Modelo probabilístico de previsão de infarto cardíaco via processos Gaussianos

Luan Coelho Vieira da Silva
Programa de Engenharia de Sistemas e Computação - PESC/ COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
luan@cos.ufrj.br

***Resumo*—O presente estudo investiga a aplicação de processos Gaussianos a um conjunto de dados composto por informações de 799 pacientes anônimos. Cada paciente possui informação de colesterol total, glicemia, idade e um desfecho binário (indicando se o paciente infartou ou não dentro de um período de 120 dias da realização dos exames). Nosso objetivo principal é desenvolver um modelo preditivo para a ocorrência de infarto cardíaco, utilizando uma abordagem probabilística através de processos Gaussianos.**

Para seleção do modelo foi realizado um processo de validação cruzada no qual foram avaliados diferentes funções de covariância, a fim de selecionar um modelo com alta capacidade de prever se um novo paciente infarta ou não dentro do período de acompanhamento, conhecendo suas informações clínicas e idade.

Os resultados experimentais nos levaram a selecionar uma função de covariância Matérn para ser utilizada na priori do processo Gaussiano. Utilizando um limiar de probabilidade de 50% para tomada de decisão, a precisão encontrada para ocorrência de infarto foi de 77%, e a taxa de revocação ficou em 47%. O modelo permite flexibilidade para uma decisão informada da equipe médica.

***Palavras-chave*—processos Gaussianos, Infarto, Classificação binária**

I. Introdução

De acordo com a Organização Mundial de Saúde (OMS), doenças cardiovasculares são a principal causa de morte em todo o mundo. Hábitos não saudáveis podem levam a características indicadoras de aumentado risco de ataque cardíaco, como aumento da pressão arterial, aumento da glicemia, aumento dos lipídios no sangue, e obesidade [1].

Diversos estudos buscaram, no contexto de aprendizado de máquina, encontrar maneiras eficientes de prever o risco de infarto com base em dados clínicos dos pacientes. Takci [2] verifica a capacidade de diferentes modelos e o efeito da seleção de variáveis na qualidade da predição. Waqar et al [3] faz uso de técnicas de aprendizado profundo alinhado ao método de sobreamostragem *SMOTE* para prever ocorrências de infarto em dados desbalanceados.

O presente artigo propõe o uso de processos Gaussianos [4] através de variáveis explicativas contidas em um conjunto de dados com 799 pacientes anônimos. Foi utilizada a biblioteca *scikit-learn* [5], que utiliza a técnica de Laplace [6] para aproximar a distribuição a posteriori dos dados.

A. Conjunto de dados

O conjunto de dados utilizado neste trabalho foi obtido através de informações clínicas - colesterol total e glicemia -, idade e desfecho (ocorrência ou não de infarto em até 120 dias após a consulta) de 799 pacientes anônimos.

A Figura 1 mostra a distribuição das variáveis explicativas usadas na predição de ocorrência ou não de infarto. O colesterol total da amostra varia entre 150 e 279 mg/dL, com metade dos pacientes apresentando valores iguais ou inferiores a $219mg/dL$, e também metade possuindo valores iguais ou superiores a esse limite. Já a idade está compreendida entre 10 e 86 anos, sendo 46 anos a mediana. Por fim, a glicemia varia entre 64 e 169 mg/dL, com medida central de 109 mg/dL.

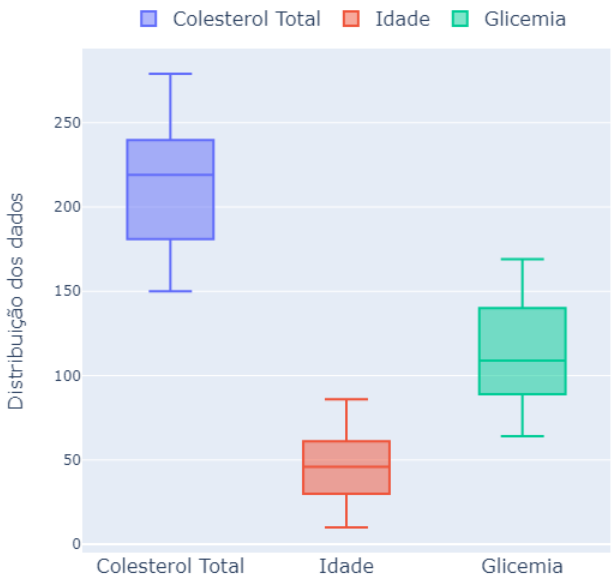

Figura 1. Boxplot da distribuição das variáveis explicativas

A Tabela I mostra como se comporta a variável binária de interesse neste estudo: o desfecho. Cerca de 67% dos pacientes não tiveram infarto em até 120 dias após a realização dos exames. Há, portanto, um desbalanceamento entre as classes a serem previstas. Vale notar também que um modelo trivial que atribua para todos os pacientes a não ocorrência de infarto acertaria quase 67% das ocasiões. Desta forma, espera-se que um bom modelo apresente *performance* superior.

