REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

# PREVISÃO DA SOP USANDO TÉCNICAS DE *MACHINE LEARNING*: UMA ANÁLISE COMPARATIVA DE MODELOS E APLICAÇÃO PRÁTICA

## *PCOS PREDICTION USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE ANALYSIS OF MODELS AND PRACTICAL APPLICATION*

## *PREDICCIÓN DEL SOP USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO: UN ANÁLISIS COMPARATIVO DE MODELOS Y APLICACIÓN PRÁCTICA*

Luiz Fernando da Cunha Silva[1], Wesley dos Santos Silva[1], Samara Martins Nascimento Gonçalves[1], Verônica Maria Lima Silva[2]

e666546
https://doi.org/10.47820/recima21.v6i6.6546

PUBLICADO: 6/2025

**RESUMO**

A Síndrome dos Ovários Policísticos (SOP) é um distúrbio endócrino que afeta mulheres em idade reprodutiva, de difícil diagnóstico devido à sua heterogeneidade clínica e sobreposição de sintomas com outras condições. Este trabalho investiga o uso de técnicas de *Machine Learning* (ML) para aprimorar a precisão diagnóstica da SOP, utilizando um conjunto de dados públicos e comparando os classificadores *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT) e *Random Forest* (RF). Além disso, são aplicadas técnicas de seleção de características e balanceamento de dados para otimização dos modelos. O modelo RF, após a etapa de otimização, apresentou o melhor desempenho, com uma acurácia e F1-*Score* de 93%. Com base nesses resultados, propõe-se a SOP ASSIST, uma API que disponibiliza o diagnóstico preditivo da paciente utilizando o modelo otimizado. Dessa forma, o uso de ML pode contribuir significativamente para o diagnóstico mais preciso da SOP, além disso, a API desenvolvida representa uma ferramenta de apoio à decisão clínica com potencial de aplicação prática na área da saúde.

**PALAVRAS-CHAVE**: Síndrome dos Ovários Policísticos. Saúde Reprodutiva. *Machine Learning.* Algoritmos de Classificação.

***ABSTRACT***

*Polycystic Ovary Syndrome (PCOS) is an endocrine disorder that affects women of reproductive age and is difficult to diagnose due to its clinical heterogeneity and symptom overlap with other conditions. This study investigates the use of Machine Learning (ML) techniques to improve diagnostic accuracy for PCOS, using a public dataset and comparing the classifiers Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF). Additionally, feature selection and data balancing techniques were applied to optimize model performance. The RF model, after the optimization step, showed the best performance, with an accuracy and F1-Score of 93%. Based on these results, we propose SOP ASSIST, an API that provides predictive diagnosis for patients using the optimized model. Thus, the use of ML can significantly contribute to a more accurate diagnosis of PCOS. Furthermore, the developed API represents a clinical decision support tool with practical application potential in healthcare.*

***KEYWORDS:*** *Polycystic Ovary Syndrome. Reproductive Health. Machine Learning. Classification Algorithms.*

[1] Universidade Federal Rural do Semi-Árido (UFERSA).
[2] Universidade Federal da Paraíba - UFPB.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



***RESUMEN***
*El Síndrome de Ovario Poliquístico (SOP) es un trastorno endocrino que afecta a mujeres en edad reproductiva y cuyo diagnóstico es difícil debido a su heterogeneidad clínica y la superposición de síntomas con otras condiciones. Este trabajo investiga el uso de técnicas de Machine Learning (ML) para mejorar la precisión diagnóstica del SOP, utilizando un conjunto de datos públicos y comparando los clasificadores Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT) y Random Forest (RF). Además, se aplicaron técnicas de selección de características y balanceo de datos para optimizar el rendimiento de los modelos. El modelo RF, tras la etapa de optimización, presentó el mejor desempeño, con una precisión y F1-Score del 93%. Con base en estos resultados, se propone SOP ASSIST, una API que proporciona el diagnóstico predictivo de la paciente utilizando el modelo optimizado. De esta manera, el uso de ML puede contribuir significativamente a un diagnóstico más preciso del SOP. Además, la API desarrollada representa una herramienta de apoyo a la decisión clínica con potencial de aplicación práctica en el área de la salud.*

***PALABRAS CLAVE****: Síndrome de Ovario Poliquístico. Salud Reproductiva. Aprendizaje Automático. Algoritmos de Clasificación.*

## INTRODUÇÃO

A Síndrome dos Ovários Policísticos (SOP) é um distúrbio endócrino caracterizado por um conjunto de sintomas, os quais incluem irregularidade menstrual, hiperandrogenismo e a presença de múltiplos cistos nos ovários, sendo comum em mulheres com idade reprodutiva (Rangel *et al*., 2024). Segundo Silva *et al*. (2024), essa síndrome está frequentemente associada a condições como resistência à insulina, obesidade e um maior risco de doenças cardiovasculares, tornando o seu diagnóstico precoce essencial para a adoção de estratégias de manejo adequadas.

O diagnóstico da SOP, conforme apresentado por Che *et al*., (2023), representa um grande desafio médico, em especial devido à sua heterogeneidade clínica e etiológica. Dessa forma, a sobreposição de sintomas com outras condições endócrinas e metabólicas torna a distinção da SOP uma tarefa complexa (Che *et al*., 2023). Sendo assim, as técnicas de *Machine Learning* (ML) surgem como uma abordagem promissora para aprimorar a precisão do diagnóstico dessa síndrome, através do reconhecimento de padrões em dados clínicos, oferecendo suporte à tomada de decisão médica (Suha; Islam, 2023).

