_i \leq C,$$

introduzida pelo parâmetro C, permite ajustar o *trade-off* entre a margem máxima e os erros de classificação permitidos, conferindo flexibilidade ao modelo para lidar com dados ruidosos.

2.3 Relativizando as Margens (*Soft-Margin* SVM)

Quando os dados não são perfeitamente (linearmente) separáveis, adicionamos variáveis *slack* ξ_i para lidar com violações nas restrições.

2.3.1 Introdução das Variáveis de Relaxação (ξ_i)

No problema de *Hard Margin*, considera-se todos os pontos como linearmente separáveis, o que nem sempre é realista ou possível. Para resolver isso, são adicionadas variáveis de relaxação/*slack* $\xi_i \geq 0$, que permitem que os pontos violem os limites da margem linear.

A nova restrição torna-se:

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n\,,$$

em que

- $\xi_i = 0$: O ponto está corretamente classificado e dentro da margem.
- $0 < \xi_i \leq 1$: O ponto está dentro da margem, mas classificado corretamente.
- $\xi_i > 1$: O ponto está mal classificado.

2.3.2 Função Objetivo Ajustada

A função objetivo, que inclui um termo de penalização para os ξ_i, controlado por um hiperparâmetro C, de custo, gera como formulação primal:

$$\min_{x,b,\xi} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \xi_i,$$

sujeito a:

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

em que $\frac{1}{2}\|\boldsymbol{w}\|^2$ maximiza a margem do SVM e $C\sum_{i=1}^{n} \xi_i$ penaliza violações da margem do SVM. Assim, ajusta o *trade-off* entre margem e classificações incorretas. Basicamente, cria um balanceamento dentro das condições da classificação.

2.3.3 Condições de Otimalidade KKT

As condições de Karush-Kuhn-Tucker (KKT) garantem que a solução do problema primal e dual seja *ótima*. As condições são:

1. **Estacionaridade**:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \implies \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \tag{2.3.1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \implies \sum_{i=1}^{n} \alpha_i y_i = 0, \tag{2.3.2}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \implies \alpha_i + \beta_i = C, \quad \forall i = 1, \ldots, n\ . \tag{2.3.3}$$

2. **Complementaridade**:

$$\alpha_i \left[y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1 + \xi_i \right] = 0, \tag{2.3.4}$$

$$\beta_i \xi_i = 0, \quad \forall i = 1, \ldots, n\ . \tag{2.3.5}$$

3. **Dualidade**:

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, n\ . \tag{2.3.6}$$

4. **Não negatividade**:

$$\alpha_i \geq 0, \quad \beta_i \geq 0, \quad \forall i = 1, \ldots, n\ . \tag{2.3.7}$$

Interpretação Matemática

- **Estacionaridade**: O primeiro ponto mostra que o vetor de pesos $\boldsymbol{w}$ é uma combinação linear dos vetores de entrada $\boldsymbol{x}_i$, balanceados pelos multiplicadores α_i e pelos rótulos das categorias y_i. Isso significa que apenas os vetores de suporte (aqueles com $\alpha_i > 0$) contribuem para a definição de $\boldsymbol{w}$. Basicamente, o algoritmo avalia o que ele quer considerar.

- **Complementaridade**: A condição $\alpha_i \left[y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1 + \xi_i\right] = 0$ garante que, se $\alpha_i > 0$, então o vetor $\boldsymbol{x}_i$ está exatamente sobre a margem ou no lado errado da margem permitido pela variável slack ξ_i. Caso contrário, $\alpha_i = 0$, e o ponto não contribui para a solução. Ou seja, o ponto saiu dos limites permitidos para que fosse considerado.

- **Dualidade**: A soma dos multiplicadores α_i ponderada pelos rótulos y_i deve ser zero, garantindo que a solução respeite o balanço entre as classes.

- **Não negatividade**: As condições $\alpha_i \geq 0$ e $\beta_i \geq 0$ garantem que as penalidades aplicadas às restrições sejam consistentes com o problema de otimização do SVM nas condições do *Soft Margin*.

2.3.4 Resolvendo a Lagrangiana para Obtenção da Forma Dual

Conforme os resultados obtidos pela propriedade de estacionariedade, foi aplicado os multiplicadores de Lagrange α_i e β_i para incorporar as restrições no problema de otimização. A função Lagrangiana é:

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \alpha, \beta) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1 + \xi_i\right] - \sum_{i=1}^{n} \beta_i \xi_i,$$

em que:

- $\alpha_i \geq 0$: Multiplicadores de Lagrange associados às restrições $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i$.
- $\beta_i \geq 0$: Multiplicadores de Lagrange associados às restrições $\xi_i \geq 0$.

A função de Lagrange incorpora as restrições do problema primal na forma de penalidades. Com isso, obtemos as condições de otimalidade que permitem construir a forma dual:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle,$$

sujeito a:

$$0 \leq \alpha_i \leq C, \forall i = 1, ..., n,$$

e

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

2.3.5 Significado das Condições de Otimalidade

As condições de otimalidade indicam que os multiplicadores de Lagrange α_i controlam a influência de cada ponto de dado na solução final. Especificamente:

- $\alpha_i = 0$: O ponto não contribui para o hiperplano final (não é um vetor de suporte).
- $0 < \alpha_i < C$: O ponto está na margem e é um vetor de suporte.
- $\alpha_i = C$: O ponto está mal classificado ou exatamente na margem com a máxima penalização permitida.

A restrição $\sum_{i=1}^{n} \alpha_i y_i = 0$ garante o equilíbrio entre as classes, enquanto $0 \leq \alpha_i \leq C$ define o limite superior para a influência de cada ponto, dado que ele avalia o máximo que o ponto pode avançar. Essas condições asseguram que o problema esteja bem definido e que a solução final maximize a margem enquanto permite uma certa flexibilidade para lidar com os possíveis erros de classificação. Pode-se dizer que as margens ficam mais frouxas e suscetíveis as formas das amostras dos dados.

Esse formalismo fornece a base matemática para entender como o SVM cria essa ponderação entre a margem e os erros de classificação que poderiam ser evitados, permitindo generalizações sólidas em dados reais.

Um modelo de SVM bem treinado apresenta as seguintes vantagens:

1. **Consistência à generalização**: Ao maximizar a margem, o SVM reduz a chance de *overfitting* (sobreajuste dos dados de treinamento).
2. **Eficiência computacional**: O SVM utiliza uma formulação dual que simplifica a solução do problema em espaços de alta dimensionalidade.
3. **Flexibilidade**: A introdução de *kernels* permite lidar com dados não linearmente separáveis e com padrões que podem ser difíceis de detectar de maneira simplista.

2.4 Funções *Kernel* para SVM

O *kernel* é uma função que transforma os dados de entrada para um espaço de características de maior dimensionalidade, permitindo que o SVM encontre um hiperplano de separação adequado, mesmo em casos onde os dados não são linearmente separáveis no espaço original.

Matematicamente, o *kernel* substitui o produto escalar $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ por $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, onde K é a função *kernel*.

Definição 1 *Considere um subespaço $\mathcal{X} \subseteq \mathbb{R}^d$. Dizemos que uma função $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ é um kernel se, para todo $n \in \mathbb{N}$ e $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, temos que*

$$\sum_{i,j} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \tag{2.4.1}$$

para todo $c_1, \cdots, c_n \in \mathbb{R}$. A igualdade em (2.4.1) ocorre quando $c_1 = \ldots = c_n = 0$. Dizemos que a matriz $(K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j}$ é positiva semidefinida.

2.4.1 Tipos de *Kernels*

1. ***Kernel* Linear**

   É o mais simples e indicado para dados que já são aproximadamente linearmente separáveis, sendo escrito dessa forma:

   $$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle.$$

   Neste caso, o SVM trabalha diretamente no espaço original e 'puro' dos dados.

