Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Comparação do Support Vector Machine com outros métodos de classificação: uma aplicação a dados de nascimentos prematuros

Letícia Bernardes Sartori

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Comparação do Support Vector Machine com outros métodos de classificação: uma aplicação a dados de nascimentos prematuros

Letícia Bernardes Sartori
Orientador: Marcio Alves Diniz

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Comparison of Support Vector Machine with Other Classification Methods: An Application to Premature Birth Data

Letícia Bernardes Sartori
Advisor: Marcio Alves Diniz

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
July 2025

Letícia Bernardes Sartori

Comparação do Support Vector Machine com outros métodos de classificação: uma aplicação a dados de nascimentos prematuros

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Letícia Bernardes Sartori e aprovado pela banca examinadora.

Aprovado em 18 de Novembro de 2025

Banca Examinadora:

- Prof. Dr. Marcio Alves Diniz
- Prof. Dr. Rafael Izbicki
- Prof. Dr. Gustavo Henrique de Araujo Pereira

Aos meus pais, Marcelo Luís Sartori e Maria Betânia Bernardes dos Santos Tamborini, e meu irmão Rodrigo Bernardes Tamborini pelo apoio na minha trajetória.

Agradecimentos

Agradeço a Deus pelas bênçãos e proteção ao longo da minha vida.

Agradeço à minha mãe, Betânia, ao meu pai, Marcelo, e ao meu irmão, Rodrigo, pelas oportunidades e por sempre me apoiarem e estarem presentes em todos os momentos. Com o incentivo deles, amadureci e estou me tornando uma pessoa mais forte para enfrentar os desafios, me ajudando a não desistir, sendo minha inspiração para conquistar meus sonhos e acreditar que tudo é possível. Minha profunda gratidão por serem minha família!!

Agradeço aos professores e funcionários do Departamento de Estatística da UFSCar por todos os ensinamentos e momentos durante esses anos. Em especial, ao meu orientador de Iniciação Científica e de Trabalho de Conclusão de Curso, Marcio Alves Diniz, pelos conselhos, aprendizados, conquistas e paciência durante toda a graduação. Minha banca, Rafael Izbicki e Gustavo Henrique de Araujo Pereira, pelas sugestões e pela avaliação do meu trabalho. Meus amigos que compartilharam esse ciclo comigo e serão lembrados com muito carinho.

Muito obrigada a todos que fizeram parte desta conquista!!

"Dificuldades preparam pessoas comuns para destinos extraordinários"

(As Crônicas de Nárnia)

Resumo

No mundo, milhares de partos são de crianças prematuras, causando danos muitas vezes irreversíveis à saúde das crianças e/ou das mães. Este projeto pretende utilizar técnicas estatísticas, mais especificamente um método de aprendizado de máquina supervisionado para classificação binária (em que a variável de interesse pode assumir apenas dois valores, como 1, caso a criança seja prematura e 0, caso contrário) para identificar e quantificar fatores da mãe e da criança que podem estar associados à prematuridade dos nascimentos. Para isso, pretende-se utilizar o método Support Vector Machine (SVM) e comparar a perfomance preditiva com os modelos paramétricos tradicionais, como a regressão logística, e com o modelo semi-paramétrico baseado em processos gaussianos.

Palavras-chave: *Support Vector Machine (SVM), prematuridade, classificação binária.*

Abstract

Worldwide, thousands of births are premature, often causing irreversible damage to the health of the children and/or their mothers. This project aims to apply statistical techniques, more specifically a supervised machine learning method for binary classification (where the target variable can only assume two values, such as 1 if the baby is premature and 0 otherwise), to identify and quantify maternal and neonatal factors that may be associated with premature births. To this end, the Support Vector Machine (SVM) method will be used and its predictive performance will be compared with traditional parametric models, such as logistic regression, and with a semi-parametric model based on Gaussian processes.

Keywords: *Support Vector Machine (SVM), prematurity, binary classification.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A inferência estatística tem como principais objetivos: (i) realizar previsões a respeito de variáveis ainda não observadas; e (ii) fazer afirmações inferenciais sobre parâmetros de interesse de determinado modelo estatístico. Isso permite que as ferramentas desenvolvidas pela teoria estatística sejam amplamente utilizadas nas mais diversas áreas da ciência.

Dentre essas áreas, será abordada a ciência da saúde, que demanda o conhecimento de certas variáveis relevantes para a previsão de determinados fenômenos. Um problema importante dessa área é a investigação da ocorrência de partos prematuros, que ocorrem quando a criança nasce antes da 37ª semana de gestação, podendo causar morbidades e mortalidade. Por ano, 15 milhões de nascimentos no mundo são prematuros, sendo que 1 milhão deles vêm a óbito devido à complicações da prematuridade (Galo *et al.* (2023)). Nessas situações, quando se conhecem o estado das variáveis explicativas relevantes para o fenômeno, pode-se utilizar métodos de aprendizado de máquina supervisionado para atribuir alguma probabilidade de que uma criança será de um parto prematuro.

Para atingir esse objetivo, será estudado um método de aprendizado de máquina supervisionado para dados binários que podem ser aplicados nas mais diversas áreas do conhecimento, como por exemplo nas áreas tributária e financeira, estimando a probabilidade de um indivíduo ser inadimplente ou adimplente após o parcelamento de tributos ou dívidas. No caso deste projeto, tais modelos serão usados para identificar as variáveis relevantes (da mãe ou da criança) para avaliar se o nascimento será prematuro ou não e para fornecer uma probabilidade de que isso ocorra dados os atributos da mãe e da criança.

A fim de obter esses resultados, será necessário estudar e compreender o método Sup-

port Vector Machine (SVM) para variáveis explicadas ou dependentes binárias, e após isso, estimar esse modelo para um conjunto de dados de nascimento de crianças prematuras. Além disso, pretende-se comparar a perfomance preditiva com outros métodos de classificação, como: regressão logística, regressão logística bayesiana e processos gaussianos, que já foram estudados em Sartori (2024). Essa comparação será feita através dos seguintes métodos: AUC (área sob a curva ROC), escore de Brier e escore logarítmico.

Sendo assim, como mencionado acima, é possível fazer previsões ou inferências sobre o efeito de determinadas variáveis explicativas, e consequentemente, usar esse resultado para os diversos campos da ciência, nesse caso em específico, da saúde, auxiliando os formuladores de políticas públicas a identificar possíveis fatores de risco para partos prematuros.

1.1 Objetivos

Esse projeto tem como principal objetivo o estudo de um método de aprendizado de máquina supervisionado para dados binários, mais especificamente, Support Vector Machine (SVM). A estimação desse modelo será feita em um conjunto de dados referente ao nascimento de prematuros,[1] permitindo a identificação e quantificação dos fatores (variáveis) associados à prematuridade dos nascimentos. Além disso, será aplicada métricas de comparação de perfomance preditiva (AUC, escore de Brier e escore logarítmico) para comparar o SVM com outros modelos para classificação binária (regressão logística, regressão logística bayesiana e processos gaussianos).

Essas conclusões serão apresentadas na forma de afirmações inferenciais, ou seja, baseadas em técnicas de inferência estatística e nos conjuntos de dados disponíveis. Assim sendo, serão afirmações de caráter empírico e, portanto, válidas apenas no recorte geográfico e temporal promovido pelo conjunto de dados.

Para esse estudo, no capítulo 2 será descrita a metodologia necessária, ou seja, os modelos de regressão logística, estimados por máxima verossimilhança e Bayesiana, processos gaussianos, SVM e as métricas de comparação de perfomance preditiva. No capítulo 3 serão descritas as análises descritivas dos dados de nascimentos de bebês prematuros, os resultados preliminares e finais de alguns modelos e as métricas de comparação. O capítulo 4 traz as conclusões do trabalho e possíveis extensões futuras.

[1] Banco de dados público disponível no portal DATASUS.

Capítulo 2

Metodologia

Para implementar os objetivos mencionados acima, será feito o estudo das seguintes metodologias usadas na literatura de inferência estatística.

2.1 Regressão Logística

A regressão logística é o modelo estatístico mais utilizado para modelar variáveis binárias, ou seja, que assumem apenas valores 0 ou 1, sendo interessante para a aplicação desse estudo como função de um conjunto de covariáveis. Dessa forma, seja Y o vetor $(Y_1, Y_2, \ldots, Y_n)^\top \in \{0,1\}^n$ em que n é a quantidade de observações contendo os valores $Y_i = 1$ se o i-ésimo parto foi prematuro e 0 caso contrário. Além disso, o vetor x_i contém as variáveis referentes ao i-ésimo parto, ou seja, $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$, sendo a primeira componente do vetor correspondente ao termo constante do preditor linear, como se verá a seguir. Neste projeto k é a quantidade de covariáveis que serão selecionadas a partir da tabela 3.1 na seção 3.1. As variáveis que serão escolhidas para o estudo são as mesmas utilizadas no projeto de Galo (2020) e Galo *et al.* (2023) permitindo a comparação dos resultados obtidos.

Como modelo estatístico, supõe-se que $Y_i \mid x_i \sim \text{Bernoulli}(p_i), \quad 0 \leq p_i \leq 1$ e, portanto, $p_i = \text{E}(Y_i \mid x_i)$. Uma vez que esse valor esperado deve pertencer ao intervalo $[0,1]$, não se deve supor que ele seja, como nos modelos lineares tradicionais de regressão, idêntica ao preditor linear, η_i, a combinação linear das covariáveis:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik}.$$

Para que isso ocorra, considera-se uma função do preditor linear que associe a qualquer valor real, um número ente 0 e 1. Funções de distribuição têm exatamente essa propriedade e, portanto,

$$p_i = \mathrm{E}(Y_i \mid x_i) = F(\eta_i) = F(x_i\boldsymbol{\beta}) \Longrightarrow \eta_i = x_i\boldsymbol{\beta} = F^{-1}(p_i), \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$$

sendo $F(\cdot)$ uma função de distribuição e $F^{-1}(\cdot)$ é chamada de função de ligação. Se F é a função de distribuição logística, temos o modelo de regressão logística, mas além dele é possível usar outras funções de distribuição como a da normal padrão (modelo probit) ou da Cauchy (modelo Cauchit).

Entretanto, para o estudo de nascimentos prematuros a regressão logística pode apresentar limitações, pois a prematuridade pode ser considerada um evento raro. Como mostra a tabela 3.1 na seção 3.1, o conjunto de dados utilizado neste projeto é desbalanceado (apenas 11.8% dos nascimentos observados são prematuros) e, sendo assim, a regressão logística pode levar a resultados equivocados e/ou viesados. Para solucionar essa limitação é possível utilizar funções de ligações assimétricas, que podem apresentar um melhor ajuste para o modelo. Outras funções utilizadas estão na tabela 2 de Galo *et al.* (2023) que possui uma lista de funções de ligações assimétricas, como exemplo, a potência e a potência reversa do modelo probit. Neste trabalho, ao invés de usar essas funções de ligação, serão utilizados um modelo semi-paramétrico baseado em processos gaussianos, sendo mais flexível do que os modelos paramétricos tradicionais e um modelo de aprendizado de máquina para estudos de classificação binária (SVM) cujas performances preditivas serão comparadas com a da regressão logística.

Para estimar os coeficientes de regressão logística dados por $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_8)^\top$ o método frequentista mais utilizado é o estimador de máxima verossimilhança. Para obtê-lo é preciso escrever a função de verossimilhança:

$$L(\boldsymbol{\beta} \mid \mathbf{y}) = \prod_{i=1}^{n} (p_i)^{y_i} (1 - p_i)^{1-y_i}, \quad i = 1, \ldots, n$$

em que $\mathbf{y} = (y_1, \ldots, y_n)$ e p_i é definida como acima. No modelo de regressão logística,

$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

em que η_i é o preditor linear do i-ésimo indivíduo (parto). Sendo assim, a estimativas

dos parâmetros são os valores de $\beta_0, \ldots, \beta_k$ que maximizam a função de verossimilhança. Em problemas como esse a maximização é feita utilizando algum método numérico, por exemplo o de Newton-Raphson.

