Técnicas Computacionais para Diagnóstico de Síndrome dos Ovários Policísticos: Uma Análise Comparativa

Roney Nogueira de Sousa[1], **Ana Júlia Lopes de Brito**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) — Campus Fortaleza
Av. Treze de Maio, 2081 — Benfica, Fortaleza–CE, 60040-531

[2]Universidade Federal do Ceará (UFC) — Campus do Porangabussu
R. Alexandre Baraúna, 994 — Rodolfo Teófilo, Fortaleza–CE, 60430-160

{nogueiraroney453, julialopes130703}@gmail.com

Abstract. *This article compares different machine learning classifiers for identifying patients with polycystic ovary syndrome using a public dataset. Techniques such as z-score normalization, outlier removal, and normalization through SMOTE were applied, along with k-fold cross-validation. The Random Forest algorithm stood out, achieving an accuracy of 93.20%.*

Resumo. *Este artigo compara diferentes classificadores de machine learning para identificar pacientes com síndrome de ovários policísticos, utilizando uma base de dados pública. Foram aplicadas técnicas como normalização z-score, remoção de outliers e balanceamento com SMOTE, com validação cruzada k-fold. O algoritmo Random Forest se destacou, atingindo uma acurácia de 93,20%.*

1. Introdução

A síndrome dos ovários policísticos consiste em um distúrbio hormonal comum caracterizado pela hipertrofia dos estromas e surgimento de cistos no córtex ovariano, eventos decorrentes da produção elevada de hormônios androgênicos, como a testosterona [de Souza Pena et al. 2022]. Também é, provavelmente, a maior causa de infertilidade por anovulação, com prevalência aproximada de 2 a 26% das mulheres na faixa etária de 18 a 44 anos, período que corresponde à idade reprodutiva da mulher, em geral [Dabravolski et al. 2021].

Os principais sintomas apresentados são: infertilidade, irregularidade menstrual, aumento no crescimento de pelos, queda de cabelo, acne e manchas na pele. Além disso, a síndrome de ovários policísticos pode acarretar complicações como abortos espontâneos e cânceres, especialmente o endometrial [Raperport and Homburg 2019]. O risco de doenças cardiovasculares, obesidade e a predisposição para diabetes mellitus, que em pacientes com SOP aumenta em até 8,8 vezes, advém da síndrome estar bastante relacionada a um quadro de resistência insulínica [Piccini et al. 2020].

A dosagem de níveis séricos para verificação de hiperandrogenismo é a forma de diagnóstico mais utilizada para identificação da síndrome, juntamente da avaliação clínica e realização da ultrassonografia. Por fim, é essencial frisar a importância do diagnóstico precoce, a fim de evitar tais complicações de longo prazo que comprometem a qualidade de vida dos pacientes [de Souza Pena et al. 2022].

Com base nesses dados apresentados, o objetivo deste estudo foi realizar uma análise comparativa entre os classificadores *Random Forest*, *Multilayer Perceptron* (MLP), k-NN e *Support Vector Machine* (SVM) com diferentes *kernels* (Linear, Polinomial e RBF) para a tarefa de classificação da Síndrome de Ovários Policísticos (SOP). Utilizamos uma base de dados pública denominada "*Polycystic ovary syndrome (PCOS)*"[1], composta por parâmetros físicos e clínicos relevantes para determinar a SOP. Esses dados foram coletados em 10 hospitais distintos na cidade de Kerala, Índia.

2. Metodologia

A presente seção visa mostrar a metodologia adotada no presente estudo.

2.1. Base de dados

O conjunto de dados analisado é intitulado como *"Polycystic Ovary Syndrome (PCOS)"*, e abrange uma ampla gama de parâmetros físicos e clínicos de vários pacientes para a identificação da SOP. Esses dados foram coletados dados de 541 pacientes em 10 hospitais distintos na cidade de Kerala, Índia.

Dentre os parâmetros contidos nesta base de dados, destaca-se a análise de vários hormônios, incluindo, mas não se limitando a, TSH, FSH, entre outros. Além disso, são registradas informações físicas do paciente, abrangendo aspectos como idade e desenvolvimento de pelos no corpo, entre outros.

2.2. Preparação de dados

A base de dados passou por um processo de pré-processamento, no qual os atributos com valores ausentes foram tratados através da inserção da média da coluna, sendo estes correspondentes a apenas 2% da base de dados. Além disso, foram removidos os atributos que não tinham relação com a classificação ou apresentavam redundância. Em seguida, implementou-se a normalização utilizando a técnica *z-score*, seguida pela remoção de *outliers* e balanceamento.

Para o balanceamento dos dados, empregou-se o método *SMOTE*, que gera dados sintéticos para a classe minoritária com base nos vizinhos, sendo esta classe minoritária a classe a qual pacientes não apresentavam a SOP. Adicionalmente, adotou-se o método de validação cruzada *k-fold*, utilizando k=10 *folds*. No que diz respeito aos classificadores utilizados, destaca-se a exploração de vários hiperparâmetros, ajustados por meio da estratégia *GridSearch* para alcançar os melhores resultados.