2. ***Kernel* Não-Lineares**

   Indicado para dados que não são linearmente separáveis. Exemplos comuns:

   (a) RBF (*Radial Basis Function*):

   $$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right).$$

   Um exemplo de aplicação do kernel RBF pode ser encontrado em (ESKANDAR, 2023).

(b) Polinomial:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^d,$$

onde c é uma constante e d é o grau do polinômio. Uma aplicação prática do kernel polinomial é apresentada em (CHANG et al., 2010).

(c) Sigmoidal:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\kappa \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c).$$

O comportamento e a aplicabilidade do kernel sigmoidal são discutidos em (LIN; LIN, 2003).

(d) *Spline Kernel*: Utilizado para dados com dependências não lineares suaves. Uma aplicação do spline kernel em simulação computacional é mostrada em (SZYMANSKI et al., 2006).

(e) *Gaussian Kernel*: Variante do RBF, com foco em proximidade. Um exemplo de generalização e uso do kernel Gaussiano está em (CHAKRABORTY et al., 2023).

(f) *Laplacian Kernel*:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right).$$

Abordagens de mapeamento de características com esse kernel são discutidas em (AHIR; PANDIT, 2024).

(g) ANOVA *Kernel*:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{d} \exp\left(-\sigma(x_{i,k} - x_{j,k})^2\right).$$

Um uso prático do ANOVA kernel em análise de sensibilidade é detalhado em (DURRANDE et al., 2013).

(h) *Chi-Square Kernel*: Utilizado principalmente em aplicações de visão computacional:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{1}{2} \sum_{k} \frac{(x_{i,k} - x_{j,k})^2}{x_{i,k} + x_{j,k}}.$$

A aplicação do chi-square kernel em análise de sentimentos é explorada em (HOKIJULIANDY; NAPITUPULU; FIRDANIZA, 2023).

Esses kernels fornecem flexibilidade para ajustar o SVM a diferentes tipos de dados e distribuições.

3 Resultados

3.1 Conjunto de Dados

O conjunto de dados *Breast Cancer Wisconsin (Diagnostic)* foi criado em 1992 por pesquisadores da Universidade de Wisconsin, incluindo o Dr. William H. Wolberg, W. Nick Street e Olvi L. Mangasarian. Ele contém informações de 699 pacientes, dos quais 458 apresentam tumores benignos e 241 malignos. A base já foi utilizada em centenas (se não milhares) de estudos acadêmicos para análises comparativas de algoritmos de aprendizado de máquina. Veja por exemplo, a aplicação de SVM em detecção de cancer de mama por (BENNETT; MANGASARIAN, 1992) ou em (ABRAHAM; JAIN; YANG, 2005), que utilizou de algoritmos híbridos, com base em SVM, para melhorar o diagnóstico de câncer. Além destes exemplo, temos também (SAHAN et al., 2007), que também propõe um modelo híbrido, utilizando-se de redes neurais artificiais e K-NN (*K-nearest neighbors*) para analisar a referida base de dados.

Esse histórico estabelece um precedente sobre sua credibilidade e relevância, visto que é uma base de dados amplamente utilizada na comunidade acadêmica e científica. Pode-se citar trabalhos como o de (COWSIK; CLARK, 2019) sobre RNA's (Redes Neurais Artificiais) de duas camadas que usa de correlações entre variáveis de entrada para classificar tumores de mama como malignos ou benignos, utilizando a base de dados *Breast Cancer Wisconsin.* Um outro trabalho muito interessante é o de (AGARAP, 2018), que compara 6 diferentes tipos de algoritmos que foram aplicados à base de dados, para medir sua acurácia na classificação de câncer de mama. Estes são apenas exemplos pontuais do diversos trabalho que existem com a utilização da base de dados.

Os dados da base foram obtidos pela UCI *Machine Learning Repository, Breast Cancer Wisconsin (Diagnostic) Data Set*[4].

3.1.1 Composição dos Dados

Os dados foram obtidos a partir de imagens digitalizadas de exames de punção aspirativa por agulha fina (FNA) das massas mamárias. As amostras celulares coletadas foram analisadas microscopicamente para extrair características morfológicas dos núcleos, sendo classificadas em valores discretos de 1 a 10 por especialistas. Cada amostra é descrita por 10 variáveis numéricas discretas que representam aspectos visuais como espessura dos aglomerados celulares, uniformidade de tamanho e forma celular, presença de núcleos

[4]disponível em: ⟨https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original⟩, acesso em 07/05/2025.

desprotegidos, entre outros (WOLBERG, 1990).

Essa abordagem estatística permite representar de forma compacta e compreensível um conjunto de dados que, de outra forma, seria massivo e redundante. O uso de estatísticas como média, desvio padrão e o máximo são amplamente adotadas em contextos clínicos para facilitar a interpretação dos dados pelos médicos. Por exemplo, os trabalho recentes de (YEŞILBAŞ et al., 2024) e (DONG et al., 2024), que utilizam de estatísticas descritivas como média e desvio padrão para analisar resultados clínicos.

3.1.2 Variáveis Presentes

O conjunto de dados inclui as seguintes variáveis:

1. **ID**: Número de identificação da amostra.
2. **Classe**: Diagnóstico do tumor:
   (a) ***Benign***: Benigno
   (b) ***Malignant***: Maligno
3. **Características celulares (10 variáveis):**
   (a) **Cl.thickness** – Espessura dos aglomerados celulares (*Clump Thickness*)
   (b) **Cell.size** – Uniformidade do tamanho celular (*Uniformity of Cell Size*)
   (c) **Cell.shape** – Uniformidade da forma celular (*Uniformity of Cell Shape*)
   (d) **Marg.adhesion** – Aderência marginal entre as células (*Marginal Adhesion*)
   (e) **Epith.c.size** – Tamanho das células epiteliais isoladas (*Single Epithelial Cell Size*)
   (f) **Bare.nuclei** – Presença de núcleos desprotegidos (*Bare Nuclei*)
   (g) **Bl.cromatin** – Aparência da cromatina (*Bland Chromatin*)
   (h) **Normal.nucleoli** – Presença de nucléolos normais (*Normal Nucleoli*)
   (i) **Mitoses** – Número de mitoses (*Mitoses*)

Essas variáveis fornecem uma descrição detalhada das características dos núcleos celulares, auxiliando na distinção entre tumores benignos e malignos. Este conjunto de dados tem sido amplamente utilizado em pesquisas de aprendizado de máquina para desenvolver e avaliar modelos de classificação na detecção de câncer de mama como o trabalho de (AGARAP, 2018), que aplicou alguns algoritmos na base de dados.

3.2 Preparação dos Dados

O conjunto de dados utilizado é composto por informações clínicas relacionadas a características celulares obtidas a partir de imagens digitalizadas de massas mamárias, com a respectiva classificação entre tumores benignos e malignos.

Inicialmente, o conjunto de dados passou por um processo de pré-processamento. Primeiramente, a coluna de identificação dos pacientes foi removida, por não conter informação preditiva relevante. Em seguida, todos os registros que apresentavam valores ausentes foram eliminados, a fim de garantir a integridade e a consistência das análises subsequentes.

As variáveis preditoras, originalmente do tipo categórica, foram convertidas para o formato numérico, permitindo a aplicação dos algoritmos de classificação. A variável resposta foi transformada em um fator binário, distinguindo as duas categorias de interesse: benigno e maligno (0 e 1).

Para a divisão dos dados, optou-se por reservar aproximadamente 35% das observações para o conjunto de treino, utilizando a função de particionamento estratificado para manter a proporção entre as classes. Este procedimento assegura que tanto os conjuntos de treino quanto de teste possuam distribuições semelhantes, minimizando o viés na avaliação dos modelos.

A distribuição das classes nos conjuntos de treino e teste foi visualizada através de gráficos de barras e de proporção, evidenciando o equilíbrio relativo mantido após o particionamento.