2.2 Modelos Bayesianos de Classificação Binária

Como a inferência sobre os modelos baseados em processos gaussianos segue a metodologia bayesiana, é preciso estudar, inicialmente, os modelos bayesianos paramétricos. No estudo de Galo *et al.* (2023), foram obtidas as estimativas de Máxima Verossimilhança (EMV) dos parâmetros da regressão logística e as estimativas bayesianas, em que é preciso obter a distribuição a posteriori de Θ, o vetor de parâmetros do modelo, dada a amostra:

$$p_{\Theta|\mathcal{D}}(\theta \mid d) = \frac{p_\Theta(\theta)L(\theta \mid d)}{\int_\Theta p_\Theta(\theta)L(\theta \mid d)d\theta}$$

em que p_Θ é a distribuição a priori de Θ e $L(\theta \mid d)$ denota a função de verossimilhança para a amostra observada $\mathcal{D} = d$.

No artigo de referência, Galo *et al.* (2023), foi utilizado o método de Markov Chain Monte Carlo (MCMC) para estimar estocasticamente a distribuição a posteriori. Esse método consiste em criar uma cadeia de Markov onde uma sequência de amostras são geradas e a próxima amostra depende apenas da atual, e através de métodos de Monte Carlo[1] são estimadas quantidades de interesse da distribuição a posteriori como a média, a mediana e quantis para se obterem intervalos de credibilidade. Além disso, são aplicados métodos para verificar a convergência e garantir que a cadeia atinja seu estado estacionário e seja representativa da distribuição a posteriori.

Neste trabalho, para obter as estimativas bayesianas, aproximou-se a distribuição a posteriori pela normal multivariada. Essa é uma aproximação assintótica da posteriori, ou seja, dadas certas condições de regularidade, é possível demonstrar que a distribuição a posteriori dos parâmetros converge para uma normal multivariada de mesma dimensão com média dada pelos valores de Θ que maximizam a posteriori (moda a posteriori ou MAP) e matriz de variância dada pela matriz hessiana da log-posteriori avaliada no mesmo ponto.[2] Para obter a MAP foi utilizado o método numérico de Newton-Raphson, que utiliza o gradiente e a hessiana da log-posteriori (primeira e segunda derivadas parciais,

[1] Basicamente esses métodos são aplicações da lei dos grandes números a sequências de variáveis simuladas de modo a serem (aproximadamente) independentes e identicamente distribuídas.

[2] Ver Bernando e Smith (1994), seção 5.3.

respectivamente) para, iterativamente, se aproximar do valor do vetor de parâmetros que maximiza a log-posteriori (e consequentemente a posteriori).

Seguindo Galo *et al.* (2023), adotamos a priori dispersa para $\Theta = \boldsymbol{\beta} \sim N_9(0, 10^6\mathbb{I}_9)$.[3] Logo, como as condições de regulariadade são satisfeitas, a posteriori pode ser aproximada por uma normal multivariada com média $\hat{\theta}$, a moda a posteriori, e variância dada pelo inverso de $-H(\hat{\theta})$, o negativo da Hessiana da log-posteriori avaliada em $\hat{\theta}$. Para efeitos de comparação, os resultados obtidos via MCMC (disponíveis em Galo *et al.* (2023) e relatados na tabela 3.5) são comparados com os obtidos pela aproximação da posteriori pela normal multivariada, disponíveis na tabela 3.6 na seção 3.1.

2.3 Processos gaussianos

Um processo gaussiano (PG) é um processo estocástico − coleção de variáveis aleatórias $\{Z(t)\}_{t\in T}$ em que T é um conjunto de índices − caracterizado pelo fato de que, para qualquer subconjunto finito de índices, t_1, t_2, $\ldots t_n$, o vetor aleatório associado a eles ($Z(t_1)$, $Z(t_2)$, $\ldots Z(t_n)$) têm distribuição normal multivariada, dada por:

$$\begin{bmatrix} Z(t_1) \\ \vdots \\ Z(t_n) \end{bmatrix} \sim N\left(\begin{bmatrix} m(t_1) \\ \vdots \\ m(t_n) \end{bmatrix}, \begin{bmatrix} k(t_1,t_1) & \cdots & k(t_1,t_n) \\ \vdots & \ddots & \vdots \\ k(t_n,t_1) & \cdots & k(t_n,t_n) \end{bmatrix} \right)$$

em que $m(\cdot)$ é a função de média ($m : T \to \mathbb{R}$) e $k(\cdot,\cdot)$ a função de covariância do processo ($k : T \times T \to \mathbb{R}$), ou seja, fornece matrizes reais simétricas semi-definidas positivas. Assim, T pode ser um conjunto contínuo (como $\mathbb{R}$ ou $\mathbb{R}^c$, $c \in \mathbb{N}$), ou discreto (finito ou não), representando o espaço com todos os valores possíveis das variáveis explicativas.

Na abordagem Bayesiana, os PG's são usados como prioris sobre o espaço de funções de regressão, sendo atualizados por meio do Teorema de Bayes, ou seja, para se obter a distribuição a posteriori desse espaço de funções. Como método, eles podem ser classificados como semi-paramétrico (uma vez que as funções $m(\cdot)$ e $k(\cdot,\cdot)$ devem ser especificadas a priori), sendo mais flexíveis que os modelos paramétricos tradicionais, podendo obter um resultado com melhor performance preditiva que o modelo de regressão logística.

[3] Normal multivariada de dimensão 9, pois o modelo tem 8 covariáveis e um intercepto, com vetor de médias todas iguais a zero e matriz de variâncias com variâncias dos β's iguais a 10^6 e covariância zero entre eles.

2.4 Processos gaussianos em modelos de classificação binária

Para usar processos gaussianos em modelos de classificação binária adota-se um PG como priori sobre o espaço das funções $f(\mathbf{x})$, sendo $\mathbf{x}$ um vetor de covariáveis, e, usando a mesma ideia de modelos lineares generalizados, restringem-se os valores de f ao intervalo $(0, 1)$ por meio de uma função de distribuição como a função de distribuição da normal padrão, $\Phi(\cdot)$. Neste trabalho adota-se essa função com o objetivo de obter a probabilidade de um parto prematuro quando se conhecem apenas as covariáveis do indivíduo (mãe e/ou criança). Essa probabildade é denotada por $p\left(y_* = 1 \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right)$ em que $\mathbf{x}_*$ é o vetor de covariáveis de um indivíduo que não pertence à amostra usada para estimar o modelo (amostra de teste).

Como mencionado acima, a definição de PG não exclui processos com conjuntos de índices finitos, pois nesse caso eles são distribuições normais multivariadas. Neste projeto trabalhamos com $c \in \mathbb{N}$ covariáveis binárias de modo que $T = \{0, 1\}^c$.

Entretanto, apenas para ilustrar como seriam as possibilidades vamos utilizar um caso reduzido, em que consideramos que estivessem disponíveis, além da variável de interesse, apenas duas covariáveis: "Idade" e "Consultas pré-natal" disponíveis na tabela 3.1 na seção 3.1. Sejam: $\mathbf{y} = (y_1, \ldots, y_n) \in \{0, 1\}^n$, $x_1, \ldots, x_n \in T = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, ou seja, x_i é um vetor bi-variado com entradas podendo valer apenas 0 ou 1, cada uma delas indicando se no i-ésimo parto a mãe tinha mais de 35 anos (1) ou não (0) e se tinha realizado consultas pré-natal (1) ou não (0). Os valores que a função latente pode assumir podem ser reunidos em um vetor, $\mathbf{f} = [f(0, 0), f(0, 1), f(1, 0), f(1, 1)]$. Dessa forma, a função de verossimilhança é escrita como:

$$L(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} \left[\Phi\left(f\left(x_i\right)\right)\right]^{y_i} \left[1 - \Phi\left(f\left(x_i\right)\right)\right]^{1-y_i}$$

em que Φ é a função de distribuição da normal padrão. Repare que: (i) agora $p_i = \Phi\left(f\left(x_i\right)\right)$; e (ii) o vetor de parâmetros do modelo é dado por $\mathbf{f}$, e não mais $\boldsymbol{\beta}$, que podiam representar apenas as funções lineares enquanto $\mathbf{f}$ pode representar qualquer tipo de função.

Como PG a priori adotou-se, $\mathbf{f} \sim N_{|T|}(m, K)$, sendo a função de média constante e dada por $m = \Phi^{-1}(0.1)\mathbb{I}_{|T|}$, refletindo a ideia de que a probabilidade a priori de um nasci-

mento prematuro é de 10%; e a função de covariância $K = k(x_i, x_j) = \exp\left(-\frac{1}{2}|x_i - x_j|^2\right), i, j = 1, \ldots, n$, chamada de kernel gaussiano.[4]

A posteriori de $\mathbf{f}$ foi aproximada por uma normal multivariada e, para isso, mais uma vez foi utilizado o algoritmo de Newton-Raphson, já que é necessário encontrar a moda a posteriori. Para maximizar essa função foi preciso, como no caso da regressão logística bayesiana, encontrar o vetor gradiente, formado pelas primeiras derivadas parciais da mesma, dado por:

$$\nabla\Psi(\mathbf{f}) = \nabla \log L(\mathbf{y} \mid \mathbf{f}) - K^{-1}\mathbf{f}.$$

Para a primeira derivada da log-verossimilhança de um valor específico de f em x_i, $f_i = f(x_i)$, temos a seguinte expressão:

$$\frac{\partial}{\partial f_i} \log L(y_i|f_i) = \sum_{i=1}^{n} \left[y_i \frac{\phi(f(x_i))}{\Phi(x_i)} - (1 - y_i)\frac{\phi(f(x_i))}{1 - \Phi(x_i)} \right]$$

Após isso, é preciso calcular a matriz jacobiana, formada pelas segundas derivadas parciais, dada por:

$$\nabla\nabla\Psi(\mathbf{f}) = \nabla\nabla \log L(\mathbf{y} \mid \mathbf{f}) - K^{-1} = -W - K^{-1}$$

Sabemos que a matriz W (matriz jacobiana) é uma matriz diagonal, pois a segunda derivada da log-verossimilhança da distribuição de y_i depende apenas de $f_i = f(x_i)$, ou seja, não depende de f_j, com $j \neq i$. Sendo assim, em relação a outros elementos é zero.

Para a segunda derivada da log-verossimilhança de um valor específico de f em x_i temos a seguinte expressão:

$$\frac{\partial^2}{\partial f_i^2} \log L(y_i|f_i) = \sum_{i=1}^{n} \left[y_i \left(-f(x_i)\frac{\phi(f(x_i))}{\Phi(f(x_i))} - \frac{\phi(f(x_i))^2}{\Phi(f(x_i))^2} \right) - (1 - y_i) \left(-f(x_i)\frac{\phi(f(x_i))}{1 - \Phi(f(x_i))} + \frac{\phi(f(x_i))^2}{(1 - \Phi(f(x_i)))^2} \right) \right]$$

Com esses elementos, é possível implementar o algortimo de Newton-Raphson para maximizar a log-posteriori. Denotando-se por, $\mathbf{f}_\ell$ o valor de $\mathbf{f}$ obtido após a ℓ-ésima iteração do algoritmo, que é dado por:

[4] Lembrando que $|T|$ denota o número de elementos do conjunto $T = \{0, 1\}^c$.

$$\mathbf{f}_\ell = \mathbf{f}_{\ell-1} + (K^{-1} + W)^{-1}(\nabla \log L(\mathbf{y} \mid \mathbf{f}_{\ell-1})) - K^{-1}(\mathbf{f}_{\ell-1} - m)$$

em que W é a matriz descrita acima obtida na iteração $\ell - 1$. Dessa forma, encontramos o valor de $\mathbf{f}$ que maximiza a posteriori, $\hat{\mathbf{f}}$, e a aproximação de Laplace, baseada na normal multivariada, para a posteriori da seguinte forma:

$$\mathbf{f} \mid \mathbf{y}, \mathbf{x} \sim q = N_{|T|}\left(\hat{\mathbf{f}}, \left(K^{-1} + W\right)^{-1}\right)$$

em que $\hat{\mathbf{f}} = \text{argmax}_{\mathbf{f}}\, p(\mathbf{f} \mid X, \mathbf{y})$, a moda da posteriori de $\mathbf{f}$, e $W = -\nabla\nabla \log p(\mathbf{y} \mid \mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$ é Hessiana do negativo da log verossimilhança nesse ponto.