A Figura 2 apresenta a matriz de correlação entre as variáveis. As variáveis explicativas possuem baixo nível de correlação linear entre si. Em relação ao desfecho, todas

as covariáveis aprensentam valores de correlação positivos, indicando uma relação direta entre elas e a ocorrência de infarto.

Tabela I
DISTRIBUIÇÃO DA VARIÁVEL A SER PREVISTA - DESFECHO

Desfecho	Contagem	Percentual
Não ocorreu infarto	535	66,96%
Ocorreu infarto	264	33.04%

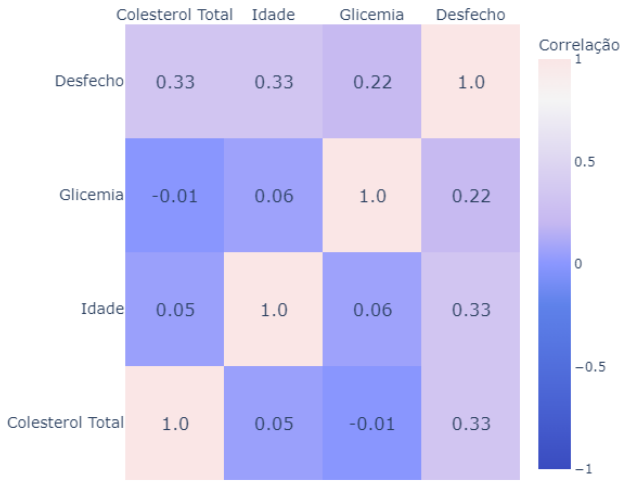

Figura 2. Matriz de correlação das variáveis dos pacientes

II. METODOLOGIA

Este trabalho busca prever a ocorrência ou não de infarto em até 120 dias após a coleta dos exames. Esta tarefa de classificação binária foi tratada utilizando processos Gaussianos, um método Bayesiano. Métricas de avaliação: precisão, revocação, *F1-score* e acurácia, foram utilizadas para avaliar o desempenho do modelo proposto. Os resultados são discutidos utilizando o caráter probabilístico da análise, permitindo maior flexibilidade na escolha do limiar de probabilidade para tomada de decisão.

A. Estatística Bayesiana

O paradigma Bayesiano de estatística [7] é frequentemente abordado sob a ótica frequentista, partindo da verossimilhança e adicionalmente assumindo como verdadeira uma distribuição a priori para as variáveis a serem estimadas. No entanto, uma maneira mais esclarecedora de compreendê-lo seria conceber um modelo generativo para a distribuição conjunta de todas as variáveis envolvidas. Isso engloba tanto as variáveis observadas, que representam os dados, quanto as não observadas, que correspondem aos parâmetros.

A estatística bayesiana postula uma distribuição conjunta para os dados e os parâmetros, permitindo a inferência sobre todas essas variáveis em vez de apenas estimativas pontuais. É importante notar que, da mesma forma que assumir uma distribuição para os parâmetros é um processo delicado que pode resultar em um modelo inadequado, o mesmo se aplica a distribuição para os dados. Uma análise cuidadosa é necessária em ambos os casos.

B. Distribuição Gaussiana

A função de densidade de probabilidade da distribuição Gaussiana multivariada de dimensão n, $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, é dada por:

$$\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$\mathbf{x}$ é o vetor de variáveis aleatórias

$\boldsymbol{\mu}$ é o vetor $n \times 1$ de média das variáveis aleatórias

$\boldsymbol{\Sigma}$ é a matriz $n \times n$ de covariância das variáveis aleatórias

$|\boldsymbol{\Sigma}|$ representa o determinante da matriz de covariância.

