ANÁLISE ESTATÍSTICA PARA A SELEÇÃO DE VARIÁVEIS EM DADOS DE CÂNCER DE MAMA PARA A CONSTRUÇÃO DE MODELOS DE REDES NEURAIS ARTIFICIAIS

LUCAS DE OLIVEIRA HERNANDES[1], MARCO AURÉLIO GRANERO[2]

[1] Graduando em Licenciatura em Matemática, IFSP, Campus São Paulo, lucas.hernandes@aluno.ifsp.edu.br.
2 Doutor, UNICAMP, Professor do IFSP, Campus São Paulo, São Paulo, SP, Brasil. E-mail:granero@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 1.02.02.07-2 Planejamento de Experimentos

RESUMO: O presente estudo tem como objetivo analisar as variáveis do conjunto de dados *Breast Cancer Wisconsin*, identificando a relação de dependência e/ou independência entre elas, possibilitando estabelecer um critério objetivo na escolha de um subconjunto de variáveis que venha a representar um problema futuro de forma simplificada, no caso de interesse dos autores, a construção de modelos de Redes Neurais Artificiais. Para isto, uma análise estatística dos dados foi realizada, buscando mensurar a variabilidade das variáveis e suas escalas, correlações lineares, independência e a tendência à normalidade de cada característica.

PALAVRAS-CHAVE: Testes de normalidade, Correlações lineares, Análise de variância, Análise de variabilidade, Seleção de variáveis.

SELECTION OF BREAST CANCER VARIABLES FOR THE CONSTRUCTION OF AN ARTIFICIAL NEURAL NETWORK MODEL THROUGH STATISTICAL ANALYSIS.

ABSTRACT: This study looked at selecting the most important characteristics of the Breast Cancer Wisconsin dataset in order to create an Artificial Neural Network models with less input variables, which would reduce the computational cost of using the algorithms. To this end, a statistical analysis of the data was carried out, seeking to measure the variability of the variables and their scales, linear correlations, independence and the tendency towards normality of each characteristic. This statistical analysis also made it possible to verify that the standardization of the data set for creating the model is necessary.

KEYWORDS: *Normality tests, Linear correlations, Analysis of variance, Variability analysis.*

INTRODUÇÃO

O câncer de mama é o tipo de neoplasia com maior incidência em mulheres no Brasil, com exceção dos tumores de pele não melanoma, sendo a principal causa de morte de mulheres no país (INCA, 2023). Chala e Urban (2023) apresentam que a identificação do câncer de mama em estágio inicial reduz a necessidade de intervenções cirúrgicas complexas. Deste modo, a identificação e classificação de tumores tornam-se relevantes no contexto científico e social brasileiro.

O Aprendizado de Máquina pode ser entendido como uma coleção de técnicas e algoritmos que visa fazer com que um computador, de forma automatizada, extraia informações, identifique padrões e encontre relações em grandes quantidades de dados, treinando e aprendendo com este conjunto de dados (Abu Mustafa, Magdon-Isma e Lin, 2012). Neste sentido a qualidade dos dados, isto é, a existência de *outliers* e/ou dados não representativos podem acarretar problemas no processo de extração de informações (Hernandes, Santos e Granero, 2024). Portanto, a verificação da disposição estatística do conjunto de dados e das relações existentes suas variáveis são de extrema importância para a construção de modelos que utilizem técnicas de Aprendizado de Máquina.

Desta forma, este trabalho apresenta um estudo acerca do comportamento e da variabilidade estatística do conjunto de dados *Breast Cancer Wisconsin*, visando à identificação e seleção das

variáveis mais significativas deste conjunto, ou seja, aquelas que sozinhas, ou em grupos menores, podem representá-lo em sua quase totalidade. Esta análise será utilizada posteriormente para o desenvolvimento de modelos de Redes Neurais Artificiais (RNA's) para o auxílio ao diagnóstico de câncer de mama.