Deseja-se estudar a capacidade preditiva desses modelos, isto é, dado um vetor $\mathbf{x}_*$ de indivíduos (partos) não pertencentes à amostra de teste, calcular $\pi_* = p\left(y_* = 1 \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right)$, ou seja, a probabilidade desse parto ser prematuro. Para isso é necessário obter a preditiva a posteriori de $f_* = f\left(\mathbf{x}_*\right)$:

$$p\left(f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right) = \int_{\mathbb{R}^{|T|}} p\left(f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*, \mathbf{f}\right) \cdot p(\mathbf{f} \mid \mathbf{y}, \mathbf{x})\; d\mathbf{f}$$

e, usando essa posteriori preditiva sobre f_*, obtem-se a previsão probabilística

$$\pi_* = \int_{\mathbb{R}} \Phi\left(f_*\right) \cdot p\left(f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right) df_*$$

A integral acima foi calculada considerando a aproximação da posteriori pela normal (denotada por q) indicada acima, e será denotada por $\overline{\pi}_*$. Para isso é preciso obter a média e a variância a posteriori para f_* dadas respectivamente por:

$$E_q\left[f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right] = \mathbf{k}_* K^{-1} \hat{\mathbf{f}}$$

e

$$\text{var}_q\left[f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right] = 1 - {\mathbf{k}_*}^\top (K + W^{-1})^{-1} \mathbf{k}_*$$

em que $\mathbf{k}_* = k\left(\mathbf{x}^*, \mathcal{X}\right)$ é o vetor de covariâncias entre o ponto da amostra de teste $\mathbf{x}^*$ e todos os vetores de covariáveis (da amostra de treinamento), $\mathcal{X}$.

Com esses resultados é possível obter a previsão do modelo de processos gaussianos

que segue a seguinte distribuição acumulada da normal padrão:[5]

$$\overline{\pi}_* \simeq E_q\left(\pi_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right) = \int_{\mathbb{R}} \Phi\left(f_*\right) \cdot q\left(f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right) df_* = \Phi\left(\frac{E_q\left[f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right]}{\sqrt{1 + \operatorname{var}_q\left(f_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_*\right)}}\right) \tag{2.1}$$

em que o subscrito q indica que a esperança e a variância são calculadas em relação à distribuição q.

2.5 Support Vector Machine

O Support Vector Machine (SVM) é um método de aprendizado de máquina utilizado principalmente para estudos de classificação. Esse classificador assume que é possível encontrar um hiperplano que separa perfeitamente as observações no conjunto de treinamento com base nos valores da variável-alvo, ou seja, busca encontrar a melhor fronteira possível que separa os dados de diferentes classes.

Dessa forma, aplicando ao nosso conjunto de dados, o objetivo é obter a melhor separação das classes, em que nesta seção, $y_i = 1$ indica que o i-ésimo nascimento foi prematuro e $y_i = -1$ indica que o i-ésimo nascimento foi a termo. Se essas classes forem separáveis então existe um hiperplano, ou seja, um conjunto dessa forma:

$$\mathbf{z} \in \mathbb{R}^c : \beta_0 + \beta_1 z_1 + \ldots + \beta_c z_c = 0$$

Em que para cada observação no conjunto de treinamento $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, tem-se que:

$$h(x) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_c x_{ic} > 0 \quad \text{se} \quad y_i = 1,$$
$$\text{e} \quad h(x) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_c x_{ic} < 0 \quad \text{se} \quad y_i = -1.$$

Ou seja, a observação de teste $\mathbf{x}_*$ é classificada com base no sinal da função h avaliada nesse ponto. Se $h(\mathbf{x}_*)$ for positivo, a observação é classificada como pertencendo à classe 1 e se $h(\mathbf{x}_*)$ for negativo, a observação é classificada como pertencendo à classe -1. Sendo assim, se $h(\mathbf{x}_*)$ está distante de zero, significa que $\mathbf{x}_*$ está longe do hiperplano e há maior

[5] No apêndice está o cálculo dessa equação resolvido analiticamente.

segurança sobre a classe que foi atribuída. Por outro lado, se $h(\mathbf{x}_*)$ está perto de zero, significa que $\mathbf{x}_*$ está próximo do hiperplano e possui menos certeza sobre a classificação das classes (James *et al.* (2023)).

Para construir $h(x)$, é preciso supor que as observações são linearmente separáveis, ou seja, existe um hiperplano que separa perfeitamente todas as observações do conjunto de treinamento de acordo com a classe. Sendo assim, abaixo segue uma imagem para ilustrar essa situação:

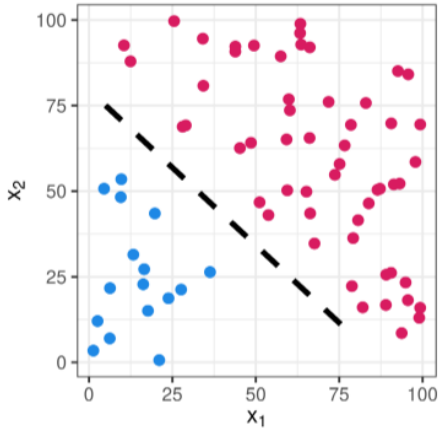

Figura 2.1: Exemplo de conjunto de dados no qual existe um hiperplano que os separa perfeitamente bem. Fonte: Izbicki e dos Santos (2020).

Um hiperplano separador tem a propriedade de que como as classes são separáveis, pode-se encontrar uma função $h(x)$ com $y_i h(x_i) > 0$ para todo $i = 1, 2, \ldots, n$. O classificador de margem máxima busca então encontrar β_0 e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_c)$ que definem o hiperplano que maximiza a margem — a distância entre os pontos mais próximos das classes de treinamento com rótulos 1 e -1. Isso é feito resolvendo o seguinte problema de otimização:

$$\underset{\beta_0,\ \boldsymbol{\beta},\ M}{\text{maximizar}}\ M \text{ sujeito a } \sum_{i=1}^{c} \beta_i^2 = 1 \text{ e } y_i \cdot h(\mathbf{x}_i) \geq M \text{ para } i = 1, \ldots, n, \tag{2.2}$$

onde M é a margem a ser maximizada.

A primeira restrição, $\sum_{i=1}^{c} \beta_i^2 = 1$, exige que o vetor $\boldsymbol{\beta}$ tenha norma unitária. Isso não é uma restrição sobre o hiperplano em si, pois se $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_c x_{ic} = 0$ define um hiperplano, então qualquer multiplicação por um escalar $\gamma(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_c x_{ic}) = 0$ com $\gamma \neq 0$ também o define. No entanto, incluir essa restrição permite uma interpretação

específica: pode-se demonstrar[6] que a distância perpendicular do i-ésimo ponto de amostra ao hiperplano é precisamente $y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_c x_{ic}) = y_i \cdot h(\mathbf{x}_i)$.

Assim, essas restrições garantem que cada observação esteja do lado correto do hiperplano, cada uma a pelo menos uma distância M dele. Essa distância M é a margem, que é maximizada ao escolher os valores ótimos de β_0 e $\boldsymbol{\beta}$. Dado que a primeira restrição tem um papel secundário, podemos reformular o problema acima sem a restrição de norma unitária sobre $\boldsymbol{\beta}$ da seguinte forma:

$$\underset{\boldsymbol{\beta}}{\text{minimizar}} \; ||\boldsymbol{\beta}|| \quad \text{sujeito a} \quad y_i \cdot h(\mathbf{x}_i) \geq 1 \quad \text{para} \quad i = 1, \ldots, n,$$

onde $||\boldsymbol{\beta}||$ é a norma L_2 de $\boldsymbol{\beta}$. Note que essa formulação implica $M = 1/||\boldsymbol{\beta}||$.

O classificador de margem máxima fornece uma abordagem intuitiva para classificação quando um hiperplano pode separar completamente as classes, o que geralmente não ocorre em conjuntos de dados reais.

Quando as classes se sobrepõem no espaço das covariáveis, ainda podemos tentar maximizar M, mas permitimos que alguns pontos fiquem do lado incorreto da margem. Isso ocorre pois diferentemente da situação anterior, em que havia uma separação perfeita das classes e era possível encontrar um hiperplano separador, nesse caso, as classes podem não ser separáveis por uma fronteira linear. Para isso, introduzimos variáveis auxiliares não negativas $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$, conhecidas como variáveis de folga (*slack variables*). Isso modifica o problema para:

$$\underset{\beta_0,\, \boldsymbol{\beta},\, M,\, \varepsilon}{\text{maximizar}} \; M \text{ sujeito } \sum_{i=1}^{c} \beta_i^2 = 1, y_i \cdot h(\mathbf{x}_i) \geq M(1 - \varepsilon_i) \text{ e } \sum_{i=1}^{n} \varepsilon_i \leq \text{constante}, \varepsilon_i \geq 0, \tag{2.3}$$

para $i = 1, \ldots, n$. A variável ε_i na restrição $y_i \cdot h(\mathbf{x}_i) \geq M(1 - \varepsilon_i)$ tem a seguinte interpretação: se $\varepsilon_i = 0$, o i-ésimo ponto de amostra está corretamente classificado e do lado correto da margem; se $0 < \varepsilon_i \leq 1$, o ponto está dentro da margem mas ainda corretamente classificado; e se $\varepsilon_i > 1$, o ponto é mal classificado. Assim, se $y_i \cdot h(\mathbf{x}_i) < 0$, essa restrição implica $M(1 - \varepsilon_i) \leq y_i \cdot h(\mathbf{x}_i) < 0$, pois $M > 0$.

Portanto, a restrição $\sum_{i=1}^{n} \varepsilon_i \leq$ constante, garante que tanto o número de erros de classificação quanto o total pelo qual os pontos estão no lado incorreto da margem sejam limitados por uma constante escolhida. A restrição de norma unitária sobre $\boldsymbol{\beta}$ pode ser

[6] Veja (Friedman *et al.*, 2009, sec. 4.5).

novamente removida, levando a uma reformulação do problema como:

$$\underset{\boldsymbol{\beta}}{\text{minimizar}}\ ||\boldsymbol{\beta}|| \quad \text{sujeito a} \quad y_i \cdot h(\mathbf{x}_i) \geq 1 - \varepsilon_i, \quad \sum_{i=1}^{n} \varepsilon_i \leq \text{constante}, \quad \varepsilon_i \geq 0, \tag{2.4}$$

para $i = 1, \ldots, n$. Conhecido como o classificador de margem suave (*soft margin classifier*), pois permite que alguns pontos violem a restrição de margem. O problema (2.4) é quadrático com restrições de desigualdade linear, sendo portanto um problema de otimização convexa. Dessa forma, a solução é via programação quadrática utilizando, por exemplo, métodos de multiplicadores de Lagrange, podendo ser reescrito como:

$$\underset{\boldsymbol{\beta}}{\text{minimizar}}\ \frac{1}{2}||\boldsymbol{\beta}||^2 + Q\sum_{i=1}^{n} \varepsilon_i, \quad \text{sujeito a} \quad \varepsilon_i \geq 0 \quad \text{e} \quad y_i \cdot h(\mathbf{x}_i) \geq 1 - \varepsilon_i, \tag{2.5}$$

para $i = 1, \ldots, n$. Onde o parâmetro Q substitui o termo "constante" no problema anterior (2.4). Além disso, o parâmetro $Q > 0$ controla o equilíbrio entre maximizar a margem e permitir erros de classificação.