C. processos Gaussianos

Um processo Gaussiano é uma coleção de variáveis aleatórias com função conjunta dada por uma distribuição Gaussiana. Trata-se de um método não paramétrico, ou seja, não faz suposições explicitas sobre a forma da função-alvo f. Ao não restringir o conjunto de hipóteses a um grupo específico de funções-alvo, há maior flexibilidade para ajustar diferentes dados. No entanto, como não há a restrição de f, uma quantidade maior de dados é geralmente necessária para garantir a convergência e generalização [8]. Este método utiliza estatística bayesiana, pois assume uma priori de distribuição Gaussiana sob o espaço de funções f. A inferência é baseada na distribuição de probabilidade a posteriori, obtida através Teorema de Bayes.

processos Gaussianos são baseados em funções de covariância (*kernels*). São univocamente definidos por uma função de média $m(\mathbf{x})$ e uma função de covariância de variáveis aleatórias dada por $k(\mathbf{x_i}, \mathbf{x_j})$. Williams & Rasmussen [9] explicam detalhadamente a aplicação dessas funções para o funcionamento da classificação. Neste trabalho foram utilizados três tipos de funções de covariância. Dados $\mathbf{x_i}$ e $\mathbf{x_j}$, eles podem ser descritos da seguinte forma:

- Função de base radial: também chamado de exponencial quadrático, é parametrizado pelo parâmetro de comprimento de escala l. Sua fórmula é dada por $\exp\left(-\frac{d(\mathbf{x_i},\mathbf{x_j})^2}{2l^2}\right)$, $d(.)$ representa a distância euclidiana.
- Matérn: possui um parâmetro adicional ν que controla a suavidade da função resultante. Sua fórmula é dada por $\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(\mathbf{x_i},\mathbf{x_j})\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}d(\mathbf{x_i},\mathbf{x_j})\right)$ $K_\nu(\cdot)$ representa uma função Bessel modificada e $\Gamma(\cdot)$ é a função gama.
- Quadrática Racional: pode ser visto como uma soma infinita de núcleos de função de base radial. Sua fórmula

é dada por $\left(1 + \frac{d(\mathbf{x_i}, \mathbf{x_j})^2}{2\alpha l^2}\right)^{-\alpha}$, α é o parâmetro de mistura de escala.

Ao realizar análise de regressão com o uso de processos Gaussianos, é prática comum assumir verossimilhança Gaussiana para os dados. Com uma priori de processo Gaussiano, tal combinação resulta em uma distribuição a posteriori de processo Gaussiano, e seu formato é expresso de forma analítica. Trata-se de um caso de priori e verrosimilhança conjugadas. No entanto, quando se trata de classificação binária, não faz sentido presumir que os dados, que podem assumir apenas dois valores, sigam uma distribuição Gaussiana.

Para classificação, são utilizadas funções de ligação como logit e probit, a fim de obter um resultado entre 0 e 1 a indicar a probabilidade a posteriori de cada classe. Para este estudo em particular, faremos uso da função logit, que é a padrão no pacote *sklearn*. Para aproximar o resultado da distribuição a posteriori para os dados, que não possui forma analítica fechada, é utilizada aproximação de Laplace.

D. Métricas de avaliação

No contexto da classificação binária, para uma dada classe prevista - neste caso sendo a ocorrência de infarto, definimos os seguintes termos:

- Verdadeiros Positivos (VP): Instâncias em que o paciente infarta dentro do período de 120 dias, e o modelo prevê corretamente.
- Falsos Positivos (FP): Instâncias em que o paciente não infarta dentro do período de 120 dias, e o modelo prevê incorretamente.
- Falsos Negativos (FN): Instâncias em que o paciente infarta dentro do período de 120 dias, e o modelo prevê incorretamente.
- Verdadeiros Negativos (VN): Instâncias em que o paciente não infarta dentro do período de 120 dias, e o modelo prevê corretamente.

Analogamente, as mesmas medidas são definidas para o caso em que o paciente não infarta dentro do período de 120 dias. Esses termos nos permitem calcular métricas de avaliação, que são comumente usadas para avaliar o desempenho de modelos em tarefas de classificação binária.

A revocação mede a capacidade do modelo identificar corretamente as ocorrências de infarto dentre todas as ocasiões em que o paciente de fato infartou. Ela é calculada como a razão entre os verdadeiros positivos e a soma dos verdadeiros positivos e falsos negativos.

