Avaliando a importância da completude dos dados clínicos em um caso real de diagnóstico de mamas

Márcio Juchneski[1], Eudoxia Lottie Silva Moura[1,2], Roger Resmini[3], Aura Conci[1]

[1]Instituto de Computação - Universidade Federal Fluminense

[2]Instituto Federal de Rondônia

[3]Instituto de Ciências Exatas e Naturais - Universidade Federal de Rondonópolis

marciojuchneski@id.uff.br, eudoxia.moura@ifro.edu.br, roger@ufr.edu.br, aconci@ic.uff.br

Abstract. *This work focuses on verifying the importance of completeness of breast metadata for diagnosis aid systems. The aim is to investigate the possibility of using already known approaches to treat missing data and how they affect diagnoses. For a real case example, the Dataset for Mastologic Research (DMR), developed and made available at IC/UFF is used. This includes, in addition to clinical data, thermal breast images, mammograms, and diagnoses. After a bibliographical review of possible techniques, the Hot Deck was considered the most appropriate to be compared with the simple exclusion of attributes of missing data to classify known cases. Its use in classification of normal or abnormal patients resulted in 94% of correction in the Area Under the Receiver Operating Characteristic Curve (AUC) versus 92% if the six attributes with the largest amount of missing data were simply disregarded.*

Resumo. *Este trabalho concentra-se na verificação da importância da completude dos metadados em sistemas de auxílio ao diagnóstico de mamas. Objetiva-se investigar a possibilidade de utilizar abordagens já conhecidas para tratar dados faltantes e verificar sua influência na classificação dos exames. Para se ter um exemplo em um caso real é utilizado o Dataset for Mastologic Research (DMR), desenvolvido e disponibilizado no IC/UFF. Esse conjunto além dos dados clínicos imagens térmicas da mama, mamografias e diagnósticos. Após uma revisão bibliográfica de possíveis técnicas a Hot Deck se mostrou a mais adequada para uma comparação com a simples exclusão dos atributos com muitos dados faltantes na classificação dos casos com diagnóstico comprovado. No entanto, seu uso na classificação entre pacientes normais ou com algum problema nas mamas, com todo o dataset fez com que o resultado final tivesse 94% de acerto na Área Sob a Curva Receiver Operating Characteristic (AUC) versus 92% nos casos em que os seis atributos com maior quantidade de dados faltantes fossem simplesmente desconsiderados.*

1. Introdução

Em exames médicos, sempre haverá algum dado clínico presente, mesmo que este seja apenas o laudo final ou o diagnóstico relativo ao exame. A importância da compreensão da completude adequada de um banco de dados médico é o aspecto explorado neste artigo. Diferentemente de outros tipos de dados, a curadoria dos dados médicos deve ser considerada com cuidado extremo, pois pode ser melhor deixá-los inalterados ou mesmo incompletos para uma utilização futura mais adequada das valiosas informações, que cada um dos seus elementos pode conter, para os sistemas de apoio à decisão clínica (ou Clinical Decision Support Systems, CDSS). Os metadados clínicos podem fornecer insights valiosos para os profissionais de saúde, permitindo uma análise mais precisa e correlações mais detalhadas do paciente e até da doença. A análise de dados (data mining) tem um papel essencial no tratamento desses dados clínicos ao aplicar técnicas que possibilitam extrair conhecimento relevante deles. Isso pode resultar em avanços significativos no diagnóstico, prognóstico e tratamento do paciente, além de auxiliar na identificação de padrões e tendências que podem passar despercebidos por métodos convencionais no entendimento das doenças.