O parâmetro Q atua como um peso para a penalização de pontos mal classificados ou que ficam dentro da margem. Valores altos de Q impõem uma penalização severa a essas violações, forçando o modelo a buscar uma separação quase perfeita dos dados de treinamento. Dessa forma, isso pode reduzir o viés, mas aumentar a variância, e consequentemente, levar ao *overfitting* (sobreajuste dos dados), especialmente na presença de ruído. Já valores baixos de Q indicam pequena penalização, ou seja, o modelo permite mais erros, favorecendo margens mais largas e soluções mais regulares, que tendem a generalizar melhor. Essa formulação define o Support Vector Classifier ou Linear Support Vector Machine, sendo o caso separável um limite quando $Q \to \infty$.

O classificador de vetores de suporte tem bom desempenho quando a fronteira entre as classes é linear, mas, em muitas aplicações, essas fronteiras são não lineares. Para lidar com isso, Boser *et al.* (1992) aplicaram os princípios do classificador de vetores de suporte a uma transformação do espaço original das covariáveis, obtendo separação (suave) linear nesse espaço transformado. Portanto, deve-se escolher uma função de transformação adequada $\phi : \mathcal{X} \to \mathcal{F}$, de modo que os rótulos das classes se tornem linearmente separáveis no espaço transformado $\phi(\mathcal{X})$. O contradomínio de ϕ, denotado por $\mathcal{F}$, é conhecido como

espaço de características (*feature space*), e pode ter alta dimensão.

Dada uma transformação ϕ, podemos definir $h(\mathbf{x}_i) = \beta_0 + \phi(\mathbf{x}_i)^\top \beta$, onde $\phi(\mathbf{x}_i) = (\phi_1(\mathbf{x}_i), \ldots, \phi_d(\mathbf{x}_i))$, com $\phi_j(\mathbf{x})$ sendo funções base. Aqui, $\beta = (\beta_1, \ldots, \beta_d)$, e d é a dimensão do espaço de características. Com essa nova definição de h, é possível resolver o problema (2.5) para obter o classificador utilizando a Lagrangiana Primal dada da seguinte forma:

$$\mathcal{L}(\beta_0, \boldsymbol{\beta}, \varepsilon) = \frac{1}{2}||\boldsymbol{\beta}||^2 + Q\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\alpha_i\left[y_i \cdot h(\mathbf{x}_i) - (1-\varepsilon_i)\right] - \sum_{i=1}^{n}\gamma_i\varepsilon_i \tag{2.6}$$

onde α_i e γ_i são Multiplicadores de Lagrange. Definindo as derivadas como zero e obtendo que as condições de primeira ordem para minimizar (2.6) são:

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \tag{2.7}$$

$$\beta = \sum_{i=1}^{n}\alpha_i \cdot y_i \cdot \phi(\mathbf{x}_i) \tag{2.8}$$

$$\alpha_i = Q - \gamma_i \tag{2.9}$$

$$\alpha_i, \gamma_i, \varepsilon_i \geq 0, \tag{2.10}$$

para $i = 1, 2, \ldots, n$. Substituindo essas expressões em (2.6) obtemos a função de Lagrange Dual, em que é mais eficiente de resolver:

$$\mathcal{L}_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j \cdot y_i y_j \cdot \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

e, dado um Kernel, em que k deve ser simétrica e positiva semi-definida podemos reescrever como:

$$\mathcal{L}_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j \cdot y_i y_j \cdot k(\mathbf{x}_i, \mathbf{x}_j). \tag{2.11}$$

A equação (2.11) fornece um limite inferior para a função objetivo do problema (2.5) para qualquer ponto viável. Além disso, é quadrática e convexa, permitindo usar algoritmos de programação quadrática para encontrar a solução. Maximizando $\mathcal{L}_D$ sujeito a $0 \leq \alpha_i \leq Q$ e $\sum_{i=1}^{n}\alpha_i y_i = 0$, as condições de Karush-Kuhn-Tucker são:

$$\alpha_i \left[y_i \cdot h(\mathbf{x}_i) - (1 - \varepsilon_i) \right] = 0 \tag{2.12}$$

$$\gamma_i \varepsilon_i = 0 \tag{2.13}$$

$$y_i \cdot h(\mathbf{x}_i) - (1 - \varepsilon_i) \geq 0, \tag{2.14}$$

para $i = 1, 2, \ldots, n$, além das equações (2.7) − (2.10). As equações (2.7) − (2.14) caracterizam a solução das formas primal e dual. A partir da equação (2.8) temos que a solução para β é:

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(\mathbf{x}_i)$$

e, consequentemente,

$$\hat{h}(\mathbf{x}) = \hat{\beta}_0 + \phi(\mathbf{x})^\top \hat{\beta} = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \cdot y_i \cdot \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \cdot y_i \cdot k(\mathbf{x}, \mathbf{x}_i) \tag{2.15}$$

onde $\hat{\alpha}_i$ são diferentes de zero apenas para as observações i que satisfazem a equação (2.14).

Os vetores de suporte são os pontos que estão na borda da margem e os pontos que estão do lado errado da margem, assim, apenas esses pontos são os que determinam o hiperplano, a maioria dos outros pontos não influenciam diretamente na solução. Essas observações compõem os vetores de suporte, já que $\hat{\alpha}_i$ é escrito em termos delas.

Além disso, $\langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle$ representa o produto interno $\phi(\mathbf{x})^\top \cdot \phi(\mathbf{x}_i)$, e os coeficientes $\hat{\alpha}_i$ são derivados da formulação dual do problema (2.5). À primeira vista, essa abordagem parece exigir conhecimento explícito de ϕ, mas ele aparece no produto interno, e por isso, o conhecimento de ϕ é evitado por meio de uma técnica conhecida como truque do kernel (*kernel trick*). Para compreendê-la, primeiro introduziremos os kernels.

Definição [Mercer kernels]. Uma função com valores reais $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ é chamada de kernel se satisfaz as seguintes propriedades:

1. $k(\mathbf{u}, \mathbf{v}) = k(\mathbf{v}, \mathbf{u})$ para todo $\mathbf{u}, \mathbf{v} \in \mathcal{X}$ (ou seja, é simétrica), e

2. a matriz $[k(\mathbf{v}_i, \mathbf{v}_j)]_{i,j=1}^{\ell}$ é semi-definida positiva para qualquer escolha de $\mathbf{v}_1, \ldots, \mathbf{v}_\ell \in$

$\mathcal{X}$ and $\ell \in \mathbb{N}$.

O truque do kernel é baseado no seguinte teorema.

Teorema [Truque do Kernel].[7] Se k é um kernel, então existe um espaço com produto interno $\mathcal{F}$ e uma função de transformação ϕ tal que $k(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle$ para qualquer $\mathbf{u}, \mathbf{v} \in \mathcal{X}$.

Como a equação (2.15) mostra que o classificador depende de ϕ apenas por meio de seu produto interno, o teorema do truque do kernel garante que especificar uma função kernel é equivalente a definir implicitamente a função de mapeamento ϕ.

Assim, o classificador baseia-se na equação (2.15): para uma amostra de teste $\mathbf{x}_*$, classifica-se como classe 1 se $\hat{h}(\mathbf{x}_*) > 0$ e como classe -1 se $\hat{h}(\mathbf{x}_*) < 0$. Usando o truque do kernel, isso pode ser expresso como:

$$\hat{C}(\mathbf{x}_*) = \text{sign}\left[\hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \cdot y_i \cdot k(\mathbf{x}_*, \mathbf{x}_i)\right]. \tag{2.16}$$

Como mostrado na equação (2.16), os SVMs produzem rótulos previstos, mas não fornecem diretamente saídas probabilísticas. Para estimar probabilidades de classe, utilizamos a *Platt's scaling* (Platt, 1999), que aplica uma transformação logística aos escores do classificador. Essa abordagem envolve o cálculo dos parâmetros A e B tais que:

$$P(y^* = 1 \mid \mathbf{x}_*) = \frac{1}{1 + \exp\left(A\hat{h}(\mathbf{x}_*) + B\right)}, \tag{2.17}$$

onde $\mathbf{x}_*$ é o vetor de covariáveis de uma observação de teste. Na prática A e B são estimados ajustando-se um modelo de regressão logística sobre os escores do classificador, $\hat{h}(\cdot)$, obtidos no conjunto de treinamento, utilizando o método da máxima verossimilhança.

Dessa forma, temos que o SVM passa a ser capaz de aprender decisões de fronteiras não lineares, mantendo as garantias teóricas e a robustez do caso linear. As funções kernels mais utilizadas são:

- **Linear:** $k(x_i, x_j) = \langle x_i, x_j \rangle$
- **Polinomial:** $k(x_i, x_j) = (\gamma \langle x_i, x_j \rangle + c)^d$

[7] (Izbicki e dos Santos, 2020, p. 69)

- **Gaussiano ou** *Radial Basis Function* **(RBF):** $k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$

A escolha dos hiperparâmetros está relacionada com a perfomance do modelo SVM, tanto o hiperparâmetro Q que controla a complexidade do modelo e o erro de treinamento, quanto os hiperparâmetros das funções kernels.

Assim, esses valores podem ser escolhidos através da validação cruzada, como o *k-fold cross-validation*, em que a ideia é dividir o conjunto de dados em k subconjuntos de tamanhos aproximadamente iguais. Em cada uma das k iterações, utilizamos $k-1$ desses subconjuntos para treinar o modelo e o subconjunto restante para validar. No final, o erro de validação é calculado como a média dos erros obtidos em cada uma das iterações. Esse procedimento será detalhado na seção a seguir.

2.6 Validação Cruzada

A validação cruzada é uma técnica de reamostragem, em que é possível extrair diversas amostras diferentes do conjunto de treinamento e ajustar um modelo em cada uma delas, permitindo obter mais informações e podendo comparar os resultados das estimativas. Assim, fazer a reamostragem permite obter resultados mais precisos do que realizar apenas uma estimativa, obtida com o conjunto de treinamento completo.

Para esse método é preciso entender os conceitos de erro de teste e de erro de treinamento. O erro de teste, ou erro de generalização, é o erro de predição de uma nova observação em uma amostra de teste independente e que não foi usada no treinamento do modelo, ou seja, é desejado que o erro de teste seja baixo. O erro de treinamento é a perda (ou *escore*) média sobre a amostra de treinamento e não é uma boa estimativa do erro de teste, pois quando a complexidade do modelo aumenta, o modelo tende a descrever bem os dados de treinamento (reduzindo o viés e aumentando a variância), e com isso, o erro de treinamento diminui, mas está superajustado (*overfitting*) podendo apresentar baixa capacidade de generalização e erro de teste superior.

Dessa forma, a validação cruzada busca estimar o erro de teste retirando um subconjunto das observações de treinamento do processo de estimação, ou seja, utiliza o máximo de dados possível para o treinamento em cada repetição, reduzindo a variância da estimativa do erro de generalização.

Quando estimamos o erro de teste para avaliar modelos preditivos temos o objetivo de selecionar um modelo, ou seja, escolher aquele que apresenta o melhor desempenho

preditivo, segundo uma métrica, que é especificada por uma função de perda ou escore.