MATERIAL E MÉTODOS

O *Breast Cancer Wisconsin* é um conjunto de dados que consolida características obtidas a partir da análise e tratamento computacional de imagens obtidas de tecido tumoral. O tecido tumoral utilizado foi obtido por meio de uma técnica conhecida como FNA (*Fine Needles Aspiration*), um método utilizado para analisar uma pequena quantidade de tecido de um tumor. Uma gota de fluido foi retirada do tecido tumoral, colocada em lâminas e colorizada (Street, 1992). Com o auxílio de um microscópio e de câmeras são geradas imagens que são analisadas por meio de processos computacionais. Após o tratamento computacional foram extraídas dez características das imagens: raio, perímetro, área, compactação, suavidade, concavidade, pontos côncavos, simetria, textura e dimensão fractal. Estas características são utilizadas para avaliar o tecido tumoral e, consequentemente, seu diagnóstico e tratamento. Elas também podem ser utilizadas para o treinamento de modelos de Redes Neurais Artificiais com o objetivo de criar modelos que auxiliem no diagnóstico médico, como feito em Marcano-Cedeño, Quintanilla-Domínguez e Andina (2011).

A construção de modelos de RNA's tem início pela seleção de variáveis de entrada do modelo e, levando em conta apenas duas das características acima mencionadas, o número de modelos possíveis de serem construídos é dado pela combinação de 10 tomados 2 a 2, num total de 45 modelos, um número elevado de modelos, mas possível serem analisados.

Porém para construir um modelo com três características de entrada seriam necessários no mínimo 120 modelos, tornado inviável devido ao custo computacional e, à necessidade de se interpretar os resultados de cada um destes modelos.

Deste modo, escolher variáveis mais representativas significa diminuir o número de modelos necessários para análise.

Para isto, serão abordadas quatro estratégias de análise: (I) descritiva, (II) de correlações lineares, (III) de normalidade e (IV) de dependência entre variáveis.

A análise descritiva do conjunto de dados permite a identificação de *outliers*, o estudo da variabilidade e das amplitudes de escalas dos dados. Géron (2021) expõe que algoritmos de aprendizagem de máquina, sobretudo as RNA's, tem redução de desempenho quando valores numéricos apresentam muitos *outliers* ou estão em escalas muito diferentes. Nesta análise, são utilizados os parâmetros convencionais da estatística descritiva, como média, mediana, desvio-padrão, além dos gráficos de *boxplot*, comparativos que podem indicar uma forma eficiente de se comparar a variabilidade em um conjunto de dados (Devore, 2018). Caso os dados estejam em uma escala de grande variabilidade para a construção do modelo, deverá ser feita uma etapa de pré-processamento, utilizando a normalização ou padronização dos dados. No caso dos *outliers*, será analisado se há uma relação entre a existência destes valores e uma maior malignidade dos tumores.

As correlações lineares serão usadas, a partir do coeficiente r de Pearson, para verificar quais variáveis tem uma maior correlação com a saída desejada do conjunto de dados e o quão significativas são as variáveis para o treinamento do modelo, uma vez que, se duas variáveis têm correlação alta é possível que elas representem uma mesma informação. Em geral, pode-se utilizar uma regra informal para definir uma interpretação para o valor de r (Devore, 2018):

a) Se $-0{,}5 \leq \mathrm{r} \leq 0{,}5$ a correlação é fraca.
b) Se $-0{,}8 < \mathrm{r} < -0{,}5$ ou $0{,}5 < \mathrm{r} < 0{,}8$ a correlação é moderada.
c) Se $\mathrm{r} \leq -0{,}8$ ou $r \geq 0{,}8$, a correlação é forte.

A análise de distribuições consiste em verificar as variáveis que possuem uma distribuição normal de frequências, para isso serão utilizados os métodos QQ-plot e o teste de Shapiro-Wilk.