O câncer de mama é o tipo de câncer mais comum entre as mulheres. De acordo com o Instituto Nacional de Câncer, 73.610 novos casos de câncer de mama foram diagnosticados no país em 2023 [1]. Até 2040, prevê-se que esse diagnóstico aumente para cerca de 3 milhões [2]. Ultrassonografia, mamografia, ressonância magnética e biópsia são os exames mais usados no sistema público de saúde. Embora a mamografia seja a mais usada para detecção inicial de microcalcificações, ela possui pontos negativos em relação ao desconforto, exposição à radiação e limitações na identificação de achados em pacientes mais jovens. Neste sentido, a termografia é uma alternativa sem riscos à saúde, indolor e capaz de identificar características fisiológicas precoces, especialmente em mulheres jovens, gestantes, puérperas e lactantes. Como nos demais exames de imagens, na termografia, há capturas das mamas, mas também dados clínicos dos pacientes. Como na maioria dos exames realizados no Brasil em locais que atendem gratuitamente nossa população, devido a diversos motivos, muitas vezes esses dados clínicos são negligenciados quanto ao seu preenchimento na ficha do paciente. Essa falta acaba muitas vezes apenas sendo notada quando posteriormente estatísticas ou análises com os mesmos são realizadas. Outro problema é que como há campos com preenchimento livre, muitas vezes é um desafio completá-los de forma adequada para uma visualização uniforme de determinado aspecto dos exames.

Diversas técnicas têm sido usadas nesta etapa de curadoria dos dados, e a relevância desta completude nas análises deste tipo específico de dados é o objetivo deste trabalho. Neste sentido, será comparada a metodologia Hot Deck para inclusão dos dados faltantes com resultados obtidos pela não utilização do dado ou não preenchimento deles em sistemas de auxílio ao diagnóstico da mama. Primeiramente, são relacionados os principais trabalhos que sugerem técnicas de preenchimento automático de dados médicos faltantes para diagnósticos por imagem (Seção 2 - Trabalhos Relacionados). Em seguida, apresenta-se a metodologia de Hot Deck, o banco de dados usado como exemplo e quais análises serão feitas sob esses dados (Seção 3 - Materiais e Métodos). Os resultados obtidos são apresentados em tabelas para facilitar sua comparação (na Seção 4 - Resultados), bem como as medidas utilizadas em suas avaliações quantitativas. Finalmente, é feita uma discussão dos resultados obtidos (Seção 5: Conclusões e Prosseguimentos).

2. Trabalhos Relacionados

Uma parcela significativa dos esforços na análise de dados está ligada à compreensão e organização dos dados desorganizados (brutos) para explorá-los eficazmente. No processo de organização de dados, surgem desafios ligados à necessidade de codificá-los adequadamente e à identificação de problemas para possibilitar sua exploração. A etapa de exploração de dados desempenha um papel fundamental nesse processo, abrangendo esforços dedicados a descobrir características inesperadas nos dados, que podem gerar insights valiosos e agregar valor à análise dos mesmos. Através da exploração minuciosa, busca-se identificar padrões, tendências e relações ocultas, que podem fornecer informações essenciais para tomadas de decisão na área dos mesmos. Uma compreensão profunda dos dados e uma habilidade apurada para explorá-los adequadamente são essenciais para garantir uma análise de qualidade e constituem uma etapa crucial na capacidade de identificar e aproveitar o potencial dos dados [3].

Para construir os escores do APACHE II, um sistema de pontuação para pacientes em unidades de terapia intensiva (UTI), que fornece uma estimativa padronizada e imparcial da probabilidade de morte hospitalar, foi utilizada uma base de dados de 23 pacientes de UTIs, onde foram utilizadas várias técnicas de imputação de valores aos dados faltantes e entre elas a técnica Hot Deck [4]. O número de dados faltantes era bastante grande em relação ao número de dados analisados: foram relatadas um total de 86 variáveis, sendo apenas duas sem valores faltantes, ou seja, apenas cerca de 2,3% das variáveis estavam com todas as informações completas, tendo um total de 1.540 dados ausentes observados no conjunto de dados da pesquisa. Os autores concluem que a técnica Hot Deck foi eficiente no preenchimento, mas levou a alguns vieses no estudo.

Em [5] apresentaram comparações de vários procedimentos para imputação de valores no estudo do colesterol total e análise do risco de mortalidade coronariana para um aumento no colesterol. Os critérios utilizados na comparação foram a média e o desvio padrão. Quando há de 10 a 60% de dados faltantes, houve diferenças claras entre os métodos, sendo que a imputação múltipla com Hot Deck foi o melhor método. No entanto, com falta de dados acima de 60%, nenhum método de imputação pareceu ser satisfatório. Os autores concluem que a técnica Hot Deck foi a mais confiável que outras mais simples.