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

A precisão mede a capacidade do modelo de identificar corretamente a ocorrência de infarto. Ela é calculada como a razão entre os verdadeiros positivos e a soma dos verdadeiros positivos e falsos positivos.

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Tabela II
RESULTADOS DE MÉDIA E DESVIO PADRÃO DA ACURÁCIA DOS *splits* DURANTE O PROCEDIMENTO DE VALIDAÇÃO

	Splits 1 a 5	
Função de covariância inicial	Média	Desvio-padrão
Função de base radial($l = 1$)	0,754	0,036
Matérn($l = 1, \nu = 1.5$)	0,757	0,034
Quadrática Racional($\alpha = 1, l = 1$)	0,754	0,036

O *F1-score* é a média harmônica da precisão e da revocação. Fornece uma medida equilibrada de ambas as métricas.

$$\textit{F1-score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Além disso, a acurácia é outra métrica comumente usada em tarefas de classificação, que mede a proporção de todas as previsões corretas feitas pelo modelo em relação ao número total de instâncias.

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

III. EXPERIMENTOS

Os dados foram divididos em treinamento e teste, na proporção aproximada de 80% e 20%, respectivamente, sendo previamente realizado um embaralhamento para evitar possível viés na ordenação original dos dados. Dentro dos dados de treinamento, foi utilizado o procedimento de validação cruzada para selecionar o modelo e estimar o erro fora da amostra, isto é, para novos pacientes cujos dados não tenham sido utilizados durante o treinamento. Ainda, as covariáveis foram padronizadas, sendo subtraídas pelas médias e em seguida divididas pelo desvio-padrão de seus respectivos dados de treinamento.

Para a validação cruzada, os dados de treino foram divididos em cinco conjuntos (*splits*) - sendo um deles separado para validação. O processo foi repetido cinco vezes, utilizando todos os *splits* para validação uma vez. Os resultados de validação em cada uma das divisões, bem como a média e o desvio-padrão entre os *splits* é descrito na Tabela II. Os resultados obtidos usando as diferentes funções de covariância são próximos, estando a menos de um desvio-padrão de distância entre si. Por esta razão, não é possível afirmar que uma escolha de função de covariância é melhor.

Escolhemos seguir a análise com a função de covariância Matérn, que apresentou a maior média de acurácia por *split*. Primeiramente, foram ajustados todos os dados de treinamento usando este núcleo. A função de covariância otimizada após o aprendizado passou a ter a forma $2,25^2 \times \text{Matérn}(l = 4,47, \nu = 1,5)$. Em seguida, foram realizadas previsões utilizando os dados de teste. Além das observações dos dados de teste não terem sido utilizadas em momento algum durante o treinamento, quando analisamos os resultados com este conjunto de dados, o modelo já está definido - isto é, há apenas

uma hipótese para os dados. Com isso, a desigualdade de Hoeffding é válida [10], possibilitando uma melhor generalização para o erro fora da amostra.

IV. Resultados e Discussão

Obter altos resultados nas métricas de previsão de infarto não é um fim por si só. O interesse por este tópico está relacionado a possibilidade de auxiliar a equipe médica a orientar os pacientes da melhor forma - e consequentemente salvar vidas.

Por padrão, o modelo Gaussiano prevê a classe com a maior probabilidade de ocorrência a posteriori. No caso de classificação binária, é atribuída ao paciente a classe cuja probabilidade excede 50%. Como visto na Tabela I, os dados deste estudo apresentam desbalanceamento entre as classes, e é esperado que a previsão de não ocorrência de infarto seja preferida mediante incerteza. O caráter probabilístico dos modelos Bayesianos permite maior flexibilidade para que a equipe médica possa escolher o limiar de probabilidade adequado para a predição.

A. Resultados

A Tabela III mostra os resultados de precisão, revocação, *F1-score* e acurácia utilizando 3 limiares de probabilidade diferentes de forma que é prevista a ocorrência de infarto caso a probabilidade a posteriori de infarto encontrada no nosso modelo escolhido seja superior a cada um destes limiares.

Percebe-se, em relação a ocorrência de infarto, o aumento da precisão e diminuição da revocação conforme o limiar de probabilidade aumenta. A escolha de um limiar de probabilidade inferior faz sentido, portanto, caso a equipe médica entenda que um diagnóstico positivo erroneamente não é tão grave. Por exemplo, caso seja previsto que o paciente infartará dentro de 120 dias, e a única consequência imediata seja uma orientação para hábitos mais saudáveis, esta escolha pode ser preferida. Por outro lado, caso o responsável acredite que uma predição de infarto incorreta deva ser fortemente evitada, um limiar de probabilidade maior seria capaz de alcançar este objetivo.