Existem diversos métodos para estimar o erro de teste. Um modo eficiente de resolver esse problema é dividir o conjunto de observações disponíveis em dois subconjuntos de tamanhos parecidos: o conjunto de treinamento, usado para estimar o modelo, e o conjunto de validação, reservado para avaliar a performance do modelo ajustado, cujas estimativas se baseiam apenas no conjunto de treinamento.

Nesse caso, para problemas de classificação, a taxa de erro no conjunto de validação é usada para estimar o erro de teste, e para problemas com respostas quantitativas pode-se utilizar o Erro Quadrático Médio (MSE, de *Mean Squared Error*). Esse método possui duas limitações: a estimativa do erro de teste depende das observações que foram escolhidas no conjunto de treinamento e no conjunto de validação, podendo ter resultados variáveis. Além disso, quando há poucas observações no conjunto de treinamento a taxa de erro do conjunto de validação pode superestimar a taxa de erro de teste do modelo ajustado com todo o conjunto de dados.

Outro método utilizado é a Validação Cruzada *Leave-One-Out* (LOOCV), que busca melhorar o método anterior. O conjunto de observações é dividido utilizando apenas uma observação para o conjunto de validação e o restante para o conjunto de treinamento, ou seja, o modelo é ajustado usando $n - 1$ observações. A observação que não é utilizada no conjunto de treinamento é utilizada para calcular o MSE, sendo uma estimativa não viesada mas muito variável, pois depende apenas de uma observação. Dessa forma, esse procedimento é repetido n vezes até passar por todas as observações, e com isso, obtemos $MSE_1, MSE_2, \ldots, MSE_n$ e podemos calcular a média deles. Essas estimativas possuem vantagens em relação ao método anterior: possui um menor viés, pois nos conjuntos de treinamento possuem $n - 1$ observações, permitindo estimar o modelo com quase todos os dados, e com isso, o modelo tende a não ter *overfitting*, e além disso, como não possui aleatoriedade nas divisões permite que os resultados sejam mais confiáveis.

O método de Validação Cruzada *k-fold* possui a mesma ideia do método anterior mas a divisão é feita em k grupos ou *folds* de tamanhos próximos. Dessa forma, o primeiro grupo é o conjunto de validação e os $k - 1$ grupos restantes compõem o conjunto de treinamento. O MSE é então calculado para o respectivo conjunto de validação. Esse procedimento é realizado k vezes, ou seja, é possível perceber que o método LOOCV é um caso particular do *k-fold*, em que $k = n$. Logo, validação cruzada *k-fold* possui uma vantagem computacional uma vez que não é preciso estimar o modelo n vezes, o que é

vantajoso principalmente para amostras com muitas observações.

O valor que se escolhe para k influencia o viés e a variância do modelo. Para o LOOCV temos que $k = n$ indicando que o viés será baixo, pois o modelo é estimado com quase todos os dados disponíveis, mas a variância tende a ser alta devido a cada conjunto de treinamento ser diferente por apenas uma observação levando a erros altamente correlacionados e a média deles terá alta variância. Como a validação cruzada *k-fold* possui $k < n$, os modelos são estimados com menos dados levando a estimativas do MSE que têm maior viés. No entanto, como os conjuntos de treinamento são menos semelhantes, o método gera estimativas do MSE que são menos correlacionadas entrei si, ou seja, têm menor variância (James *et al.* (2023)).

Para problemas de classificação, em que a variável resposta é qualitativa, como no caso desse estudo, a validação cruzada pode ser utilizada da mesma forma, mas no lugar de usar o MSE podemos utilizar o número de observações classificadas incorretamente para estimar o erro de teste ou qualquer outra métrica, como a AUC e a acurácia, que foi a métrica utilizada na validação cruzada do modelo SVM.

Portanto, utilizando a validação cruzada *k-fold* temos que o erro de treinamento tende a diminuir à medida que a complexidade do modelo aumenta, enquanto o erro de teste tende a diminuir até atingir um mínimo e depois volta a aumentar, lembrando o formato de uma curva em forma de "U". Dessa forma, temos que o erro de treinamento não é indicado para escolher o valor ótimo. Nesse caso, é preferível optar pelo erro de teste, pois ele atinge um mínimo muito próximo do melhor valor.

Para o método validação cruzada *k-fold* temos a estimativa do erro de predição de modelo parametrizado como $y = f(x, \alpha)$ em que x indica o conjunto de covariáveis utilizados no modelo e α um vetor de hiperparâmetros. Denotamos por $\hat{f}^{-k}(x, \alpha)$ o valor previsto pelo modelo considerando o vetor de hiperparâmetros α e estimado com uma amostra que não contém a k-ésima parte ou *fold* dos dados. Dessa forma, definimos:

$$CV(\hat{f}, \alpha) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-k(i)}(x_i, \alpha))$$

como a estimativa do erro de teste. O valor do vetor de hiperparâmetros α que minimiza essa estimativa é o utilizado na estimativa calculada usando todos os dados (Friedman *et al.* (2009)).

2.6.1 A validação cruzada e a escolha dos hiperparâmetros

Como mencionado acima, a validação cruzada pode ser utilizada também para a escolha dos valores dos hiperparâmetros. Nesse trabalho vamos utilizar esse procedimento para obter o melhor valor de dois hiperparâmetros, o hiperparâmetro Q, utilizado como penalização de pontos mal classificados ou que ficam fora da margem no SVM, ou seja, controla o equilíbrio entre maximizar a margem e permitir erros de classificação e o hiperparâmetro γ do kernel Gaussiano (RBF).

Dessa forma, vamos exemplificar os detalhes da validação cruzada *k-fold*, deixando mais claro as etapas desse método. Supondo um conjunto de dados de $n = 5000$ e $k = 5$, ou seja, 5 *folds* e separando 70% para treinamento e 30% para teste, temos os seguintes conjuntos: 3500 observações para treinamento e 1500 observações para teste. Aplicando a validação cruzada no conjunto de treinamento temos que as 3500 observações são divididas em 5 *folds*, ou seja, possuimos *folds* de tamanho igual a 700 observações. Sendo assim, o modelo é ajustado com $k - 1 = 4$ grupos ou *folds*, ou seja, 2800 observações e a validação é feita usando as outras 700 observações separadas e utilizadas para calcular a taxa de erro. Esse procedimento é realizado para os 5 *folds*, obtendo-se 5 estimativas da taxa de erro, e através disso podemos calcular a função $CV(\hat{f}, \alpha)$ com a média das 5 taxas de erros.

Para encontrar os melhores valores dos hiperparâmetros, a função $CV(\hat{f}, \alpha)$ é calculada para cada combinação de hiperparâmetros, pois no caso desse trabalho temos dois hiperparâmetros (por exemplo, para 5 valores de Q e 5 valores de γ temos 25 combinações para testar). Escolhemos a combinação $\hat{\alpha}$ que minimiza a função $CV(\hat{f}, \alpha)$.

Após escolher os melhores hiperparâmetros o modelo final $f(x, \hat{\alpha})$ é novamente estimado em todo o conjunto de treinamento (3500 observações) e avaliado no conjunto de teste (1500 observações), fornecendo uma estimativa final do desempenho com observações fora da amostra de treinamento.

Portanto, para esse exemplo com $n = 5000$ observações é possível obter um bom resultado, mas se o número de observações fosse menor, poderia superestimar o erro verdadeiro. Além disso, com $k = 5$ obtemos um equilíbrio entre o viés e a variância para a taxa de erro estimada.

2.7 Métricas

Para comparação dos modelos estudados vamos utilizar três métricas para analisar a perfomance preditiva, sendo AUC (área sob a curva ROC), escore de Brier e escore logarítmico, para assim, obter o melhor modelo para esse conjunto de dados.

2.7.1 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) é uma ferramenta gráfica utilizada para avaliar o desempenho de classificadores binários, sendo construída variando o limiar de decisão para classificar uma categoria como "positivo" (rótulo 1) ou "negativo" (rótulo 0). Essa medida relaciona a taxa de verdadeiros positivos com a taxa de falsos positivos. Essas métricas são dadas por:

- Taxa de Verdadeiros Positivos (TVP) ou Sensibilidade: mede a proporção de positivos corretamente identificados.

$$TVP = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}}$$

- Taxa de Falsos Positivos (TFP): mede a proporção de negativos incorretamente classificados como positivos.

$$TFP = \frac{\text{Falsos Positivos}}{\text{Falsos Positivos} + \text{Verdadeiros Negativos}}$$

A curva ROC é um gráfico, em que a TFP está no eixo horizontal e a TVP está no eixo vertical. Além disso, cada limiar define o ponto a partir do qual uma observação é classificada como positiva ou negativa. Um modelo que prevê perfeitamente todas as instâncias em todos os limiares tem TVP igual a 1 e TFP igual a 0 em todos os pontos. Entretanto, um classificador completamente aleatório, que prevê as classes positiva e negativa com probabilidades iguais, produz uma curva ROC idêntica à diagonal que vai de (0, 0) até (1, 1), ou seja, em um classificador aleatório, a TVP é igual a TFP pois ela produz o mesmo número de verdadeiros e falsos positivos em qualquer limiar.

Dessa forma, quanto mais próxima a curva ROC estiver do canto superior esquerdo, melho será a capacidade do modelo de distinguir os rótulos. A imagem a seguir ilustra a curva ROC de um classificador binário.

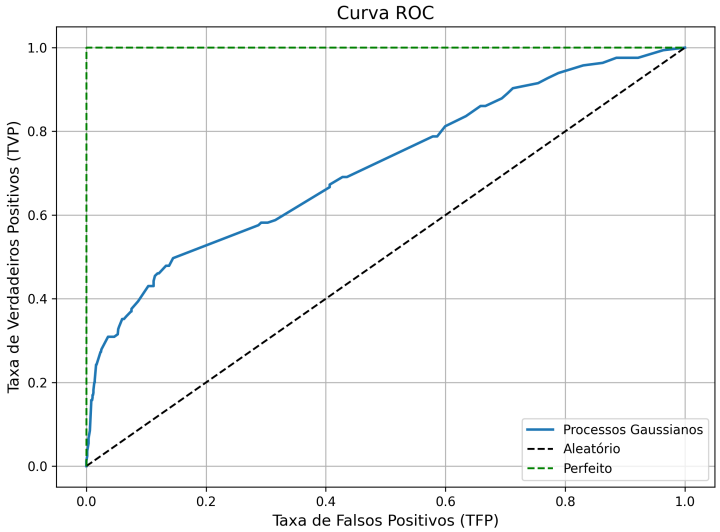

Figura 2.2: Exemplo de uma Curva ROC.

A AUC (*Area Under the Curve* é a área sob a curva ROC que resume em um número, entre 0 e 1, a capacidade do modelo de discriminar duas classes, sendo que:

- AUC $= 1$: indica um classificador ótimo;
- AUC $= 0.5$: indica um classificador aleatório;
- AUC < 0.5: indica um classificador pior que o acaso, ou seja, é melhor inverter os rótulos preditos.

Dessa forma, quanto mais próximo de 1 melhor é a capacidade preditiva do modelo. Essa medida é interessante pois não precisamos escolher um valor de limiar de decisão, pois avalia o desempenho global, sem se restringir a um valor específico.

2.7.2 Escore de Brier

O escore de Brier é dado por:

$$BS = \frac{1}{n^*} \sum_{i=1}^{n^*} (p_i^* - y_i^*)^2,$$

em que, p_i^* é a previsão probabilística, y_i^* é o resultado observado do indivíduo i (0 se o evento não ocorre e 1 se o evento ocorre) e n^* é o tamanho da amostra de teste.

O escore de Brier pode ser interpretado como o erro quadrático médio da previsão, oferecendo uma maneira simples de avaliar a precisão de previsões probabilísticas.