Em [6] compararam métodos para lidar com dados faltantes no sistema de gerenciamento de depressão em pessoas idosas. Utilizaram a imputação múltipla com Hot Deck, a correspondência preditiva da média para dados faltantes e a abordagem bayesiana aproximada para dados não respondidos.

O último método é baseado em um modelo normal multivariado (MVN) usando o PROC MI com o software SAS V8.2. Mais de quinhentas variáveis foram coletadas, sendo que a maioria das variáveis apresentou taxas de itens ausentes menor do que 2%. Os autores afirmam que a técnica Hot Deck teve um bom desempenho para alcançar o objetivo proposto pela pesquisa. Baseados nestes trabalhos, considera-se que a técnica Hot Deck deveria ser utilizada para o tratamento dos dados faltantes nesta pesquisa, visto que estudos anteriores

fornecem evidências de ser uma abordagem promissora para o tratamento de dados faltantes no contexto de dados médicos clínicos.

3. Materiais e Métodos

A utilização de bases de dados na área médica é essencial para o avanço das pesquisas e a obtenção de resultados confiáveis. No entanto, é importante ressaltar que a disponibilidade dessas bases de dados pode ser limitada. Em especial no caso de exames de mamas, que ainda é uma área de pesquisa incipiente, especialmente no caso de ter anotações e classificação diagnóstica confirmada pela comunidade interessada [7].

De acordo com [8], existem 16 bases de dados relacionadas à termografia, sendo que, entre essas, apenas o DMR (Dataset for Mastologic Research) é aberto para acesso público. O trabalho desenvolvido por [9] foi fundamental para a divulgação do DMR como uma base de dados pública de exames da mama que disponibilizou a pesquisadores da área acesso a um conjunto de exames com oportunidades para investigações em diferentes contextos. Antes do DMR, as imagens de termografia eram, em geral, de acesso restrito às pesquisas internas dos hospitais [10] e [11]. Os estudos ocorriam em bases de dados privadas sem possibilidade de se comparar adequadamente os resultados obtidos [12]. A publicação em formato open access é de suma importância para a pesquisa científica [13].

A ausência de dados é uma complicação recorrente em estudos que envolvem análise de dados do mundo real. As razões para a falta de dados podem ser diversas, algumas são relacionadas aos contextos próprios dados necessários ao estudo e outras são de natureza aleatória. Variáveis podem apresentar lacunas de preenchimento devido à ausência de importância dada aos mesmos na coleta de determinados campos, ou os dados podem estar faltantes devido à recusa dos participantes em fornecer algumas informações, muitas vezes motivadas pela necessidade de preservar a confidencialidade, seu desconhecimento ou a pressa em ser atendido. Outras fontes de dados faltantes são intrínsecas do Sistema de Informação (SI) e podem incluir problemas na extração de um banco de dados, falhas na integração de informações, entre outros.

Neste trabalho, compara-se objetivamente a importância de todos os dados clínicos de exames de mama estarem completos ou não no diagnóstico de normalidade ou anormalidade dos pacientes. Após tomar conhecimento de detalhes da pesquisa, cada paciente participou voluntariamente da mesma, assinando o Termo de Consentimento Livre e Esclarecido (TCLE) do projeto, que foi aprovado pelo Comitê de Ética em Pesquisa (CEP) e registrado na plataforma Brasil do Ministério da Saúde sob número de Certificado de Apresentação para Apreciação Ética (CAAE) 01042812.0.0000.5243. Para o registro das imagens termográficas por infravemelho foi utilizada uma câmera da marca Flir modelo SC 620.

Há dois conjuntos de metadados nestas aquisições realizadas entre os anos de 2012 e 2021, com 367 exames de 306 pacientes: um relativo a dados dos pacientes e outro dos exames realizados em uma determinada data (ou visitas). O dataset patients contém 6 tipos de informações (Tabelas 1) e o dataset visits tem 32 tipos de informações (atributos), com nomes de atributos significativos e valores bastante variados, como sintomas, queixas, datas (último exame, da primeira e última

menstruação), temperatura corporal da pessoa na hora do exame, histórico familiar de câncer, hábitos alimentares, se a pessoa já fez biópsias, cirurgias, tratamentos radioterápicos, diagnóstico de cada mama e do paciente, etc. Para o processamento dos dados, foram utilizados a ferramenta de codificação Visual Studio Code 1.74.2, Python 3.6, com as bibliotecas Pandas 1.1.5, Plotly 4.14.13 e Dash 1.20.0.