3.3 Construção e Avaliação Inicial dos Modelos SVM

3.3.1 Treinamento e Avaliação dos Modelos

Com os dados devidamente preparados, procedeu-se à construção dos primeiros modelos de classificação utilizando a técnica do SVM. Quatro tipos de *kernels* foram empregados, sendo o *kernels* linear uma base para comparação. Os outros três *kernels* utilizados foram: Sigmoidal, Radial e Polinomial. Todos os modelos foram aplicados a uma malha de valores avaliados sob a lógica de qual é o modelo que gera menor número de falsos positivos.

Figura 2: Malha de Valores - Modelo SVM com *Kernel* Linear

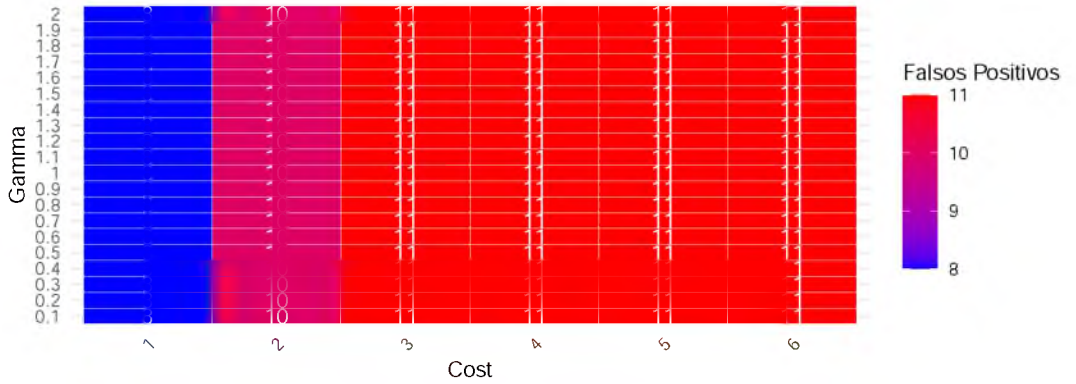


Figura 3: Malha de Valores - Modelo SVM com *Kernel* Radial

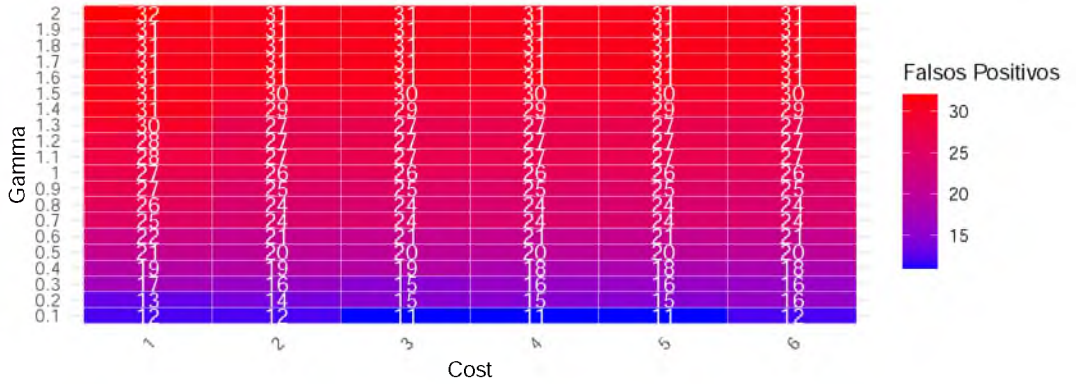


Figura 4: Malha de Valores - Modelo SVM com *Kernel* Sigmoidal

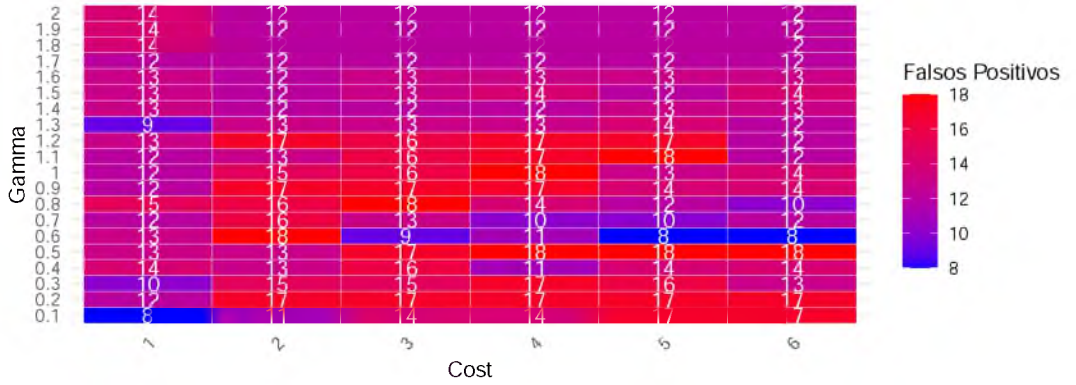
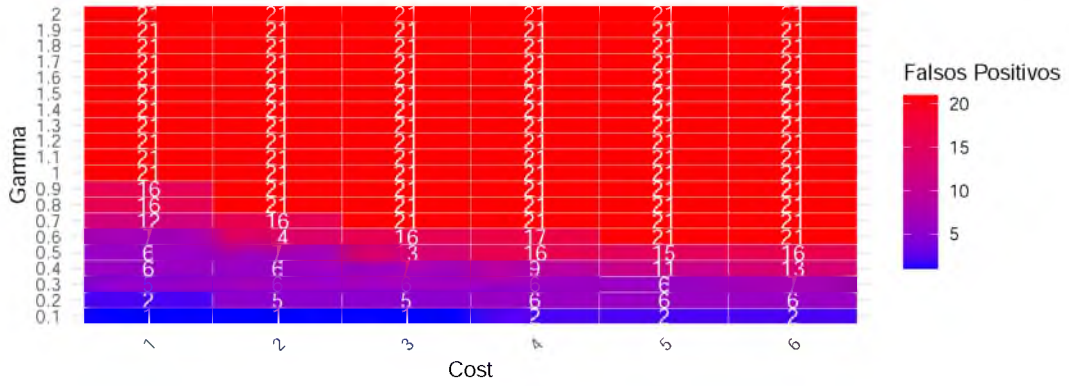


Figura 5: Malha de Valores - Modelo SVM com *Kernel* Polinomial



Os melhores resultados foram obtidos com *kernel* polinomial, onde foi possível identificar uma diferença relevante do número de falsos positivos entre os modelos como mostram as Figuras 2, 3, 4 e 5. Por sua vez, o modelo linear foi ajustado para otimização de desempenho assim como foi feito com os outros modelos.

A avaliação dos modelos foi realizada com base em métricas tradicionais de classificação, incluindo a acurácia, a sensibilidade (*recall*), a especificidade, a curva ROC (*Receiver Operating Characteristic*) e a estatística KS (Kolmogorov-Smirnov).

As Tabelas a seguir comparam os resultados de teste dos dois modelos:

Tabela 1: Matriz de confusão SVM com Kernel Linear

Predição	**Benigno**	**Maligno**	**Total**
Benigno	198	8	206
Maligno	3	100	103
Total	201	108	309

Tabela 2: Matriz de confusão SVM com Kernel Polinomial

Predição	**Benigno**	**Maligno**	**Total**
Benigno	200	14	214
Maligno	1	94	138
Total	201	108	309

Tabela 3: Indicadores de Desempenho dos Modelos

Métrica	**Modelo Linear**	**Modelo Polinomial**
Acurácia	0,9644	0,9514
Sensibilidade	0,9850	0,9950
Especificidade	0,9259	0,8703
Acurácia Balanciada	0,9555	0,9326
F1 Score	0,9729	0,9638
Kappa	0,9208	0,8901

Os resultados indicaram que ambos os modelos alcançaram valores elevados de acurácia, situando-se próximo de 95%, demonstrando boa capacidade de generalização para o conjunto de teste. De maneira geral, os resultados das métricas de avaliação se mostraram muito próximos, com uma diferença mais acentuada nas especificidades dos modelos.