2.7.3 Escore Logarítmico

O escore logarítmico, ou de entropia cruzada, é dado por:

$$LS = -\frac{1}{n^*} \sum_{i=1}^{n^*} \left[y_i^* \ln(p_i^*) + (1 - y_i^*) \ln(1 - p_i^*) \right],$$

sendo calculado com o logaritmo da probabilidade atribuída ao resultado observado. A interpretação desse escore é baseado na teoria da informação: dado que $p_i^* = P(y_i^* = 1 \mid \mathbf{x}^*)$, o valor esperado de LS pode ser escrito como $-\frac{1}{n^*} \sum_{i=1}^{n^*} p_i^* \ln(p_i^*)$, o que está relacionado intimamente à entropia das probabilidades previstas. Uma limitação dessa métrica é que ela é indefinida quando p_i^* é exatamente 0 ou 1.

O escore de Brier e o escore logarítmico são métricas utilizadas para avaliar de maneira quantitativa a qualidade das previsões em problemas de classificação binária. Essas métricas medem o quão próximas as probabilidades previstas estão nos rótulos reais das classes.

Os valores que essas duas métricas podem assumir também estão entre 0 e 1, entretanto, quanto mais próximo de 0 melhor é o desempenho do modelo, diferentemente da AUC.

O escore logarítmico penaliza mais fortemente previsões erradas com alta confiança, pois utiliza o logaritmo das probabilidades. Além disso, é sensível à calibração das probabilidades, diferentemente da AUC que mede a capacidade da separação das classes, mas não considera a calibração.[8]

[8] Um modelo tem previsões bem calibradas quando, ao atribuir uma probabilidade p ao rótulo 1, essa proporção é observada em aproximadamente $p\%$ das observações.

Capítulo 3

Resultados

Nesse capítulo, na seção 3.1 será abordado sobre o conjunto de dados de nascimentos de bebês prematuros com uma análise descritiva das variáveis, mostrando as proporções da variável de interesse "Prematuridade"de acordo com as categorias de cada covariável binária. Além disso, será feito o teste χ^2 de independência entre a variável de interesse e as covariáveis binárias para verificar se há associação entre elas.

Na seção 3.2 serão abordados os resultados preliminares que já foram obtidos na Iniciação Científica Sartori (2024) para as estimativas da regressão logística por Máxima Verossimilhança e Bayesiana.

Para finalizar, na seção 3.3 serão abordadas as métricas de comparação de perfomance preditiva, como: AUC, escore de Brier e escore logarítmico, calculadas para a regressão logística, os processos gaussianos e o SVM, permitindo comparar esses modelos.

3.1 Conjunto de dados

Conforme mencionado anteriormente, o objetivo final do trabalho é estimar um método de aprendizado de máquina de classificação binária, mais especificamente, Support Vector Machine (SVM) considerando como covariáveis atributos da mãe, do feto e do parto para comparar a performance preditiva desse modelo com a dos modelos paramétricos tradicionais já estudados por Oliveira (2015); Galo *et al.* (2023) e com o modelo semi-paramétrico baseado em processos Gaussianos já estudado em (Sartori, 2024).

3.1.1 Análise descritiva

O conjunto de dados, publicado por Oliveira (2015), foi gentilmente cedido pelos autores, sendo o registro de 5.060 nascidos vivos de mães parturientes em Maringá (PR) em 2017. Todas as variáveis registradas são categóricas e suas proporções, segundo cada categoria, são mostradas na Tabela 3.1.

Tabela 3.1: Proporções de acordo com a ocorrência de prematuridade em recém-nascidos, Maringá, Paraná, Brasil, 2017

Relação	Variáveis	Categorias	Prematuridade	
			0: não	1: sim
			%	%
Mãe	Idade (Anos)	0: < 35	88.9	11.2
		1: ≥ 35	85.8	14.2
	Raça/Cor da pele	0: branca	87.4	12.6
		1: não branca	90.2	9.8
	Parceiro	0: sim	87.8	12.2
		1: não	89.2	10.8
	Educação (Anos)	0: ≥ 12	87.2	12.8
		1: ≤ 12	88.8	11.2
	Paridade	0: múltipla	88.9	11.1
		1: primária	87.1	12.9
Gravidez e Parto	Tipo de gravidez	0: única	90.2	9.8
		1: múltipla	25.0	75.0
	Tipo de parto	0: vaginal	91.6	8.4
		1: cesariana	87.2	12.8
	Consultas pré-natal	0: ≥ 7	90.1	9.9
		1: ≤ 7	77.6	22.4
	Local do parto	0: hospital	88.2	11.8
		1: outros	100.0	0.0
Recém-nascido	Sexo	0: feminino	88.2	11.8
		1: masculino	88.2	11.8
	Má-formação congênita	0: não	88.3	11.7
		1: sim	75.5	24.5
Total			88.2	11.8

Paridade se refere ao número de partos que a mulher teve após 20 semanas de gestação, "múltipla" se não foi o primeiro parto e "primária" se foi a primeiro.

Sendo assim, temos a variável de interesse ("prematuridade" na Tabela 3.1) e foram escolhidas 8 covariáveis binárias, sendo as mesmas utilizadas no projeto de Galo (2020) e Galo *et al.* (2023) permitindo a comparação dos resultados obtidos, que assumem apenas os valores 0 ou 1: "Idade", "Parceiro", "Paridade", "Tipo de gravidez", "Tipo de parto", "Consultas pré-natal", "Raça/Cor da pele" e "Má-formação congênita".

Abaixo temos duas tabelas de contingência para as variáveis “Prematuridade”e “Idade”e “Prematuridade”e “Consultas pré-natal”, respectivamente.

Tabela 3.2: Tabela de contigência para as variáveis Prematuridade e Idade da mãe

Idade \ Prematuridade	Não	Sim	Total
<35 anos	3408	424	3832
≥ 35 anos	1054	174	1228
Total	4462	598	5060

Tabela 3.3: Tabela de contigência para as variáveis Prematuridade e Consultas pré-natal

Consultas pré-natal \ Prematuridade	Não	Sim	Total
≥ 7 consultas	3850	421	4271
≤ 7 consultas	612	177	789
Total	4462	598	5060

Na Tabela 3.2 temos as frequências da variável de interesse “Prematuridade”segundo as categorias da covariável binária “Idade (Anos)”, e com isso, calculamos a estatística χ^2 para testar a independência entre as variáveis. O valor obtido para a estatística foi 8.3061 com 1 grau de liberdade. Dessa forma, o *p-valor* resultante é de 0.0040 indicando que podemos rejeitar a hipótese nula de independência, ou seja, concluímos que há associação relevante entre “Idade”e “Prematuridade”.

Na Tabela 3.3 temos as frequências da variável de interesse “Prematuridade”segundo as categorias da covariável binária“Consultas pré-natal”, e com isso, calculamos a estatística χ^2 para testar a independência entre as variáveis. O valor obtido para a estatística foi 99.8689 com 1 grau de liberdade. Dessa forma, o *p-valor* resultante é muito próximo de zero, indicadndo que podemos rejeitar a hipótese nula de independência, ou seja, concluímos que há associação relevante entre “Consultas pré-natal”e “Prematuridade”.

Tabela 3.4: *P-valores* entre Prematuridade e covariáveis binárias

Covariáveis binárias	*P-valores*
Idade (Anos)	0.0040
Raça/Cor da pele	0.0062
Parceiro	0.1916
Paridade	0.0468
Tipo de gravidez	0.0000
Tipo de parto	0.0001
Consultas pré-natal	0.0000
Má-formação congênita	0.0070

A Tabela 3.4 traz os *P-valores* entre Prematuridade e covariáveis binárias. Portanto, há evidência empírica para afirmar que existe associação entre a variável de interesse "Prematuridade"e essas covariáveis binárias que são utilizadas no estudo, pois os *p-valores* são muito pequenos indicando que podemos rejeitar a hipótese nula de indepêndencia, com exceção da covariável binária "Parceiro"que não tem associação relevante com a variável de interesse "Prematuridade". Entretanto, esse estudo tem a intenção de seguir as mesmas covariáveis binárias utilizadas em Galo (2020) e Galo *et al.* (2023) e por isso manteremos essa covariável binária para comparação dos resultados com os diferentes métodos. A tabelas de contingência para as demais variáveis que serão utilizadas nesse estudo e que deram origem aos p-valores mostrados na Tabela 3.4 estão no apêndice B.

Dessa forma, podemos aplicar as metodologias estudadas nesse conjunto de dados de nascimentos de bebês prematuros para estimar os diferentes modelos (regressão logística, regressão logística Bayesiana, processos gaussianos e SVM) e comparar a perfomance preditiva (segundo seus valores para a AUC, escore de Brier e escore logarítmico) para esses modelos de classificação binária.

3.2 Resultados preliminares

Para a regressão logística temos os resultados das estimativas no conjunto de dados nas Tabelas 3.5 e 3.6, respectivamente.

Na Tabela 3.5 temos os resultados do estudo anterior para as estimativas da regressão logística Bayesiana via MCMC e disponíveis em Galo *et al.* (2023). Na Tabela 3.6 seguem as estimativas obtidas em (Sartori, 2024). Nela também estão as estimativas de Máxima Verossimilhança da regressão logística e da regressão logística Bayesiana por aproximação

pela distribuição normal para que as estimativas obtidas pelas abordagens frequentista e Bayesiana possam ser comparadas numericamente.

Pelas tabelas percebemos que os dois métodos de estimação (via MCMC e aproximação pela normal) levaram a resultados muito próximos, indicando que a normal multivariada aproxima adequadamente a posteriori pois esta parece ser simétrica em todas as dimensões (o que também é sugerido pelo fato das médias e medianas obtidas via MCMC serem próximas dos valores encontrados para a média a posteriori pela aproximação assintótica da posteriori).

Dessa forma, para a comparação dos modelos vamos utilizar apenas a regressão logística estimada por Máxima Verossimilhança, visto que as estimativas por Máxima Verossimilhança, aproximação pela normal e por MCMC obtiveram resultados muito semelhantes.

Além disso, no estudo da Iniciação Científica (Sartori, 2024) também foi feita a estimação dos processos gaussianos e o cálculo das métricas para 10 amostras de treino diferentes para comparação da regressão logística e dos processos gaussianos.

Para o caso deste projeto nos processos gaussianos temos $|T|= 256$, portanto temos 256 combinações possíveis contendo 8 entradas em cada uma delas assumindo 0 ou 1. Entretanto, apenas 108 dessas combinações aparecem nesse conjunto de dados, sendo assim, a dimensão é reduzida, passa a ser 108.