2.3. Modelos de classificação

Os modelos de classificação foram configurados para otimizar o desempenho e a precisão na tarefa de classificação. A seguir, são detalhadas as configurações para cada modelo:

1. ***Random Forest***: Utilizou-se o critério de entropia em uma floresta com 100 árvores, sem limitação de profundidade. Por fim, estabeleceu-se que o número mínimo de amostras em uma folha da árvore deveria ser igual a 2;

[1] `https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos`

2. ***MLP***: Implementou-se o classificador com duas camadas ocultas, cada uma composta por 50 neurônios, com função de ativação *ReLu*;
3. ***k-NN***: O número de vizinhos foi definido como 5, com todos os vizinhos considerados igualmente. Além disso, utilizou-se distância de Manhattan para calcular a proximidade entre os pontos;
4. ***SVM Linear***: A SVM linear foi definida com fator de regularização igual a 1;
5. ***SVM Polinomial***: A SVM polinomial foi ajustada com um grau de polinômio igual a 2 e fator de regularização igual a 1;
6. ***SVM RBF***: Na SVM RBF, estabeleceu-se um fator de regularização igual a 1 e gamma igual a 0,1.

3. Resultados e Discussões

Os resultados analisados compreenderam as médias das métricas de acurácia, precisão, *recall*, *F1-Score* e AUC, derivadas da aplicação de 10 *folds* à base de dados. A Tabela 1 apresenta detalhadamente os resultados destas métricas.

Tabela 1. Média das métricas após os 10 folds

Modelo	Acurácia (%)	Recall (%)	Precisão (%)	F1-Score (%)	AUC (%)
Random Forest	**93,20**	**93,25**	**93,41**	**93,03**	**97,16**
MLP	75,73	75,53	81,58	76,62	85,61
K-nn	61,17	61,18	65,28	62,44	64,11
SVM Linear	87,38	87,68	87,83	87,53	94,81
SVM Polinomial	31,00	31,07	03,12	14,73	45,25
SVM RBF	65,05	64,85	59,12	60,18	50,00

Analisando as métricas dos classificadores na tarefa de classificação de pacientes com síndrome do ovário policístico, observa-se que o *Random Forest* demonstrou o melhor desempenho global. Destacando-se com uma acurácia de 93,20%, um recall, precisão, e F1-Score consistentes em 93%, evidenciando uma capacidade robusta em identificar corretamente os casos positivos. Além disso, sua AUC atingiu 97,16%, indicando uma excelente capacidade de discriminação entre as classes.

Outros classificadores, como SVM Linear e MLP, também apresentaram desempenhos notáveis, com acurácias de 75% e AUCs superiores a 85%. No entanto, o *Random Forest* se destaca como a escolha mais sólida para essa tarefa específica, considerando sua consistência em todas as métricas avaliadas.

4. Conclusão e Trabalhos Futuros

O artigo sintetiza um trabalho de comparação entre modelos de classificação sobre a SOP, partindo da avaliação de parâmetros médicos contidos em uma base de dados com registros de pacientes. A SOP, por configurar um acometimento complexo e comum a uma parcela significativa da população feminina, evidencia a importância do presente estudo, caso aplicado em práticas médicas, para otimizar o processo de diagnóstico da síndrome.

Feito o tratamento dos dados, que constavam taxas hormonais e aspectos físicos de 541 pacientes, e a comparação entre classificadores, destacou-se o modelo *Random Forest*, por suas métricas consistentemente acima de 93%. Propõe-se, portanto, a utilização

do modelo citado como meio alternativo e complementar para o diagnóstico da SOP, dadas as limitações de outros classificadores, como o *SVM Polinomial* e *SVM RBF*, que apresentaram taxas abaixo do desejável.

Para trabalhos futuros, seria interessante explorar a aplicação desses modelos em um ambiente clínico real para observar o seu comportamento. Além disso, a exploração de outras técnicas de pré-processamento de dados e o ajuste de hiperparâmetros podem oferecer oportunidades para melhorar ainda mais o desempenho dos modelos. Finalmente, a comparação com outros modelos de classificação avançados pode fornecer dados adicionais sobre a melhor abordagem para a tarefa de classificação da SOP.

Referências

Dabravolski, S. A., Nikiforov, N. G., Eid, A. H., Nedosugova, L. V., Starodubova, A. V., Popkova, T. V., Bezsonov, E. E., and Orekhov, A. N. (2021). Mitochondrial dysfunction and chronic inflammation in polycystic ovary syndrome. *International journal of molecular sciences*, 22(8):3923.

de Souza Pena, V., Gonçalves, A. C. R., Vieira, I. R., de Sousa, M. R., de Souza, A. C. D., La Croix, L. M. d. O., Fernandes, B. B., and da Cunha Gonçalves, S. J. (2022). Uma análise sobre as características da síndrome dos ovários policísticos: uma revisão de literatura. *Revista Eletrônica Acervo Médico*, 4:e9996–e9996.

Piccini, C. D., dos Santos Michelon, E., Medeiros, A. G., Heinen, C. A., Maranghelli, G. S., and von Eye Corleta, H. (2020). Síndrome dos ovários policísticos, complicações metabólicas, cardiovasculares, psíquicas e neoplásicas de longo prazo: uma revisão sistematizada. *Clinical and Biomedical Research*, 40(3).

Raperport, C. and Homburg, R. (2019). The source of polycystic ovarian syndrome. *Clinical Medicine Insights: Reproductive Health*, 13:1179558119871467.