A avaliação de modelos preditivos, especialmente em contextos médicos, exige métricas que capturem não apenas a acurácia global, mas também a capacidade do modelo em distinguir entre classes de interesse (por exemplo, maligno e benigno).

Duas métricas fundamentais para essa avaliação são a Área sob a Curva ROC (AUC) e a estatística de KS.

A AUC é uma medida que resume o desempenho de um classificador em todos os possíveis limiares de decisão. Ela representa a probabilidade de que o modelo atribua uma pontuação mais alta a uma observação positiva do que a uma negativa. Em termos práticos, uma AUC próxima de 1 indica excelente capacidade discriminativa, enquanto valores próximos de 0,5 sugerem desempenho equivalente ao acaso. Essa avaliação é crucial por contemplar tanto a sensibilidade (verdadeiros positivos) quanto a especificidade (verdadeiros negativos), permitindo uma visão abrangente da capacidade de generalização do modelo, independentemente do ponto de corte utilizado.

Já a estatística de KS mede a maior diferença entre as funções de distribuição acumulada das pontuações previstas para as classes positiva e negativa. Em outras palavras, o KS quantifica o quão bem separadas estão as distribuições de previsão para cada classe. Um valor de KS próximo a 1 indica que as classes são altamente separáveis.

Portanto, a análise conjunta de AUC e KS fornece uma avaliação consistente do desempenho do modelo, assegurando tanto a capacidade discriminativa quanto a separabilidade das classes, sendo aspectos essenciais para garantir confiança em cenários sensíveis como diagnósticos médicos.

Conforme as Figuras 6 e 7 abaixo, podemos avaliar os resultados obtidos por cada um dos modelos.

Figura 6: Curva ROC para o Modelo SVM com Kernel Linear.

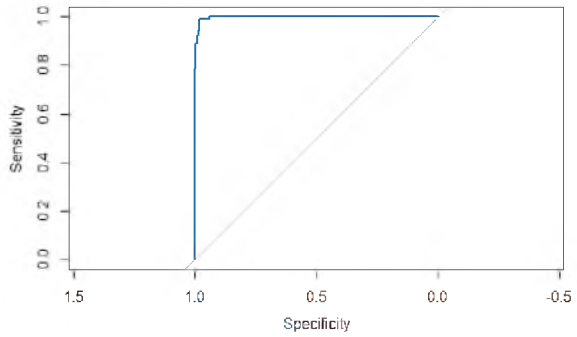
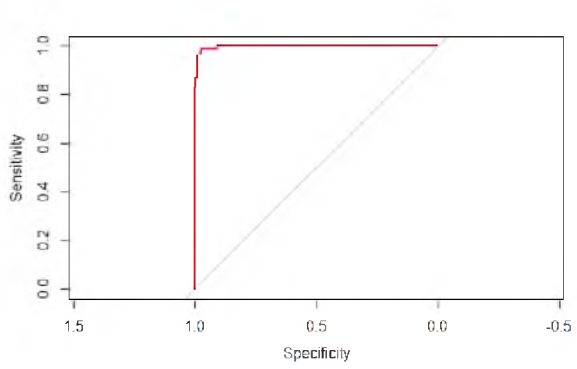


Figura 7: Curva ROC para o Modelo SVM com Kernel Polinomial.



O modelo SVM com kernel polinomial apresentou uma AUC de 0,997, conforme a Figura 7 assim como o modelo SVM com kernel linear, que também obteve uma AUC de 0,997, de acordo com a Figura 6, ambos indicando excelente capacidade discriminativa. Na prática, a AUC quantifica a habilidade do modelo em distinguir corretamente entre as classes positivas (malignas) e negativas (benignas). Um valor de AUC próximo a 1 sugere que o modelo atribui probabilidades mais altas a exemplos positivos do que a exemplos negativos na grande maioria dos casos, o que é crucial em contextos como o diagnóstico médico, onde erros de classificação podem ter consequências severas.

A estatística de Kolmogorov–Smirnov (KS) quantifica a distância máxima entre as funções de distribuição acumulada empírica e teórica, ou entre duas distribuições empíricas. Essa distância corresponde ao maior desvio vertical observado entre as duas curvas, sendo formalmente definida como:

$$D_n = \sup_x |F_n(x) - F(x)|$$

no caso do teste de uma amostra, em que $F_n(x)$ é a função de distribuição acumulada empírica e $F(x)$ é a função de distribuição acumulada teórica.

Para o teste de duas amostras, a estatística é dada por:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

em que $F_n(x)$ e $G_m(x)$ são as funções de distribuição acumulada empírica das duas amostras. O valor de D é utilizado como estatística de teste para avaliar a aderência entre as distribuições consideradas.

Nas Figuras 8 e 9 é possível conferir o desempenho das curvas de distribuição empírica, utilizada na avaliação da estatística KS.

Figura 8: Curvas de Distribuição Acumulada Empírica para Cálculo do KS - Linear.

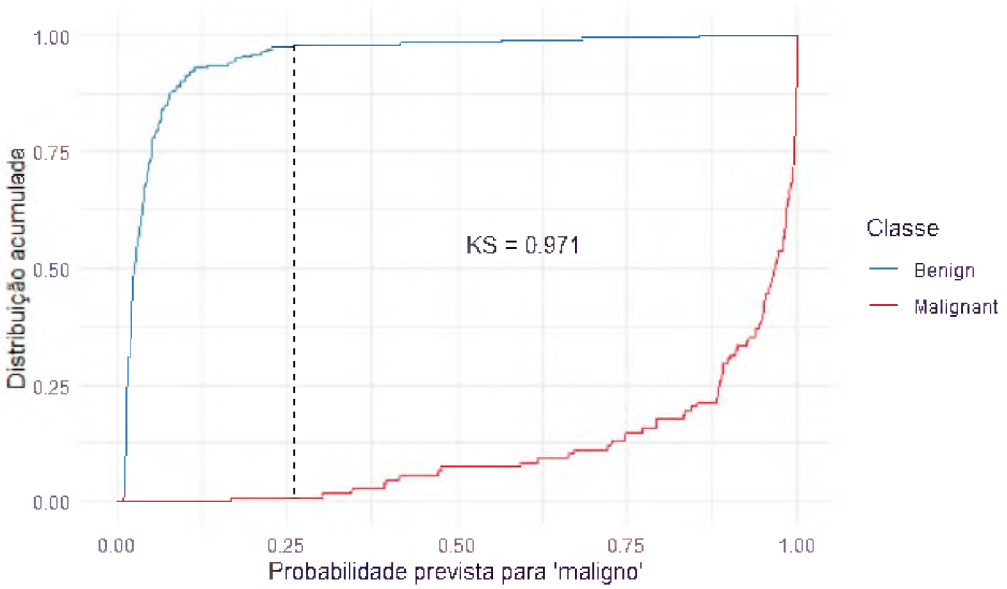
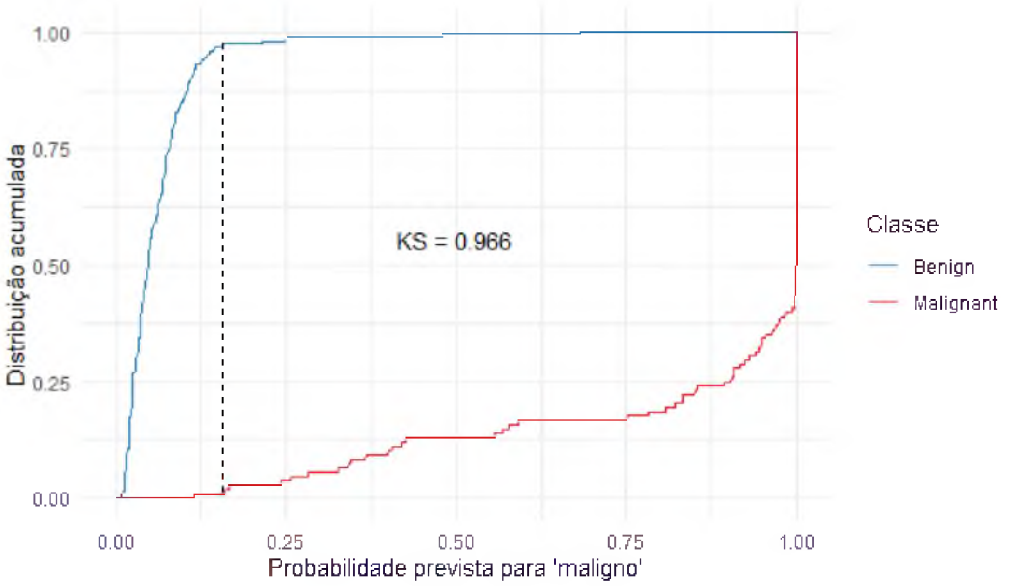