Tabela 1. Nomes e valores dos atributos ligados aos pacientes: dataset *Patients.*

Nome do atributo	**Valor possível do atributo**
id	Número inteiro positivo
date_birth	dia/mês/ano
marital_status	widow, married, single, divorced ou viúvo, solteiro, casado, divorciado ou viúva, solteira, casada, divorciada
race	(asian, black, indigenous, mulatto, white) ou (amarela, negra, parda, branca e indígena)
type	(p ou v) ja patient ou voluntario extra
record_date	Data do registro como (dia/mês/ano hora:minuto)

A análise inicial desses dados mostrou ter apenas 4,96% do total dos registros completos. Esse resultado indica uma lacuna significativa no preenchimento dos dados. Como estratégia inicial, tentou-se verificar a influência de uniformizar os dados de cada classe, mas manter os dados faltantes no conjunto de dados e verificar o quanto isso modificaria os resultados de predições de diagnóstico, visto que um dado irreal facilmente introduz vieses durante a análise. A Tabela 2 apresenta os atributos com a maior quantidade de dados faltantes. Na coluna "Porcentagem faltante", são apresentados o número dos dados faltantes para cada atributo pelo total de registros no dataset. Pode-se perceber que o atributo "bc_family", que indica a ocorrência de câncer de mama na família, é o atributo que mais tem dados faltantes. Essa falta ocorre porque os examinados frequentemente não têm conhecimento sobre o histórico familiar, seguido de se o paciente está ou não na menopausa (isso porque se considera que o fornecimento da data de última menstruação tornava óbvio se a pessoa estava ou não na menopausa). Os demais atributos (histórico de doenças em geral da família, sintomas

que a pessoa esteja sentindo, sinais nas mamas e queixas em geral) geralmente não estavam preenchidos porque não ocorriam para a pessoa em exame. Foram incluídos nesta tabela atributos com pelo menos falta em 100 voluntários (ou seja, mais de 27% do total dos pacientes da base de dados).

Tabela 2. Atributos com mais dados faltantes no dataset *Visits.*

Nome do atributo	**Porcentagem faltante** %
bc_family	72
menopause	56
familys_history	49
symptoms	47
signs	35
complaints	33

Dentre as técnicas de imputação existentes, a técnica Hot Deck (ou Hot Deck Imputation) é uma abordagem amplamente utilizada para estimar valores ausentes por meio da identificação e substituição por observações existentes semelhantes. A técnica Hot Deck é um método de imputação de dados faltantes que busca estimar valores ausentes por meio da identificação de observações completas semelhantes (referidas como "doadores") dentro do conjunto de dados. Essa abordagem é baseada na premissa de que os valores ausentes podem ser substituídos por observações semelhantes em termos de características relevantes. A técnica Hot Deck é aplicada quando há uma razão válida para acreditar que observações similares apresentam valores ausentes semelhantes. Essa abordagem é particularmente útil em casos em que a estrutura dos dados sugere uma dependência entre as observações, sendo bastante utilizada na área médica e epidemiológica. Consiste basicamente em separar as amostras do conjunto de dados em que todos os indivíduos tenham respondido daqueles que não responderam. Com essa separação, é feita uma comparação de similaridade entre as características dos dados e então o dado faltante é preenchido aleatoriamente pelo dado do grupo que tiver maior similaridade [14].