O QQ-plot é uma representação gráfica que compara os quartis da amostra original com os quartis de uma amostra normal hipotética, entretanto este método tem uma perda de confiabilidade para amostras com menos de 5000 elementos, devendo ser utilizado o teste de Shapiro-Wilk, (Miot,

2007). O teste de Shapiro-Wilk consiste em comparar uma distribuição qualquer com uma distribuição normal, tendo como hipótese nula, a distribuição testada tender à normalidade. Caso os dados não estejam normalmente distribuídos, eles serão padronizados e submetidos ao teste ANOVA. A padronização de uma variável x em uma variável z segue a equação 1:

$$z = \frac{x - \bar{x}}{s} \quad (1)$$

Para a análise de dependência entre as variáveis será utilizado o teste ANOVA (*Analysis of variance*). O ANOVA é um teste paramétrico que assume que os dados populacionais estão normalmente distribuídos (Castanheira, 2023), para verificar se as médias populacionais são iguais, neste sentido a análise de variância tem como hipótese nula as médias populacionais das variáveis serem iguais, a hipótese alternativa será de que as médias populacionais são distintas (Devore, 2018), de modo que se as médias são distintas, então é provável que as variáveis sejam menos dependentes. Para isso é calculado um fator F, dado pela razão entre a variância de todo o conjunto de dados pela média de variância entre as variáveis (Castanheira, 2023). Neste caso, quanto maior o valor de F, maior a chance de se rejeitar a hipótese nula, e quanto mais próximo de 1, maior a chance de se aceitar a hipótese nula. A análise de variância será importante para verificar quais variáveis são mais independentes entre si, ou seja, quais variáveis são independentes em relação à saída obtida (presença ou não de tumor), além de se analisar quais variáveis são mais independentes entre si, possibilitando assim a identificação das características de maior independência, reduzindo a redundância das variáveis, aumentando assim a significância no treinamento de modelos de RNA's.

RESULTADOS E DISCUSSÃO

I) Estatísticas descritivas.

A tabela 1 apresenta um resumo estatístico descritivo das características presentes no conjunto de dados, são apresentadas a média, mediana, desvio padrão e os quartis para cada uma das variáveis.

TABELA 1. Estatísticas descritivas das variáveis do conjunto de dados.

Variáveis	Média	Mediana	Desvio padrão	Q1	Q2	Q3
Raio	14,1273	13,3700	3,5210	11,7000	13,3700	15,7800
Textura	19,2896	18,8400	4,2973	16,1700	18,8400	21,8000
Perímetro	91,9690	86,2400	24,2776	75,1700	86,2400	104,1000
Área	654,8891	551,1000	351,6048	420,3000	551,1000	782,7000
Suavidade	0,0964	0,0959	0,0141	0,0864	0,0959	0,1053
Compactação	0,1043	0,0926	0,0528	0,0649	0,0926	0,1304
Concavidade	0,0888	0,0615	0,0796	0,0296	0,0615	0,1307
Pontos côncavos	0,0489	0,0335	0,0388	0,0203	0,0335	0,0740
Simetria	0,1812	0,1792	0,0274	0,1619	0,1792	0,1957
Dimensão fractal	0,0628	0,0615	0,0071	0,0577	0,0615	0,0661

Nota-se que as variáveis de raio, textura, perímetro e área possuem parâmetros estatísticos em uma escala maior, em valores absolutos, do que as outras variáveis. Casos estas variáveis sejam utilizadas diretamente em um modelo de RNA, haverá uma tendência de menor desempenho dos modelos caso os dados não sejam reescalonados, Géron (2021).

A área é a característica do conjunto de dados com maior número de *outliers*, tendo 30 valores extremos no total, já a suavidade possui 7 *outliers*, sendo a variável com menos valores extremos do conjunto de dados. A Figura 1 apresenta os gráficos *boxplot* das variáveis área e suavidade, respectivamente, onde é possível identificar a dispersão dos dados bem como seus *outliers*.