Tratando especificamente do caso do limiar de probabilidade de 50%, padrão deste tipo de classificação, observamos que a revocação é maior para o caso de não ocorrência de infarto, 92% contra 75% obtida para o caso de ocorrência de infarto. Isso é esperado, devido ao desbalanceamento das classes nos dados. No entanto, para o mesmo limiar de probabilidade, a precisão foi maior dentro da classe de ocorrência de infarto, indicando que quando ocorre a previsão de infarto há uma grande probabilidade dela ser concretizada.

B. Curvas de probabilidade a posteriori

Nosso modelo fornece a distribuição a posteriori para os dados, definindo portanto, de forma unívoca, uma probabilidade a posteriori de desfecho - ocorrência ou não de infarto - para cada conjunto de observações - idade, glicemia e colesterol total - dos pacientes. Portanto, é possível pensarmos a probabilidade a posteriori de infarto como função da idade, glicemia e colesterol total - denotada por g. Como são três variáveis explicativas, a visualização dos dados em três dimensões seria dificultada.

A Figura 3 mostra as curvas de probabilidades preditivas a posteriori obtida através das interseções da função g com os hiperplanos dados por Glicemia = 87 mg/dL **(a)**, Idade = 40 anos **(b)** e Colesterol Total = 241 mg/dL **(c)**. Os pontos vermelhos indicam pacientes do conjunto de teste que de fato sofreram um infarto nos 120 dias subsequentes à coleta dos exames, enquanto os pontos azuis representam pacientes que não tiveram a interrupção do fluxo sanguíneo para o coração neste período.

O item **(a)** da Figuta 3 indica a situação dos pacientes dos dados de treinamento cuja glicemia tenha sido medida em 87 mg/dL. Apenas um deles teve infarto em até 120 dias após a medição, e o modelo previu que ele teria a maior probabilidade de infarto dentre todos os pares na mesma condição, com cerca de 63,6% de probabilidade. Já o item **(b)** indica um situação na qual todos os pacientes de idade 40 anos tiveram uma probabilidade infarto inferior a 50% prevista pelo modelo, no entanto, dois deles infartaram dentro do período de acompanhamento, e, embora baixas, suas probabilidades de ataque cardíaco foram as maiores dentro deste grupo. Por fim, o item **(c)** retrata a situação dos pacientes com colesterol total de 241 mg/dL na base de teste - valor acima dos níveis de referência. Quatro dos nove pacientes neste cenário infartaram, e o modelo atribuiu probabilidade superior a 50% para o infarto de três deles. Um indivíduo cuja probabilidade predita de ataque cardíaco foi superior a meio não infartou dentro do período considerado.

V. Conclusão

A previsão de infarto cardíaco desempenha um papel fundamental no campo da medicina e na assistência médica moderna. Doenças cardiovasculares representam a principal causa de morte em todo o mundo, a capacidade de antecipar o risco de infarto com base em dados clínicos dos pacientes é uma ferramenta valiosa para profissionais de saúde. Essa previsão não apenas permite a identificação precoce de pacientes em risco, mas também auxilia os médicos na tomada de decisões informadas sobre o tratamento e na alocação de recursos de forma mais eficiente.

Neste estudo foi empregado o uso de processos Gaussianos para previsão de ocorrência de infarto em até 120 dias após a realização de exames clínicos. Os dados foram divididos em treinamento e teste, e foi realizada validação cruzada para seleção de modelo e estimativa de erro fora da amostra. Após a validação foi selecionado um modelo com função convariância inicial $1^2 \times \text{Matérn}(l = 1, \nu = 1,5)$. Este modelo foi então treinado utilizando todos os dados de treinamento, resultando em uma função de covariância otimizada $2,25^2 \times \text{Matérn}(l = 4,47, \nu = 1,5)$.