Figura 9: Curvas de Distribuição Acumulada Empírica para Cálculo do KS - Polinomial.



Além disso, a estatística KS foi calculada para ambos os modelos, resultando em 0,966 para o SVM polinomial e 0,971 para o SVM linear, de acordo com as Figuras 8 e 9. A estatística KS mede a maior diferença entre as distribuições acumuladas das classes positiva e negativa. Em outras palavras, ela avalia o quão separáveis são as distribuições das pontuações previstas para cada classe. Um valor de KS próximo de 1 indica excelente separabilidade, o que confirma que o modelo consegue distinguir com grande eficácia entre amostras benignas e malignas.

A avaliação do KS é crítica porque permite detectar, de maneira intuitiva, se há sobreposição significativa entre as pontuações das classes — um problema que pode comprometer a interpretação clínica dos resultados. Em aplicações na área médica, valores de KS superiores a 0,6 já são considerados ótimos; portanto, os valores obtidos reforçam a eficácia dos modelos desenvolvidos, mesmo em um espaço de alta dimensionalidade.

Por fim, as visualizações a seguir da curva ROC e as curvas acumuladas por classe, usadas no cálculo do KS, reforçam visualmente o bom desempenho do modelo, mostrando uma separação clara e forte entre as classes. Essas análises complementam a avaliação numérica e sustentam a decisão pela adoção do modelo polinomial como principal solução deste trabalho pela sua alta capacidade de diferenciação entre as classes, alta acertividade e a baixíssima quantidade de falsos-positivos detectada pelo modelo.

3.3.2 Avaliação da Importância das Variáveis

A interpretação dos modelos de SVM não é tão direta quanto em modelos lineares ou baseados em árvores de decisão. No entanto, é possível estimar a **importância relativa das variáveis** por meio de abordagens empíricas, como ao criar modelos retirando uma variável afim de avaliar o impacto de cada das variáeis na performance preditiva do modelo, baseando-se na acurácia.

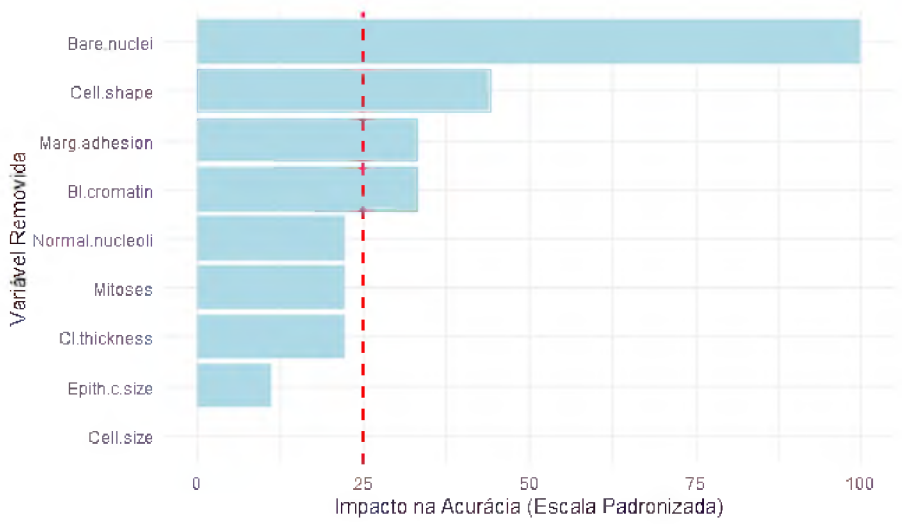
A Figura 10 apresenta os valores estimados de importância das variáveis no modelo SVM com *kernel* polinomial. Os valores foram padronizados para facilitar a interpretação. Nota-se que algumas variáveis se destacam significativamente em relação às demais, indicando maior influência na definição da fronteira de decisão do classificador.

Entre os atributos com maior relevância, destacam-se:

- Uniformidade da forma celular (Cell.shape);
- Grau de adesão da margem celular das células tumorais (Marg.adhesion);
- Presença de núcleos desprotegidos (Bare.nuclei);
- Cromatina suave (Bl.cromatin).

Estes atributos estão diretamente relacionados a características das células, frequentemente utilizadas por profissionais da saúde para avaliar suspeitas de malignidade. O fato de o modelo estatístico reconhecer essas variáveis como relevantes valida a coerência da análise computacional a favor da integração entre estatística e medicina.

Figura 10: Importância Relativa das Variáveis no Modelo SVM com Kernel Polinomial.



A presença de outras variáveis com relevância moderada também evidencia que o modelo leva em consideração uma **combinação de múltiplas características** para realizar a classificação — característica desejável em situações de alta complexidade, como o diagnóstico de câncer de mama. Não somente as variáveis ajudam a explicar como a entender toda a situação, que no caso deste estudo, foi observada a significativa diferença de níveis de importância dos tamanhos celulares para diagnosticar corretamente e explicar o problema.

3.3.3 Visualização do SVM - Variáveis Selecionadas

Para explorar a capacidade de separação proporcionada pelas variáveis mais importantes, foram construídas visualizações de fronteiras de decisão considerando a combinação dos pares dessas variáveis.

Figura 11: Fronteiras de Decisão Geradas pelo SVM com Duas das Combinações de Pares das 4 Variáveis Mais Explicativas. - Parte 1

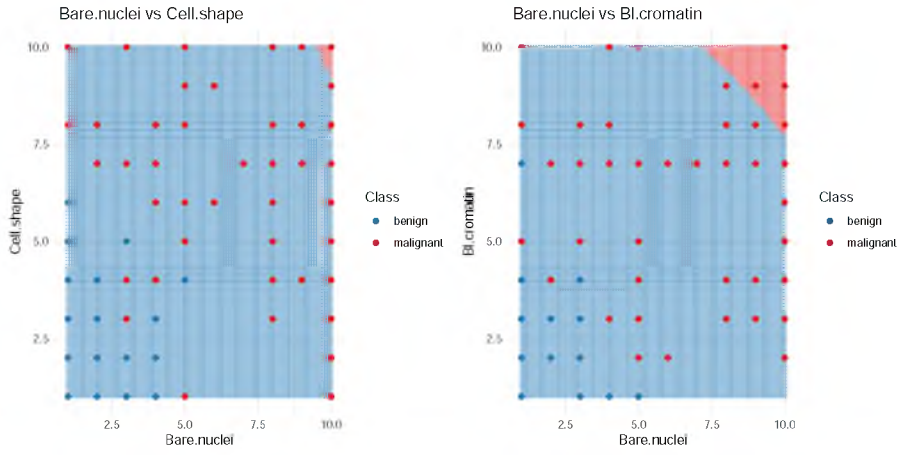


Figura 12: Fronteiras de Decisão Geradas pelo SVM com Duas das Combinações de Pares das 4 Variáveis Mais Explicativas. - Parte 2