Tabela 3.5: Estimativas da regressão logística Bayesiana (MCMC)

Coef. da variável	Logit Bayesiano (MCMC)	
	Médias Desvio-padrão	Mediana (95% HPD)
Constante	-2.770 (0.139)	-2.768 $(-3.038; -2.498)$
Idade	0.380 (0.111)	0.380 $(0.162; 0.593)$
Parceiro	-3.303 (0.116)	-0.301 $(-0.532; 0.076)$
Paridade	0.268 (0.099)	0.268 $(0.075; 0.460)$
Tipo de gravidez	3.343 (0.200)	3.338 $(2.951; 3.741)$
Tipo de parto	0.266 (0.124)	0.266 $(0.030; 0.513)$
Consultas pré-natal	1.174 (0.111)	1.174 $(0.962; 1.395)$
Raça/Cor da pele	-0.269 (0.110)	-0.267 $(-0.493; -0.060)$
Má-formação congênita	0.856 (0.370)	0.865 $(0.153; 1.591)$

Tabela 3.6: Estimativas da regressão logística (Máx. Verossimilhança e Bayesiana (Aprox. Normal))

Coef. da variável	Logit (EMV)	Logit Bayesiano (Aprox.)	
		Médias (95% HPD)	Desvio-padrão
Constante	−2.701	−2.762 (−3.013; −2.510)	0.128
Idade	0.362	0.379 (0.152; 0.606)	0.116
Parceiro	−0.296	−0.296 (−0.514; −0.079)	0.111
Paridade	0.250	0.267 (0.071; 0.462)	0.099
Tipo de gravidez	3.200	3.404 (2.862; 3.946)	0.277
Tipo de parto	0.235	0.262 (0.034; 0.491)	0.116
Consultas pré-natal	1.138	1.159 (0.897; 1.420)	0.133
Raça/Cor da pele	−0.270	−0.264 (−0.471; −0.056)	0.106
Má-formação congênita	0.780	0.858 (−0.114; 1.829)	0.496

3.3 Resultados finais

Os 5060 registros foram divididos aleatoriamente em amostras de treinamento (70%) e de teste (30% dos registros) e, considerando as covariáveis mencionadas acima ("Idade", "Parceiro", "Paridade", "Tipo de gravidez", "Tipo de parto", "Consultas pré-natal", "Raça/Cor da pele" e "Má-formação congênita") foram estimados os modelos logístico por máxima verossimilhança, processos gaussianos e SVM, visto que os modelos logístico por máxima verossimilhança, aproximação da Normal e MCMC obtiveram resultados muito semelhantes, por isso a escolha de apenas de um modelo logística para comparação com os demais.

A comparação desses modelos foi feita segundo a capacidade preditiva sobre as amostras de testes, que foram geradas a partir de 10 divisões aleatórias do conjunto de dados completo entre treinamento e teste. Posteriormente foi calculada a média aritmética das métricas obtidas para cada uma dessas divisões com o intuito de reduzir o viés na separação dos dados.

Para o SVM foi feita a validação cruzada *k-fold* com 5 *folds* para a escolha dos melhores hiperparâmetros, sendo assim os valores obtidos foram $Q = 1$ e $\gamma = 0.5$ para serem utilizados para estimar o modelo. Em cada amostra de teste foram utilizadas as métricas AUC (área sob a curva ROC), escore de Brier e escore logarítmico. A seguir estão as tabelas com as métricas calculadas para 10 amostras diferentes. Como ilustração, a Figura 3.1 mostra a comparação da curva ROC segundo os três métodos para uma das divisões geradas.

Tabela 3.7: AUC dos modelos para 10 amostras de testes diferentes

	Modelos		
Amostras	Regressão Logística	Processos Gaussianos	SVM
1	0.6699	0.6658	0.6054
2	0.6857	0.6878	0.6026
3	0.7142	0.6952	0.6086
4	0.7083	0.6854	0.6298
5	0.7052	0.7004	0.5938
6	0.6939	0.6856	0.5634
7	0.6916	0.6813	0.5155
8	0.6905	0.6951	0.5962
9	0.6687	0.6650	0.6020
10	0.7009	0.6878	0.5900
Média	0.69289	0.68494	0.59073
Desvio-Padrão	0.01520	0.01174	0.03122

Tabela 3.8: Escore logarítmico dos modelos para 10 amostras de testes diferentes

	Modelos		
Amostras	Regressão Logística	Processos Gaussianos	SVM
1	0.3097	0.3115	0.3181
2	0.3014	0.3021	0.3131
3	0.3117	0.3169	0.3246
4	0.3097	0.3166	0.3268
5	0.2942	0.2971	0.3043
6	0.3221	0.3261	0.3381
7	0.3134	0.3164	0.3243
8	0.3185	0.3183	0.3247
9	0.3552	0.3597	0.3642
10	0.3047	0.3090	0.3160
Média	0.31406	0.31737	0.32541
Desvio-Padrão	0.01654	0.01705	0.01636

Tabela 3.9: Escore de Brier dos modelos para 10 amostras de testes diferentes

	Modelos		
Amostras	Regressão Logística	Processos Gaussianos	SVM
1	0.0859	0.0864	0.0879
2	0.0832	0.0835	0.0861
3	0.0870	0.0882	0.0901
4	0.0868	0.0886	0.0910
5	0.0807	0.0812	0.0827
6	0.0916	0.0927	0.0955
7	0.0879	0.0889	0.0904
8	0.0889	0.0894	0.0903
9	0.1016	0.1024	0.1047
10	0.0848	0.0860	0.0871
Média	0.08784	0.08873	0.09058
Desvio-Padrão	0.00569	0.00577	0.00601

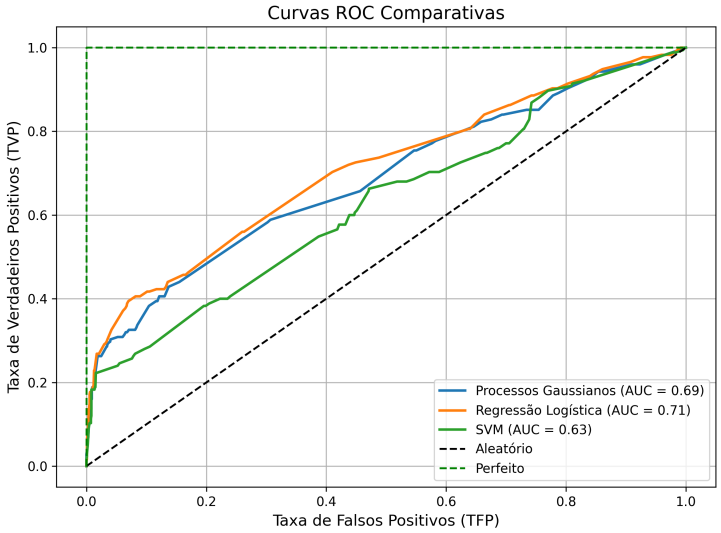

Figura 3.1: Comparação da curva ROC dos três modelos.

Destacamos novamente que, para a AUC, quanto mais perto o valor está de 1, melhor é o desempenho do modelo. Já para o escore logarítmico e para o escore de Brier quanto mais perto de 0, melhor é o desempenho do modelo.

De acordo com as Tabelas 3.7, 3.8 e 3.9 é possível perceber que os modelos de regressão logística e processos gaussianos apresentam performances preditivas muito semelhantes para esse conjunto de dados, independente da métrica de comparação utilizada (AUC, escore de Brier e escore logarítmico). Entretanto, o SVM obteve uma perfomance preditiva inferior em todas as métricas utilizadas, sendo uma diferença mais nítida na AUC, indicando menor capacidade de discriminar as classes.

Sabendo que a AUC varia entre 0 a 1, a diferença entre a regressão logística e os

processos gaussianos não apresentam diferenças significativas numericamente, mas para valores inferiores a 0.5 indica que o modelo está invertendo as classes, sendo assim é considerado o intervalo entre 0.5 e 1 para interpretação da AUC. Dessa forma, as diferenças pequenas nos valores da AUC podem ser relevantes, visto que pequenas variações podem refletir melhorias significativas na capacidade de discriminar corretamente as classes.

Como os processos gaussianos podem ser considerados um modelo semi-paramétrico, espera-se um bom desempenho preditivo, pois conseguem ser capazes de captar relações não lineares entre as covariáveis e a variável de interesse. Para o SVM é esperado também um melhor desempenho preditivo, pois é um método de aprendizado de máquina utilizado principalmente para estudos de classificação, sendo um modelo não paramétrico, visto que usando o SVM com o kernel gaussiano a fronteira de decisão depende apenas de todos os pontos de treinamento, permitindo modelar fronteiras de decisão complexas.

No entanto, para esse conjunto de dados não houve grandes diferenças em relação ao desempenho dos modelos de regressão logística e os processos gaussianos. Dessa forma, neste estudo ambos os métodos tiveram boas performances preditivas, com ligeira vantagem para a regressão logística, que obteve melhores valores médios em todas as métricas avaliadas.

Além disso, baseando-se no exemplo da Figura 3.1, percebe-se que as curvas ROC para a regressão logística e para os processos gaussianos são mais próximas entre si e para o SVM está consideravelmente abaixo, mostrando um desempenho menor para esse modelo. Essa observação confirma os resultados obtidos nas tabelas das métricas, mostrando que a regressão logística apresentou melhor desempenho nesse conjunto de dados de nascimentos de bebês prematuros. Além disso, nesse caso a regressão logística é preferível, pois os resultados foram muito semelhantes e possui a vantagem de gerar um modelo facilmente interpretável.

Sendo assim, usando a Tabela 3.5 é possível fazer a exponencial das estimativas dos parâmetros da regressão logística e obter uma interpretação dos resultados. Para a estimação por Máxima Verossimilhança, a variável "Idade"possui um coeficiente estimado de 0.362, indicando que gestantes mais velhas apresentam maior chance de ocorrência do parto prematuro. Em termos de razão de chances, esse coeficiente corresponde a um *odds ratio* de $e^{0.362} \approx 1.436$, ou seja, as mães com mais de 35 anos de idade possuem *odds ratio* aproximadamente 44% maior, mantendo-se constantes as demais covariáveis, em relação às mães com menos de 35 anos de idade.

Da mesma forma, o coeficiente estimado para a variável "Consultas pré-natal" foi positivo de 1.138, indicando que a realização de menos de sete consultas pré-natal está associada a um aumento da *odds ratio* de parto prematuro. Em termos quantitativos, esse coeficiente corresponde a um *odds ratio* de $e^{1.138} \approx 3.120$. Dessa forma, na regressão logística é possível obter uma interpretação dos resultados dos coeficientes estimados para todas as covariáveis.

É interessante comparar os resultados obtidos com um modelo completamente preditivo, ou seja, em que excluímos as variáveis que não são conhecidas antes do nascimento, sendo as variáveis "Tipo de parto" e "Má-formação congênita". Além disso, diferentemente de Galo *et al.* (2023), incluímos o nível de escolaridade da mãe (se completou 12 ou mais anos de estudo) como covariável "Educação (Anos)"nesses modelos. É importante destacar que a variável "Local do parto"não foi utilizada nesses modelos, pois não há variação da variável resposta quando a covariável assume o rótulo 1 ("outros"), o que indica que o nascimento não ocorreu em ambiente hospitalar.

Tabela 3.10: AUC dos modelos preditivos para 10 amostras de testes diferentes

	Modelos		
Amostras	Regressão Logística	Processos Gaussianos	SVM
1	0.6447	0.6289	0.5757
2	0.6880	0.6833	0.5740
3	0.6831	0.6775	0.6067
4	0.6573	0.6510	0.5912
5	0.6955	0.6740	0.6251
6	0.7047	0.6947	0.5878
7	0.6955	0.6805	0.5813
8	0.7116	0.6903	0.5737
9	0.7028	0.6899	0.6206
10	0.6533	0.6631	0.5982
Média	0.6837	0.6733	0.5934
Desvio-Padrão	0.0236	0.0205	0.0189

Tabela 3.11: Escore logarítmico dos modelos preditivos para 10 amostras de testes diferentes

	Modelos		
Amostras	Regressão Logística	Processos Gaussianos	SVM
1	0.3261	0.3295	0.3326
2	0.3420	0.3430	0.3535
3	0.3229	0.3269	0.3320
4	0.3364	0.3406	0.3383
5	0.3121	0.3215	0.3208
6	0.3040	0.3084	0.3132
7	0.3199	0.3232	0.3286
8	0.3222	0.3267	0.3360
9	0.3226	0.3247	0.3317
10	0.3298	0.3299	0.3382
Média	0.3238	0.3274	0.3325
Desvio-Padrão	0.0109	0.0097	0.0108

Tabela 3.12: Escore de Brier dos modelos preditivos para 10 amostras de testes diferentes

	Modelos		
Amostras	Regressão Logística	Processos Gaussianos	SVM
1	0.0917	0.0922	0.0935
2	0.0981	0.0982	0.1008
3	0.0902	0.0917	0.0929
4	0.0946	0.0965	0.0953
5	0.0866	0.0899	0.0890
6	0.0845	0.0859	0.0861
7	0.0897	0.0902	0.0916
8	0.0914	0.0922	0.0945
9	0.0909	0.0914	0.0928
10	0.0929	0.0932	0.0952
Média	0.0911	0.0921	0.0932
Desvio-Padrão	0.0038	0.0034	0.0039

A Tabela 3.10 mostra que, com exceção do SVM com kernel, os outros dois modelos apresentaram desempenho preditivo semelhante em termos de AUC. No entanto, ao considerar os escores logarítmico e de Brier, as Tabelas 3.11 e 3.12 indicam que o desempenho preditivo dos três modelos foi próxima. Destaca-se que a regressão logística apresentou, em média, desempenho superior aos demais modelos, embora a diferença observada tenha sido pequena.