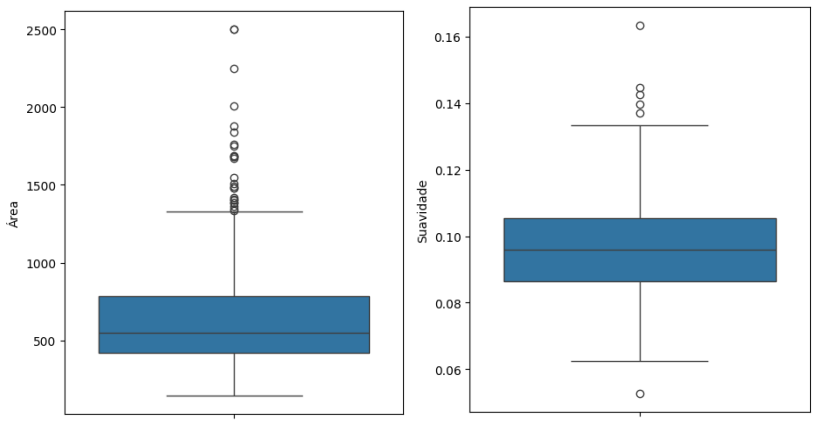

FIGURA 1. Gráficos *boxplot* para as variáveis de área e suavidade.

II) Análise de correlações.

A Figura 2 apresenta a matriz das correlações lineares obtidas para as variáveis do conjunto de dados. Nela é possível observar como todos os pares possíveis de valores de um conjunto de dados estão relacionados entre si, indicando o quão fortemente as variáveis independentes estão relacionadas.

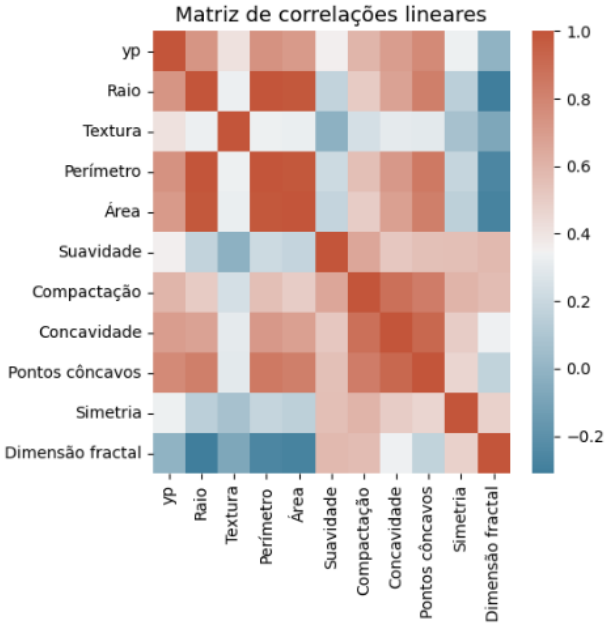

FIGURA 2. Matriz de correlações lineares das variáveis do conjunto de dados.

Em relação à saída desejada, as variáveis de textura, suavidade, simetria e dimensão fractal obtiveram correlações fracas, a dimensão fractal é a que possui a menor correlação do conjunto de dados em relação à saída, com um coeficiente r de aproximadamente -0,0128. Já as variáveis pontos côncavos, perímetro, raio, área, concavidade e compactação, tiveram maior coeficiente r ao serem comparadas com a saída desejada, sendo correlações moderadas. A característica pontos côncavos teve a maior coeficiente r do conjunto de dados, aproximadamente 0,7766.