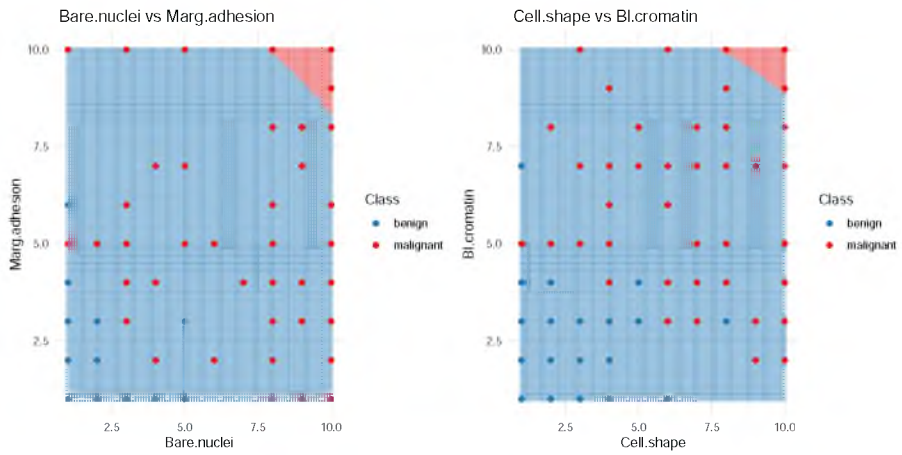
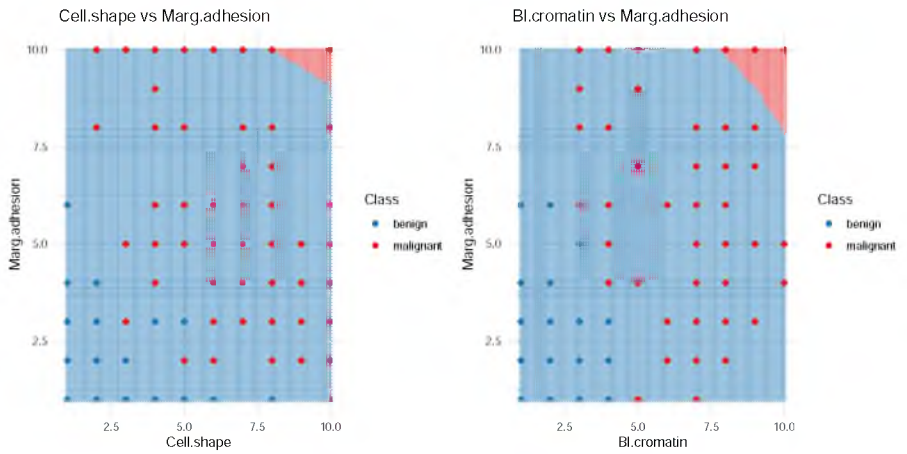


Figura 13: Fronteiras de Decisão Geradas pelo SVM com Duas das Combinaçõs de Pares das 4 Variáveis Mais Explicativas. - Parte 3



Ao observar as Figuras 11, 12 e 13, notou-se que, apesar de alguma separação entre as classes, as fronteiras não eram bem definidas, havendo considerável sobreposição entre as observações benignas e malignas. Essa dificuldade visual reforçou o diagnóstico de que as relações entre as variáveis são não lineares e de difícil separação apenas por meio de combinações simples.

Dessa forma, tornou-se necessária a adoção de uma abordagem que pudesse capturar de maneira mais eficaz as estruturas latentes dos dados, justificando a utilização da Análise de Componentes Principais (PCA), que é abordada na próxima Seção. Para mais informações sobre a PCA, consulte (JOHNSON; WICHERN et al., 2002).

3.4 Aplicação do PCA e Avaliação dos Modelos

3.4.1 Visualização dos Dados sob Ótica do PCA

Para melhor compreender a separabilidade das classes no conjunto de dados, foi aplicada uma **PCA**. O objetivo foi reduzir a dimensionalidade das variáveis explicativas e observar visualmente se há uma tendência de agrupamento entre a classe das amostras, benignas e malignas.

Tabela 4: Resumo da Análise de Componentes Principais (PCA) - Parte 1

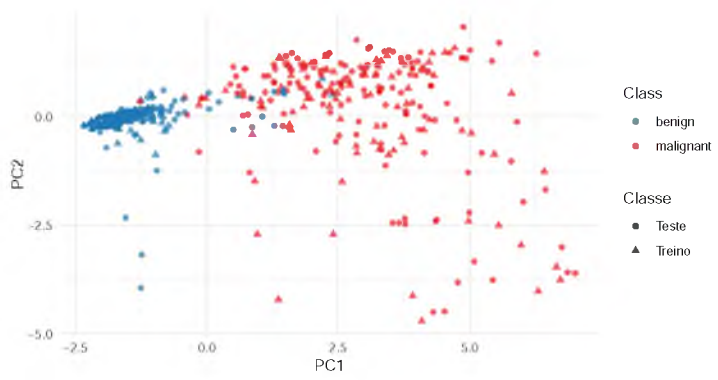
Componente	PC1	PC2	PC3	PC4	PC5
Desvio padrão	2,4382	0,8799	0,7540	0,6564	0,5794
Proporção da variância	66,05%	8,60%	6,31%	4,78%	3,73%
Proporção acumulada	66,05%	74,65%	80,97%	85,76%	89,49%

Tabela 5: Resumo da Análise de Componentes Principais (PCA) - Parte 2

Componente	PC6	PC7	PC8	PC9
Desvio padrão	0,5750	0,5335	0,4723	0,3280
Proporção da variância	3,67%	3,16%	2,47%	1,19%
Proporção acumulada	93,16%	96,32%	98,80%	100%

A PCA foi calculada com todas as variáveis numéricas do conjunto original e os dados foram projetados nas duas primeiras componentes principais (PC1 e PC2). Dentro dos resultados obtidos, nota-se, de acordo com as Tabelas 4 e 5, as duas primeiras componentes comportam aproximadamente 75% das informações dos dados. Isso nos ajuda a resumir, explicar e visualizar os dados do trabalho, como nas Figura 14:

Figura 14: Distribuição dos Dados no Espaço das Duas Primeiras Componentes Principais (PC1 e PC2).



O gráfico resultante, apresentado na Figura 14, revela uma boa separação entre as classes, embora com certo grau de sobreposição em algumas regiões. Essa sobreposição acontece bastante em dados reais onde o comportamento pode apresentar áreas não muito definidas dada uma certa projeção. A PCA foi utilizada nessa situação a fim de facilitar a aprensentação visual dos dados e tentar minimizar a possível sobreposição dos dados. Em sua grande maioria, os dados apresentam centros/centroides, uma concentração de observações em certas áreas, criando assim uma possível tendência que pode facilitar a detecção das classes pelo SVM.