Assim optou-se por usar o método Hot Deck para preencher os dados faltantes das diversas colunas do conjunto de dados *visits*, ou seja, foi usado nestas colunas os atributos que mais se aproximavam (se correlacionam com os demais), de modo a gerar uma substituição dos valores para um novo dataset sem dados faltantes (*i.e*. completo). Assim, optou-se por usar o método Hot Deck para preencher os dados faltantes das diversas colunas do conjunto de dados visits. Ou seja, foram utilizados nestas colunas os atributos que mais se aproximavam (ou seja, que se correlacionavam com os demais), de modo a gerar uma substituição dos valores para um novo dataset sem dados faltantes (ou seja, completo). Ao fazer isso, houve uma redução de 29,72% no número de colunas completas (das 37 colunas de dados, apenas 26 tinham dados completos). Esse resultado indica uma lacuna significativa no preenchimento dos dados com a utilização da técnica

Hot Deck, visto que, das 322 tuplas do conjunto de dados, apenas 16 ficaram com dados totalmente preenchidos.

Além disso, nem todos os atributos puderam ser considerados em um preenchimento automático pelo método. Os atributos “id_patient”, “menarche2”, “id” e “temperature” foram desconsiderados por serem atributos de categoria numérica. As colunas “date” e a coluna “lmp” (last menstrual period) foram removidas por serem do tipo data. A coluna “menarche”, apesar de ser um atributo categórico, apresentou 42 categorias de registros possíveis, o que ocasionou incompatibilidade para a classificação e, portanto, foi removida. Os demais atributos foram usados na forma completa e como estão (incompletos) em um software de mineração de dados para verificar a correção do diagnóstico em ambos os casos, utilizando uma medida de desempenho bastante eficiente: a AUC (área abaixo da curva ROC).

4. Resultados

Um algoritmo de classificação utiliza alguma técnica de reconhecimento de padrões a partir de dados para indicar a classe de uma amostra aleatória. Alguns dos algoritmos de classificação mais utilizados são:

1. Árvores de Decisão: uma categoria de algoritmos que utiliza o relacionamento entre os dados para construir uma árvore, na qual os nós mais próximos da raiz têm maior relevância na classificação.

2. Random Forest: uma estratégia de combinação de classificadores na qual várias árvores de decisão são aplicadas para encontrar melhores resultados médios de classificação do que uma única árvore.

3. Support Vector Machine (SVM): uma técnica estatística que separa os dados com auxílio de vetores de suporte em um espaço n-dimensional, o espaço de atributos.

4. Naïve Bayes: um algoritmo probabilístico que utiliza a lógica de Bayes como estratégia para calcular as probabilidades.

Para calcular o desempenho da classificação, podem ser utilizadas várias medidas. Uma delas é a área abaixo da curva (Area Under the Curve, AUC), que trata da Curva Característica de Operação do Receptor (Receiver Operating Characteristic, ROC). A AUC é apresentada em valores entre 0 e 1, sendo 1 o valor ótimo. Ela representa a relação entre a quantidade de acertos e a quantidade de erros e é indicada como medida de desempenho para amostras desbalanceadas, onde a quantidade de instâncias em cada classe não é a mesma ou é próxima [15].

Os resultados de classificação considerando a AUC para os quatro métodos com sete variações dos dados são mostrados na Tabela 3. Na segunda coluna desta tabela, é apresentado o resultado obtido com o dataset inteiro. Observa-se que, neste caso, o classificador Random Forest apresentou o melhor resultado com 93%. Para as análises feitas nas demais colunas, foi retirado antes da classificação algum dos dados mais faltantes do banco e a análise foi refeita.

Assim, a coluna 3 apresenta o resultado obtido sem o atributo “bc_family”, que é o que tem mais valores faltantes. Novamente, o classificador Random Forest apresentou o melhor resultado com 94% de AUC. A coluna 4 da Tabela 3 apresenta o resultado obtido sem os atributos “bc_family” e “menopause”, os dois com mais dados faltantes. Também o classificador Random Forest apresentou o melhor resultado com 93% de AUC. A coluna 5 apresenta o resultado obtido sem os atributos “bc_family”, “menopause” e “family_history”. Embora os resultados dos diversos classificadores variem, o classificador Random Forest apresentou resultados com 94% de AUC. A coluna 6 apresenta o resultado obtido sem os atributos “bc_family”, “menopause”, “family_history” e “symptoms”, em que o classificador Random Forest apresentou o melhor resultado com 94% de AUC. A coluna 7 da Tabela 3 apresenta o resultado obtido sem os atributos “bc_family”, “menopause”, “family_history”, “symptoms” e “signs”, mas isso não alterou os resultados do classificador Random Forest (94% de AUC). A coluna 8 apresenta o resultado obtido sem os atributos “bc_family”, “menopause”, “family_history”, “symptoms”, “signs” e “complaints”. O classificador Random Forest apresentou o melhor resultado com 92% de AUC.