Foram apresentados resultados obtidos através da predição sobre o conjunto de teste, utilizando três limiares de probabilidade, e diferentes curvas preditivas de probabilidade a posteriori foram analisadas. A abordagem probabilística

Tabela III
RESULTADOS OBTIDOS NO CONJUNTO DE TESTE

	Probabilidade = 0,3			Probabilidade = 0,5			Probabilidade = 0,7			
	Precisão	Revocação	*F1- score*	Precisão	Revocação	*F1- score*	Precisão	Revocação	*F1- score*	Suporte
Não ocorreu infarto	0,85	0,76	0,80	0,75	0,92	0,83	0,70	0,98	0,82	102
Ocorreu infarto	0,65	0,76	0,70	0,77	0,47	0,58	0,89	0,28	0,42	58
acurácia			0,76			0,76			0,73	160

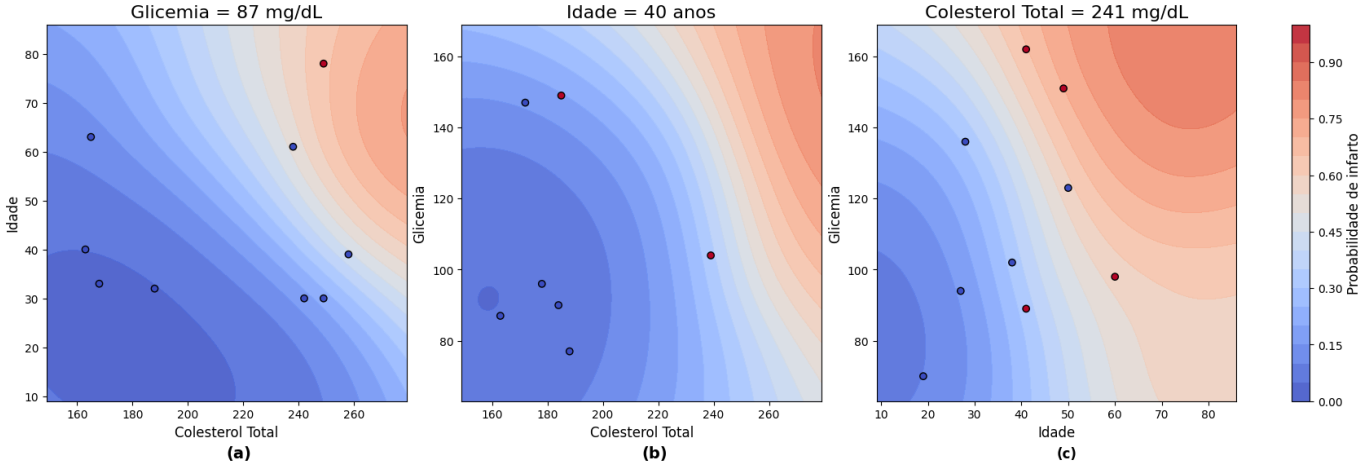

Figura 3. Curvas de probabilidade preditiva a posteriori

empregada neste artigo permite flexibilidade para a tomada de decisão médica.

A previsão de infarto é uma tarefa bastante difícil de ser realizada, pois há diversos fatores não observados que impactam no desfecho de um paciente. Por exemplo, não se sabe se o paciente adotou ou não hábitos mais saudáveis após os exames. A fim de obter previsões mais assertivas, um trabalho futuro pode considerar a combinação de diferentes funções de covariância para obter melhores estimativas de probabilidades a posteriori, buscar obter dados de mais pacientes, tendo em vista o caráter não paramétrico do modelo adotado, que demanda muita informação, porém também é capaz de gerar de aproximar funções-alvo complexas. O acompanhamento por um período superior a 120 dias, alinhado à coleta de dados referentes aos hábitos paciente pós exame também pode gerar resultados mais satisfatórios. Essas considerações têm o potencial de fornecer resultados ainda mais valiosos para a prática clínica.

REFERÊNCIAS

[1] World Health Organization, "Cardiovascular diseases (cvds)," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), accessado em: 15 de setembro de 2023.

[2] H. Takci, "Improvement of heart attack prediction by the feature selection methods," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 1, pp. 1–10, 2018.

[3] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An efficient smote-based deep learning model for heart attack prediction," *Scientific Programming*, vol. 2021, pp. 1–12, 2021.

[4] K. P. Murphy, "Chapter 18: Gaussian processes," in *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023, pp. 673–725. [Online]. Available: http://probml.github.io/book2

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] Z. Shun and P. McCullagh, "Laplace approximation of high dimensional integrals," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 4, pp. 749–760, 1995.

[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[8] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Statistical learning," in *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023, pp. 15–67.

[9] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[10] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*. AMLBook New York, 2012, vol. 4.