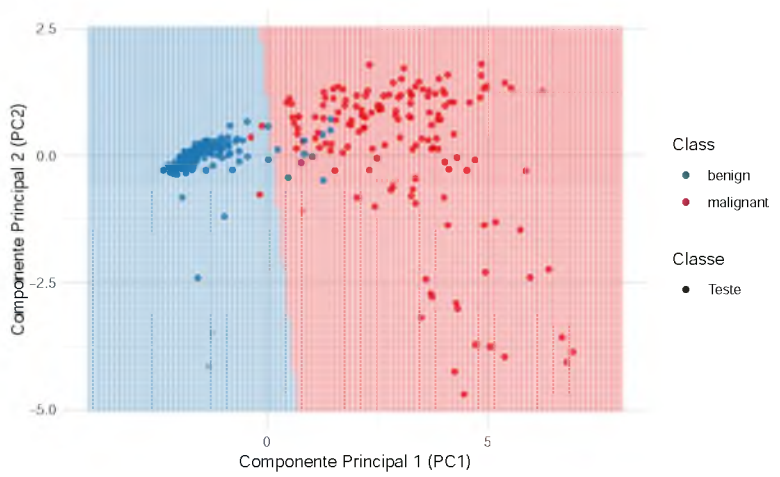
3.4.2 Aplicação de Modelo PCA e Análise dos Resultados

Com base nas duas componentes principais, foram treinados dois modelos, um com *kernel* linear e um modelo de *kernel* polinomial, com os mesmo parâmetros inicialmente utilizados pós análise nas possibilidades de parâmetros. Tais parâmetros são:

a proporção das amostras de treino e teste, as entradas do modelo SVM, avaliação das métricas. Essa abordagem permitiu a geração da **fronteira de decisão** em duas dimensões nas duas abordagens.

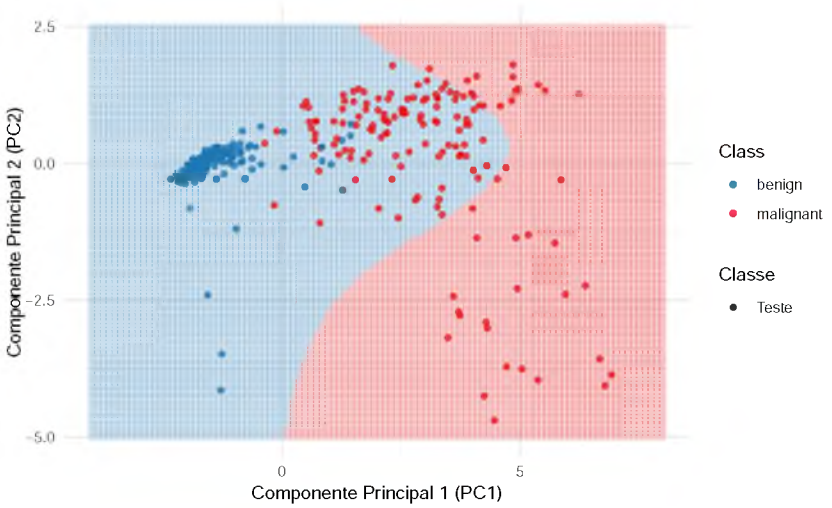
A Figura 15 mostra a fronteira gerada pelo classificador sobre o espaço das duas primeiras componentes principais no modelo linear.

Figura 15: Fronteira de Decisão do Modelo SVM com Kernel Linear no Espaço Bidimensional (PC1 e PC2).



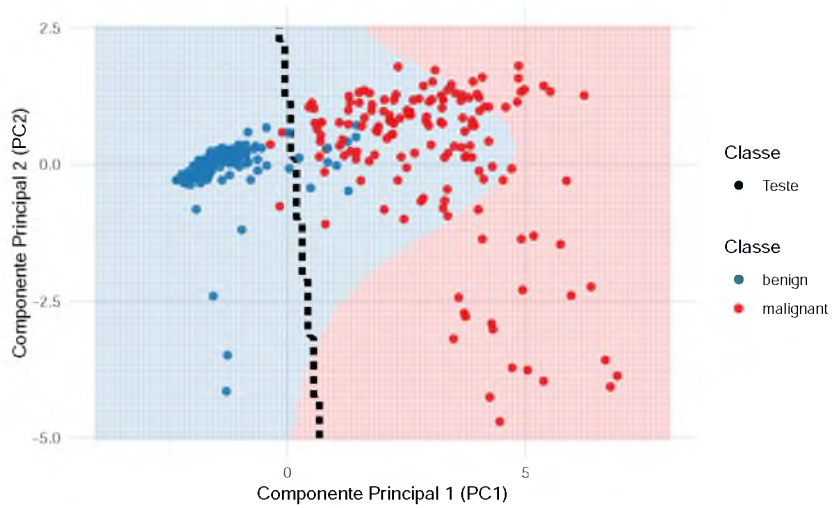
A Figura 16 mostra as possibilidades de adequação do modelo aos dados com o *kernel* polinomial.

Figura 16: Fronteira de Decisão do Modelo SVM com *Kernel* Polinomial no Espaço Bidimensional (PC1 e PC2).



Cabe destacar que o modelo SVM treinado com as duas componentes principais (PC1 e PC2) foi utilizado apenas para fins de visualização da fronteira de decisão em duas dimensões, e não substitui o modelo final treinado com todas as variáveis originais. Essa simplificação visa apenas ilustrar a capacidade do *kernel* polinomial de construir limites de decisão não lineares mesmo em projeções bidimensionais, como pode ser comparado com a Figura 17.

Figura 17: Fronteira de Decisão do Modelo SVM com *Kernel* Polinomial e Linear no Espaço Bidimensional (PC1 e PC2).



Cada um dos modelos de PCA geram suas respectivas matrizes de confusão, que podem ser conferidas com as Tabelas 6 e 7:

Tabela 6: Matriz de Confusão – Modelo Linear SVM com PC1 e PC2

Predição	**Benigno**	**Maligno**	**Total**
Benigno	281	4	285
Maligno	7	151	158
Total	288	155	443

Tabela 7: Matriz de Confusão – Modelo Polinomial SVM com PC1 e PC2

Predição	**Benigno**	**Maligno**	**Total**
Benigno	288	31	319
Maligno	0	124	124
Total	288	155	443

Não só as contagens foram realizadas mas também foram coletados as métricas de avaliação, que podem ser conferidas na Tabela 8 abaixo:

Tabela 8: Indicadores de Desempenho dos Modelos de CPA

Métrica	**Modelo Linear CPA**	**Modelo Polinomial CPA**
Acurácia	0,9751	0,9300
Sensibilidade	0,9756	1
Especificidade	0,9741	0,8000
Acurácia Balanciada	0,9749	0,9000
F1 Score	0,9808	0,9489
Kappa	0,9456	0,8387

Ao aplicar o modelo gerado pelo PCA, na predição das classes na amostra inteira, observamos resultados parecidos, de maneira geral, com o modelo polinomial sem PCA. Há de se ponderar que foi um modelo mais "radical", visto que não obteve nenhum falso-positivo, sua taxa de acurácia foi menor, mas os resultados ainda são bastante parecidos dado que foi utilizado apenas duas componentes principais em ambos os modelos.

4 Conclusão

Neste trabalho, foi investigada a aplicação de SVM na detecção de câncer de mama a partir de atributos morfológicos celulares, utilizando o conjunto de dados *Breast Cancer Wisconsin (Original)*. A pesquisa integrou fundamentos matemáticos da técnica com a implementação computacional em R, evidenciando o potencial de modelos baseados em SVM como suporte à área médica.

Do ponto de vista metodológico, foram testadas duas configurações de *kernel*: linear e polinomial. Ambas apresentaram acurácia acima dos 90%, mas uma análise mais aprofundada da matriz de confusão revelou que o modelo com *kernel* polinomial apresentou menor incidência de falsos positivos — fator decisivo na escolha final, dada a sensibilidade do contexto clínico. O modelo polinomial demonstrou ainda um excelente desempenho nas métricas quantitativas, com AUC de 0,997 e estatística KS de 0,966, indicando elevada capacidade discriminativa.

Adicionalmente, foi possível observar, por meio da PCA, uma separação visual satisfatória entre as classes no espaço projetado em duas dimensões. O modelo SVM com *kernel* polinomial mostrou-se adaptável aos dados, sendo capaz de construir fronteiras de decisão não lineares coerentes com os agrupamentos observados.