Tabela 3. Resultados de AUC com os diversos classificadores e subconjuntos do dataset “visits” sem alguns atributos.

Tipo	Todo o dataset	Sem o mais faltante	Sem os 2 mais faltantes.	Sem os 3 mais faltantes.	Sem os 4 mais faltantes.	Sem os 5 mais faltantes	Sem os 6 mais faltantes
Árvore de Decisão	0,83	0,83	0,82	0,84	0,83	0,85	0,84
SVM	0,92	0,92	0,92	0,93	0,93	0,93	0,91
Random Forest	0,93	0,94	0,93	0,94	0,94	0,94	0,92
Naïve Bayes	0,82	0,83	0,83	0,84	0,84	0,84	0,83

Como pode ser observado na Tabela 3, o classificador Random Forest mostrou-se como o melhor algoritmo de classificação. Além disso, as diversas colunas da tabela demonstraram que a remoção de um ou todos os atributos com muitos dados faltantes

não resulta em perda significativa em relação aos resultados obtidos com os atributos que têm muitos dados faltantes. Desta forma, sugere-se que os dados faltantes não sejam mantidos para evitar sobrecarregar o processamento com atributos que não contribuirão significativamente para o resultado final. Além disso, o resultado obtido sem os atributos removidos sugere que eles não têm influência na tarefa de classificação, conforme demonstrado neste trabalho.

5. Conclusões e Prosseguimentos

A disponibilidade de conjuntos de dados completos e de alta qualidade desempenha um papel fundamental na obtenção de conclusões precisas e confiáveis em pesquisas médicas. No entanto, é comum haver atributos com dados faltantes em estudos clínicos ou registros de saúde. A presença desses dados ausentes pode apresentar desafios significativos na análise e interpretação dos resultados, podendo distorcer conclusões médicas e impactar negativamente a tomada de decisões clínicas.

O objetivo deste artigo é mostrar a importância de lidar adequadamente com os dados faltantes para validar uma classificação de um estudo em medicina nos moldes dos demais métodos usados em ciência de dados. É conhecido que a ausência de dados nos metadados pode deixar de fornecer insights valiosos sobre o paciente, mas também se compreende a necessidade de explorar possíveis abordagens para lidar com esses dados ausentes de maneira adequada para o conjunto de dados e o fenômeno em estudo. Dessa forma, um tratamento adequado dos dados faltantes nos metadados médicos pode resultar em análises médicas mais robustas e confiáveis.

Para exemplificar esse cenário, foi utilizado o conjunto de dados fornecido pelo DMR (Dataset for Mastologic Research), um banco de dados para mastologia. Através dessa investigação, foi possível ilustrar de forma prática como a abordagem correta em relação aos dados faltantes nos metadados dos exames pode contribuir para a obtenção de resultados mais sólidos e conclusões médicas mais precisas.

A técnica Hot Deck foi a abordagem usada, substituindo os valores ausentes pelos mesmos valores dos dados conhecidos. Isso pode levar a uma sobreconfiança nos dados imputados, uma vez que não há uma consideração explícita da incerteza na imputação. Essa falta de incerteza pode levar a uma subestimação das estimativas de erro padrão e, consequentemente, a uma confiança excessiva nos resultados da análise.

Nos experimentos de classificação o algoritmo Random Forest apresentou os melhores resultados, sendo que na classificação com todos os atributos o resultado foi de 94% de AUC e sem os seis atributos com mais registros faltantes foi de 92% de AUC. Pode-se concluir então que os atributos com muitos dados faltantes não contribuem significativamente na classificação de diagnósticos. Por se tratar de dados de saúde, também pode-se concluir que a imputação de dados é uma tarefa extremamente complexa e a técnica de Hot Deck não foi suficientemente robusta para contribuir com a melhoria do dataset.