Portanto, é possível observar que os resultados obtidos foram próximos aos dos modelos estimados inicialmente de acordo com os projetos de Galo (2020) e Galo *et al.* (2023) que utilizam variáveis não conhecidas antes do nascimento. Sendo assim, não houve redução

na capacidade preditiva do modelo e apresentou desempenho semelhante ao dos modelos estimados nas Tabelas 3.7, 3.8 e 3.9.

Capítulo 4

Conclusão

Nesse Trabalho de Graduação foram estudadas as metodologias do método Support Vector Machine (SVM) e das métricas de comparação de perfomance preditiva (AUC, escore de Brier e escore logarítmico). Além disso, para comparação, foram utilizados os resultados da Iniciação Científica (Sartori, 2024), em que foram estimados dois modelos de classificação binária, sendo um deles paramétrico tradicional (regressão logística) e o outro semi-paramétrico baseado em processos gaussianos. Considerando covariáveis que refletem atributos da mãe, do feto e do parto foi possível comparar a perfomance preditiva desses três modelos.

As Tabelas 3.7, 3.8 e 3.9 mostram o resultado das métricas de poder preditivo de 10 amostras diferentes do conjunto de teste geradas aleatoriamente: AUC (área sob a curva ROC), escore logarítmico e escore de Brier, respectivamente. Para cada métrica foram estimados os três modelos (regressão logística, processos gaussianos e SVM) em 10 amostras de testes diferentes geradas a partir do conjunto de dados de nascimentos de bebês prematuros. Para comparação do desempenho foram estimados esses três modelos sem as variáveis não conhecidas antes do nascimento e os resultados das métricas estão nas Tabelas 3.10, 3.11 e 3.12, em que foi observado desempenhos próximos nos modelos estimados inicalmente com as variáveis de acordo com Galo (2020) e Galo *et al.* (2023) e nos modelos preditivos, por isso, o estudo continua sendo baseado nos modelos com as variáveis propostas inicialmente.

O SVM apresentou um desempenho preditivo inferior aos outros modelos, visto que nas três métricas obteve valores razoavelmente menores e pela curva ROC é possível perceber que a curva está abaixo dos demais modelos, apresentando uma AUC média de aproximadamente 0.59, enquanto a regressão logística apresentou uma AUC média de

0.693 e os processos gaussianos, 0.685. Dessa forma, temos que a regressão logística e os processos gaussianos tiveram perfomances preditivas muito semelhantes, embora marginalmente a regressão logística obteve resultados melhores neste conjunto de dados de nascimentos de bebês prematuros, sendo preferível por gerar um modelo mais facilmente interpretável.

Entretanto, em outros conjuntos de dados podemos obter resultados diferentes, por exemplo, em hipóteses de linearidade distante, ou seja, quando os dados apresentam relações não-lineares, a regressão logística tende a não ser muito eficiente, portanto, os processos gaussianos e o SVM provavelmente teriam desempenhos preditivos melhores.

A regressão logística tem como vantagens a facilidade de cálculos, permitindo identificar o risco dos estimadores e dos coeficientes, além de possuir baixa complexidade computacional, em termos de tempo de execução e uso de memória, mesmo para grandes conjuntos de dados.

Por outro lado, os processos gaussianos tem como vantagens ser um modelo semiparamétrico, ou seja, é mais flexível para se ajustar aos dados e possui melhores resultados para relações não-lineares. Além disso, permite a escolha de um kernel que seja melhor para cada conjunto de dados, e por ser bayesiano é possível obter com facilidade a distribuição a posteriori associada a cada valor da função latente, ou seja, segundo as características específicas do indivíduo em estudo (no caso desse estudo, o parto a ser realizado).

O SVM busca encontrar um hiperplano que maximiza a margem de separação entre as classes, ou seja, a distância entre as observações mais próximas de cada classe (vetores de suporte). Sendo assim, tem como vantagens a capacidade de lidar com relações não lineares, utilizando funções kernel, que mapeiam os dados para uma dimensão superior, tornando os dados linearmente separáveis nesse novo espaço.

A escolha da função kernel e dos hiperparâmetros Q de penalização e γ do kernel RBF influenciam o desempenho preditivo do modelo. Para esse trabalho, foram escolhidos os hiperparâmetros por validação cruzada para obter um melhor ajuste aos dados. Entretanto, o SVM apresentou um desempenho preditivo inferior aos demais modelos, pricipalmente em alguns conjuntos de testes, em que os hiperparâmetros escolhidos podem não ser os melhores para discriminar corretamente as classes, ou seja, podem levar a perda de generalização, resultando em desempenho inferior.

Sendo assim, como possíveis desenvolvimentos é interessante fazer um estudo da per-

fomance do SVM para esse conjunto de dados utilizando outros kernels para verificar se é possível obter uma melhora no desempenho desse modelo. Além disso, é possível utilizar técnicas para interpretar modelos complexos como o SVM, mostrando o motivo do modelo tomar determinadas decisões. Por exemplo, o Explainable AI é um conjunto de técnicas, como o Local Interpretable Model-agnostic Explanations (LIME), o Partial Dependence Plot (PDP) e o Individual Conditional Expectation (ICE) que podem ser utilizados para uma maior compreensão do SVM.

Portanto, considerando o conjunto de dados de nascimentos de bebês prematuros e a comparação entre os modelos de regressão logística, processos gaussianos e Support Vector Machine (SVM) por meio das métricas de AUC, escore logarítmico e escore de Brier, conclui-se que a regressão logística apresentou o melhor desempenho preditivo, sendo de forma ligeiramente superior aos demais. Além disso, destaca-se que, por ser um modelo simples, eficiente e de fácil interpretação, pode ser o mais útil em estudos da área da saúde.

O conjunto de dados utilizado e os códigos realizados referente aos modelos estudados nesse Trabalho de Conclusão de Curso estão no link abaixo: https://github.com/LeticiaBS7/Trabalho-de-Conclusao-de-Curso.git.

Referências Bibliográficas

Bernando, J. M. e Smith, A. F. M. (1994). *Bayesian Theory*. Wiley.

Boser, B. E., Guyon, I. M. e Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Em *Proceedings of the fifth annual workshop on Computational learning theory – COLT '92*, páginas 144–152.

Friedman, J., Hastie, T. e Tibshrani, R. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Galo, R. (2020). Modelagem bayesiana para dados de nascimentos prematuros desbalanceados. Dissertação de Mestrado, Universidade Estadual de Maringá.

Galo, R., Rossi, R. M., Alves, D. C. e Oliveira, R. R. (2023). Bayesian binary regression using power and power reverse link functions: an application to premature birth data. *Brazilian Journal of Biometrics*, **2**(41), 131–143.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, São Carlos, SP. ISBN 978-65-00-02410-4. Livro eletrônico.

James, G., Witten, D., Hastie, T., Tibshirani, R. e Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.

Oliveira, R. R. (2015). Nascimento prematuro no estado do paraná e município de maringá. Tese de Doutorado, Universidade Estadual de Maringá.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Em *Advances in Large Margin Classifiers*, páginas 61–74.

Rasmussen, C., E. e Williams, C., K. I. (2006). *Gaussian Process for Machine Learning*. MIT Press.

Sartori, L. B. (2024). Modelos semi-paramétricos de regressão binária: uma aplicação a dados de nascimentos prematuros. Relatório Final de Iniciação Científica – Universidade Federal de São Carlos, sob orientação de Márcio Alves Diniz. Projeto financiado pela FAPESP (Processo 2023/15010-3).

Apêndice A

Derivações de Momentos

Utilizando infomações do livro Rasmussen e Williams (2006) podemos resolver analiticamente a integral (2.1), que fornece a probabilidade de um indivíduo da amostra de treinamento pertencer à categoria de interesse quando possui o vetor de covariáveis $\mathbf{x}^*$.

Considere a integral de uma normal acumulada, Φ, em relação a uma normal com média μ e variância σ^2:

$$Z = \int_{-\infty}^{\infty} \Phi\left(\frac{x-m}{v}\right) N(x \mid \mu, \sigma^2) dx$$

em que,

$$\Phi(x) = \int_{-\infty}^{x} N(y \mid 0, 1) dy$$

Primeiramente, vamos fazer para o caso $v > 0$, substituir por $z = y - x + \mu - m$ e $w = x - \mu$ e trocar as ordens das integrais:

$$Z_{v>0} = \frac{1}{2\pi\sigma v} \int_{-\infty}^{\infty} \int_{-\infty}^{x} \exp\left(-\frac{(y-m)^2}{2v^2} - \frac{(x-\mu)^2}{2\sigma^2}\right) dy dx$$

$$= \frac{1}{2\pi\sigma v} \int_{-\infty}^{\mu-m} \int_{-\infty}^{\infty} \exp\left(-\frac{(z+w)^2}{2v^2} - \frac{w^2}{2\sigma^2}\right) dw\, dz$$

Ou seja, uma integral (incompleta) sobre uma normal conjunta. A integral interna corresponde a marginalização sobre w:

$$Z_{v>0} = \frac{1}{\sqrt{2\pi(v^2+\sigma^2)}} \int_{-\infty}^{\mu-m} \exp\left(-\frac{z^2}{2(v^2+\sigma^2)}\right) dz = \Phi\left(\frac{\mu-m}{\sqrt{v^2+\sigma^2}}\right)$$

Agora, para $v < 0$ podemos substituir a simetria $\Phi(-z) = 1 - \Phi(z)$ na equação inicial:

$$Z_{v<0} = 1 - \Phi\left(\frac{\mu - m}{\sqrt{v^2 + \sigma^2}}\right) = \Phi\left(-\frac{\mu - m}{\sqrt{v^2 + \sigma^2}}\right)$$

Apêndice B

Tabelas de contingência

Tabelas de contingência para as variáveis da tabela 3.1 que foram selecionadas para o estudo de acordo com Galo (2020) e Galo *et al.* (2023):

Raça/Cor da pele \ Prematuridade	Não	Sim	Total
Branca	3142	454	3596
Não branca	1320	144	1464
Total	4462	598	5060

Parceiro \ Prematuridade	Não	Sim	Total
Sim	3310	459	3769
Não	1152	139	1291
Total	4462	598	5060

Paridade \ Prematuridade	Não	Sim	Total
Múltipla	2583	320	2903
Primária	1879	278	2157
Total	4462	598	5060

Tipo de gravidez \ Prematuridade	Não	Sim	Total
Única	4423	481	4904
Múltipla	39	117	156
Total	4462	598	5060

Tipo de parto \ Prematuridade	Não	Sim	Total
Vaginal	1051	97	1148
Cesariana	3411	501	3912
Total	4462	598	5060

Má-formação congênita \ Prematuridade	Não	Sim	Total
Não	4427	586	5013
Sim	35	12	47
Total	4462	598	5060