Ao analisar as correlações lineares entre as características do conjunto de dados nota-se que as variáveis referentes ao raio, perímetro, área e pontos côncavos possuem correlações fortes entre si. Outro grupo obtido de correlações fortes é o grupo das variáveis de compactação, pontos côncavos e concavidade. Neste sentido, apesar de as correlações lineares indicarem apenas uma relação matemática entre as variáveis e não uma causalidade (Quinsler, 2022), a natureza das variáveis, com base na constituição do conjunto de dados (Street, 1992) indicam uma possível relação de dependência

entre as variáveis, o que pode as tornar menos significativas para a construção do modelo, já que elas representariam uma mesma informação.

III) Análise de normalidade.

O teste de normalidade QQ-plot indicou que apenas a textura, a suavidade e a simetria têm um comportamento próximo à normalidade, como pode ser observado na Figura 3.

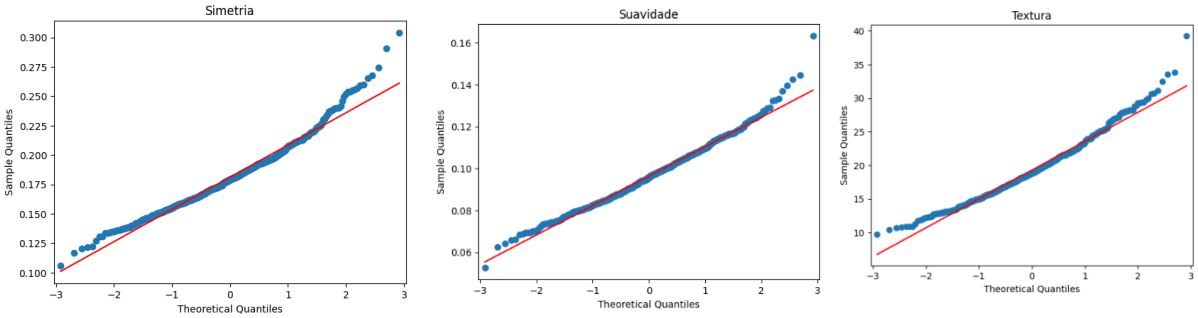

FIGURA 3. Gráficos QQ-plot das variáveis de simetria, suavidade e textura.

Para maior estabilidade dos resultados, todas as variáveis foram testadas utilizando o teste Shapiro-Wilk. Este teste indicou que nenhuma das variáveis do conjunto de dados apresenta tendência à normalidade, isto é, todas as variáveis tiveram um valor p maior menor do que 0,05, valor mínimo para que a hipótese nula fosse aceita. Deste modo será necessário padronizar as variáveis para a realização do teste ANOVA.

IV) Análise de dependência

Os valores obtidos pela análise de variância das variáveis de entrada em relação à saída desejada estão presentes na Tabela 2.

TABELA 2. Fatores F obtidos no teste ANOVA.

Variáveis padronizadas	F
Raio	646,9810
Textura	118,0961
Perímetro	697,2353
Área	573,0607
Suavidade	83,6511
Compactação	313,2331
Concavidade	533,7931
Pontos côncavos	861,6760
Simetria	69,5274
Dimensão fractal	0,0935

Pode-se observar que as variáveis com maior fator F comparando as variáveis de entrada com a saída desejada foram os pontos côncavos, perímetro, raio, área e concavidade, consequentemente estas variáveis têm uma relação de dependência maior com a saída desejada, sendo consideradas variáveis mais representativas para a construção do modelo. A dimensão fractal apresentou menor valor F, indicando assim uma menor dependência entre esta característica e a malignidade dos tumores. Ao analisar o teste ANOVA entre as variáveis de entrada, observou-se que as variáveis: área, raio e perímetro possuem um valor de F muito alto entre si, deste modo a utilização destas características para a construção do modelo será menos significativa, ima vez que apresentam uma maior dependência entre si. Para as outras variáveis, o valor de F foi menor, o que revela uma maior independência destas variáveis.

Consequentemente, as variáveis com maior significância para o treinamento das redes neurais são os pontos côncavos, perímetro, raio, área e concavidade, contudo, é importante notar que construir

um modelo com as características perímetro, raio e área em conjunto apresentam uma redundância de informações, ou seja, o ideal será escolher uma destas variáveis.