Agradecimentos

M.J.L.J. recebeu bolsa CAPES durante seu mestrado no IC/UFF. E.L.S.M. tem apoio do Instituto Federal de Educação, Ciência e Tecnologia de Rondônia (IFRO). R.R. …….. A.C. é apoiado em parte pelos Institutos Nacionais de Ciência e Tecnologia (projeto INCT - MACC), Conselho Nacional de Ciência e Tecnologia (CNPq) sob bolsa 307638/2022-79, Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) sobre Projetos CNE, e-Health Rio e Digit3D ("temáticos").

Referências

[1] Instituto Nacional de Câncer - INCA (2023) “Estatísticas de câncer”, Disponível em https://www.gov.br/inca/pt-br/assuntos/cancer/numeros.

[2] Arnold M., Morgan E., Rumgay H., Mafra A., Singh D., Laversanne M., Vignat J., Gralow J.R., Cardoso F., Siesling S., Soerjomataram I., (2022). "Current and future burden of breast cancer: Global statistics for 2020 and 2040", The Breast. https://doi.org/10.1016/j.breast.2022.08.010.

[3] Donoho, D. "50 Years of Data Science." Journal of Computational and Graphical Statistics, vol. 26, no. 4, 2017, pp. 745-766. doi:10.1080/10618600.2017.1384734.

[4] Pérez, A., Dennis, R.J., Gil, J.F.A., Rondón, M.A., and López, A. "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia." Statist. Med., vol. 21, 2002, pp. 3885-389 doi:10.1002/sim.1391.

[5] Barzi, F., Woodward, M. "Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies." American Journal of Epidemiology, vol. 160, no. 1, 2004, pp. 34-45. doi:10.1093/aje/kwh175.

[6] Tang, L., Song, J., Belin, T.R., and Unützer, J. "A comparison of imputation methods in a longitudinal randomized clinical trial." Statist. Med., vol. 24, 2005, pp. 2111-2128. doi:10.1002/sim.2099.

[7] Periyasamy, S., Prakasarao, A., Menaka, M., Venkatraman, B., Jayashree, M. (2021). Thermal Grading Scale for Classification of Breast Thermograms. IEEE Sensors Journal, 21(13), 13996–14002. https://doi.org/10.1109/JSEN.2020.3045455.

[8] Pérez-Martín, J. e Sánchez-Cauce, R., “Quality analysis of a breast thermal images database”, Health Informatics Journal, 2023, https://doi.org/10.1177/14604582231153779 https://journals.sagepub.com/doi/10.1177/14604582231153779.

[9] Silva, L. F., et al. "A new database for breast research with infrared images." Journal of Medical Imaging and Health Informatics 4.1 (2014): 92-100.

[10] Lee, C. H., Dershaw, D. D., Kopans, D., Evans, P., Monsees, B., Monticciolo, D., Burhenne, L. W. (2010). Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. Journal of the American College of Radiology, 7(1), 18- 27.

[11] Neal, C. H., Flynt, K. A., Jeffries, D. O., Helvie, M. A. (2018). Breast Imaging Outcomes following Abnormal Thermography. Academic Radiology, 25(3), 273–278. https://doi.org/10.1016/j.acra.2017.10.015.

[12] Heena, H., Durrani, S., Riaz, M., Al Fayyad, I., Tabasim, R., Parvez, G., & Abu-Shaheen, A. (2019). Knowledge, attitudes, and practices related to breast cancer screening among female health care professionals: a cross sectional study. BMC women's health, 19, 1-112019.

[13] Ragavendra, U., Gudigar, A., Rao, T. N., Ciaccio, E. J., Ng, E. Y. K., & Rajendra Acharya, U. (2019). Computer-aided diagnosis for the identification of breast cancer using thermogram images: A comprehensive review. Infrared Physics & Technology, 102, 103041. https://doi.org/10.1016/j.infrared.2019.103041.

[14] Andridge, R. R, Little R. J. "A Review of Hot Deck Imputation for Survey Non-response." Int Stat Rev, vol. 78, no. 1, 2010, pp. 40-64. doi:10.1111/j.1751-5823.2010.00103.x.

[15] Resmini, R. *et al*. Combining genetic algorithms and SVM for breast cancer diagnosis using infrared thermography. Sensors, v. 21, n. 14, p. 4802, 2021.