A análise da importância das variáveis revelou que atributos como *a Uniformidade da forma celular*, *a Uniformidade do tamanho celular*, *a Presença de núcleos desprotegidos* e *a Cromatina suave* foram os mais relevantes para a classificação, reforçando a aderência do modelo ao conhecimento biomédico já consolidado sobre alterações celulares associadas à malignidade.

Os resultados obtidos indicam que o uso de SVM, especialmente com *kernel* polinomial, é uma alternativa viável e precisa para auxiliar no diagnóstico do câncer de mama, desde que associado a ferramentas de apoio médico. Ressalta-se, no entanto, que o modelo desenvolvido neste estudo tem caráter experimental e foi aplicado a um conjunto de dados limitado, pela tipologia dos dados discretos e com apenas nove atributos. Isso abre margem para algoritmos ainda melhores caso os dados disponibilizados possuam qualidade e quantidade maiores.

Conclui-se, portanto, que a integração entre modelagem matemática e análise computacional, quando orientada por critérios técnicos e sensíveis ao contexto da área da saúde, pode contribuir significativamente para o avanço do diagnóstico clínico. Ressalta-se, entretanto, que tais aplicações devem ser desenvolvidas em colaboração com pesquisadores da área, cujas expertises acumuladas ao longo de anos de trabalho são fundamentais para garantir a relevância e a aplicabilidade dos resultados.

Referências

ABRAHAM, A.; JAIN, L.; YANG, J. Hybrid intelligent systems for breast cancer diagnosis. *International Journal of Biomedical Soft Computing and Human Sciences*, v. 10, n. 1, p. 55–63, 2005.

AGARAP, A. F. M. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In: *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*. ACM, 2018. (ICMLSC 2018), p. 5–9. Disponível em: ⟨http://dx.doi.org/10.1145/3184066.3184080⟩.

AHIR, S.; PANDIT, P. Feature maps for the laplacian kernel and its generalizations. *arXiv preprint arXiv:2502.15575*, 2024.

AZEVEDO-MARQUES, P. M. Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, 2001.

BENNETT, K. P.; MANGASARIAN, O. L. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, v. 1, n. 1, p. 23–34, 1992.

BILLINGSLEY, P. *Probability and Measure*. 3. ed. New York: John Wiley & Sons, 1995. ISBN 9780471007104.

BISHOP, C. M.; NASRABADI, N. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. v. 4.

CHAKRABORTY, S. et al. Machine learning application of generalized gaussian radial basis function kernel. *Mathematics*, MDPI, v. 12, n. 6, p. 829, 2023.

CHANG, Y.-W. et al. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, v. 11, p. 1471–1490, 2010.

COWSIK, A.; CLARK, J. W. Breast cancer diagnosis by higher-order probabilistic perceptrons. *ArXiv*, abs/1912.06969, 2019. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:209376230⟩.

CRUZ, T. N.; CRUZ, T. M.; SANTOS, W. P. Detection and classification of mammary lesions using artificial neural networks and morphological wavelets. *IEEE Latin America Transactions*, v. 16, n. 3, p. 926–932, 2018.

DONG, B. et al. Evaluation of very-high-frequency ultrasound imaging characteristics of dermatofibroma. *Clinical, Cosmetic and Investigational Dermatology*, 2024. Disponível em: ⟨https://www.tandfonline.com/doi/pdf/10.2147/CCID.S493437⟩.

DONG, R. et al. Detection and direct readout of drugs in human urine using dynamic surface-enhanced raman spectroscopy and support vector machines. *Analytical Chemistry*, v. 87, n. 5, p. 2937–2944, 2015. PMID: 25634247. Disponível em: ⟨https://doi.org/10.1021/acs.analchem.5b00137⟩.

DURRANDE, N. et al. Anova kernels and rkhs of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, Elsevier, v. 115, p. 57–67, 2013.

ESKANDAR, S. Introduction to rbf svm: A powerful machine learning algorithm for non-linear data. *Medium*, 2023. Disponível em: ⟨https://medium.com/@eskandar.sahel/introduction-to-rbf-svm-a-powerful-machine-learning-algorithm-for-non-linear-data-1d1cfb55a1a⟩.

HOKIJULIANDY, E.; NAPITUPULU, H.; FIRDANIZA. Application of svm and chi-square feature selection for sentiment analysis of indonesia's national health insurance mobile application. *Mathematics*, MDPI, v. 11, n. 17, p. 3765, 2023.

JOHNSON, R. A.; WICHERN, D. W. et al. Applied multivariate statistical analysis. Prentice hall Upper Saddle River, NJ, 2002.

LIN, H.-T.; LIN, C.-J. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. *Technical report, Department of Computer Science and Information Engineering, National Taiwan University*, 2003. Disponível em: ⟨https://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf⟩.

MALATHI, M. et al. An estimation of pca feature extraction in eeg-based emotion prediction with support vector machines. In: GUPTA, D. et al. (Ed.). *Proceedings of Data Analytics and Management*. Singapore: Springer Nature Singapore, 2022. p. 651–664. ISBN 978-981-16-6289-8.

MELLO, R. F. de; PONTI, M. A. *Machine Learning: A Practical Approach on the Statistical Learning Theory*. [S.l.]: Springer, 2018.

OLIVEIRA, S. B. *Acesso de mulheres com câncer de mama aos serviços de atenção à saúde: Perspectiva de usuárias, profissionais e gestores*. Dissertação (Dissertação de Mestrado em Saúde Coletiva) — Universidade Federal da Bahia, 2021. Disponível em: ⟨https://repositorio.ufba.br/handle/ri/33866⟩.

SAHAN, S. et al. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*, v. 37, n. 3, p. 415–423, 2007.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014. ISBN 1107057132.

SZYMANSKI, B. K. et al. Spline kernel based machine learning tool. In: *Proceedings of the 2006 Winter Simulation Conference*. [S.l.]: Springer, 2006.

THOMAS, G. B.; WEIR, M. D.; HASS, J. *Cálculo, Volume 2*. 12ª. ed. São Paulo: Pearson Education do Brasil, 2012.

TOMAZELLI, J. G. et al. Avaliação das ações de detecção precoce do câncer de mama no brasil por meio de indicadores de processo: Estudo descritivo com dados do sismama, 2010-2011. *Epidemiologia e Serviços de Saúde*, v. 26, n. 1, p. 61–70, 2017. Disponível em: ⟨https://www.scielo.br/j/ress/a/3BBXVFZMtDGZf5WVznmy9xw/⟩.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995. ISBN 9780387987804.

WOLBERG, W. *Breast Cancer Wisconsin (Original)*. 1990. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HP4Z.

YEŞILBAŞ, Ş. et al. Compliance of cleaning staff to standard precautions in hospital hygiene and affecting factors. *Family Practice and Palliative Care*, Yusuf Haydar ERTEKİN, v. 9, n. 4, p. 104–111, 2024.

5 Apêndice

5.1 Banco de Dados

Cl.thickness	Cell.size	Cell.shape	Marg.adh	Epith.size	Bare.nuc	Bl.chrom	Norm.nuc	Mitoses	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	benign
6	8	8	1	3	4	3	7	1	malignant
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10	9	7	1	malignant
				...					
3	2	1	1	1	1	2	1	1	benign
1	1	1	1	2	1	2	1	1	benign
1	1	1	1	2	1	2	1	1	benign
8	9	9	5	10	10	7	8	1	malignant
10	10	10	10	10	10	7	10	1	malignant
4	8	6	4	3	4	10	6	1	malignant

5.2 Códigos R

O código-fonte desenvolvido em linguagem R (formatado em RMarkdown) e utilizado neste trabalho pode ser acessado diretamente no seguinte link:

Aplicação de SVM na Detecção de Câncer de Mama - TCC - Lucas Menezes.Rmd