CONCLUSÕES

O presente estudo analisou os dados do conjunto de dados Breast Cancer Wisconsin, com intuito de selecionar as variáveis mais relevantes para a construção de um modelo de redes neurais. Observou-se que as variáveis estão em escalas diferentes, deste modo, faz-se necessário um pré-processamento dos dados. Além disso, nenhuma característica apresentou distribuição normal, o que pode tornar a padronização dos dados necessária para testes estatísticos paramétricos. As variáveis com maior correlação linear em relação à saída foram pontos côncavos, perímetro, raio, área, concavidade e compactação, já as características com maior dependência em relação à saída desejada foram, pontos côncavos, perímetro, raio, área e concavidade, porém nota-se que as variáveis perímetro, raio e área possuem uma correlação linear forte e uma dependência estatística alta, de modo que para um modelo de treinamento mais significativo, será necessário utilizar apenas uma destas variáveis.

CONTRIBUIÇÕES DOS AUTORES

L.O.H e M.A.G. contribuíram com a curadoria análise dos dados, assim como a metodologia, experimentos, redação e revisão do trabalho.

AGRADECIMENTOS

Agradecemos a todos que contribuíram direta ou indiretamente a produção deste trabalho.

REFERÊNCIAS

ABU-MOSTAFA, Yaser S.; MAGDON-ISMAL, Malik; LIN, Hsuan-Tien. **Learning From Data**: A short course. New York, NY, USA: AMLBook, 2012.

CASTANHEIRA, Nelson Pereira. **Estatística aplicada a todos os níveis**. 3. ed. Curitiba: Intersaberes, 2023.

CHALA, Luciano Fernandes; URBAN, Linei Augusta Brolini Delle. Rastreamento do câncer de mama. **Radiologia Brasileira**, [*s. l.*], v. 56, ed. 4, Jul/Ago 2023. DOI http://dx.doi.org/10.1590/0100-3984.2023.56.4e1. Acesso em: 4 ago. 2024.

DEVORE, Jay L. **Probabilidade e estatística para engenharia e ciências**: Fundamentos e Aplicações. 9. ed. São Paulo: Cengage, 2018.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-learn, Keras & TensorFlow**: Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. Rio de Janeiro: Alta Books, 2021.

HERNANDES, Lucas de Oliveira; SANTOS, Flavia Milo; GRANERO, Marco Aurélio. Diferenciação e distinção de flores: um estudo por meio de redes neurais artificiais. In: **SEMANA DA MATEMÁTICA E EDUCAÇÃO MATEMÁTICA**, 2024, Guarulhos. Anais ... Guarulhos: IFSP/Campus Guarulhos, 2024.

MARCANO-CEDEÑO, Alexis; QUINTANILLA-DOMÍNGUEZ, Joel; ANDINA, Diego. WBCD breast cancer database classification applying artificial metaplasticity neural network. **Expert Systems with Applications**, [s. l.], v. 38, ed. 8, 17 fev. 2011. DOI https://doi.org/10.1016/j.eswa.2011.01.167. Acesso em: 4 ago. 2024.

MIOT, Hélio Amante. Avaliação da normalidade dos dados em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, [*s. l.*], v. 16, n. 2, abril/dez 2017. DOI http://dx.doi.org/10.1590/1677-5449.041117. Acesso em: 4 ago. 2024.

QUINSLER, Aline Purcote. **Probabilidade e estatística**. 1. ed. Curitiba: Intersaberes, 2022. *E-book*. Disponível em: https://plataforma.bvirtual.com.br. Acesso em: 07 set. 2024.

STREET, W. Nick; WOLBERG, William H; MANGASARIAN, O. L. Nuclear Feature Extraction for Breast Tumor Diagnosis. San Jose: **Biomedical Image Processing and Biomedical Visualization**, 1992.