Diante desse contexto, este trabalho tem como objetivo investigar a aplicação de técnicas de ML para apoiar o diagnóstico da SOP, por meio da análise comparativa de diferentes modelos de classificação. Para isso, são exploradas estratégias de seleção de características e balanceamento de dados, com o intuito de otimizar o desempenho dos algoritmos utilizados. Além disso, desenvolveu-se uma Interface de Programação de Aplicações (API), denominada SOP ASSIST, que disponibiliza o modelo com melhor desempenho, permitindo sua integração em sistemas médicos e ampliando o potencial de uso prático na área da saúde.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



## 1. TRABALHOS RELACIONADOS

Estudos têm explorado modelos de ML aplicados ao contexto da saúde, sobretudo no diagnóstico de doenças como a SOP. Nesta seção, são elencados trabalhos que utilizam técnicas de ML para apoiar o diagnóstico da síndrome em mulheres de diferentes faixas etárias e características. Sendo assim, é possível destacar Sreejith, Khanna Nehemiah e Kannan (2022), que desenvolvem um sistema de apoio à decisão para auxiliar equipes médicas no monitoramento da SOP. A proposta combina a estratégia de seleção de características com o classificador *Random Forest* (RF), alcançando especificidade de 90,43%, sensibilidade de 89,73% e precisão de 89,81%. Já Nasim *et al*., (2022), apresentam um método para previsão da SOP utilizando o mecanismo otimizado *chi-squared* (CS-PCOS) para seleção de características. O estudo compara dez modelos de ML hiperparametrizados e destaca o desempenho do *Gaussian Naive Bayes* (GNB) que atingiu 100% de acurácia, precisão, recall e F1-*Score*, com tempo de processamento de 0,002 segundos.

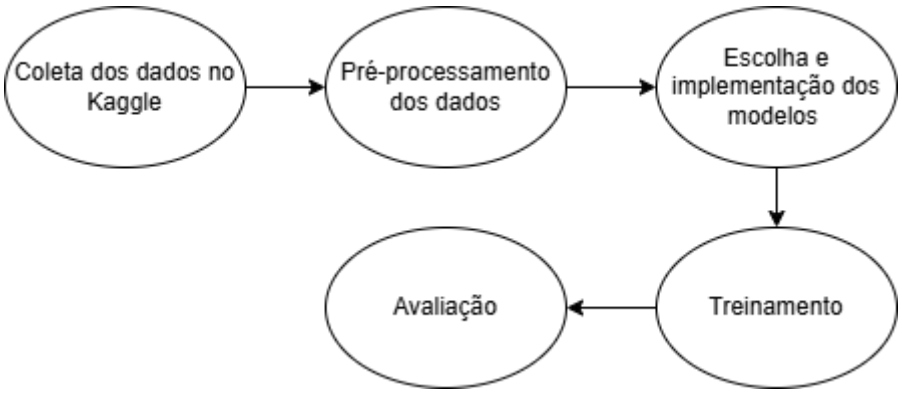
Swamy e KS (2022) propuseram um modelo híbrido, combinando os classificadores *Support Vector Machine* (SVM), RF e *XGBoosting* e avaliaram seu desempenho por meio da comparação com classificadores tradicionais, como *K-Nearest Neighbors* (KNN) e *Decision Tree* (DT). Os autores apontam que o modelo híbrido foi o de maior desempenho, alcançando 93,8% na detecção precoce da SOP. Já Tiwari *et al*., (2022) compararam o desempenho de modelos de ML usando diferentes coeficientes de correlação. Dentre os algoritmos avaliados, o RF obteve a melhor precisão (92,4%) com correlação de 0,8, seguido pelo SVM *linear* (87,1%) e pelo SVM *radial* (71,1%).

É possível destacar, ainda, Suha e Islam (2023) que utilizaram cinco classificadores tradicionais de ML e o modelo de conjunto baseado em boosting como meta-classificador. Os resultados apontam que modelo *stacking* com *Gradient Boosting* (GB) obteve a melhor acurácia, atingindo 95,7% ao utilizar os 25 principais atributos selecionados pela técnica *Principal Component Analysis* (PCA).

## 2. MÉTODOS

A proposta deste estudo incluiu as etapas ilustradas na Figura 1, as quais representam o fluxo de trabalho conduzido, abrangendo a coleta e pré-processamento de dados, bem como a escolha dos algoritmos de ML utilizados para comparação a partir das métricas de desempenho.

**Figura 1:** Configuração do estudo




**Fonte**: Autores.

### 2.1. PCOS *Dataset* e pré-processamento dos dados

Para esse estudo foi utilizado o *dataset Polycystic Ovary Syndrome* (PCOS), que apresenta uma série de 42 características a respeito da síndrome, com 541 mulheres de 10 diferentes hospitais, os quais estão situados na região de Querala, na Índia.

A base de dados PCOS possui variáveis numéricas e categóricas que buscam representar a SOP amplamente, incluindo dados geográficos e físicos, informações clínicas, características menstruais e reprodutivas, além de sinais e sintomas. Além disso, o *dataset* apresenta uma variável-alvo, a PCOS (Y/N), desbalanceada, com 364 pacientes incluídos na classe de não-SOP e 177 na de detecção da síndrome. Dessa forma, é importante perceber que tal configuração pode acabar impactando os resultados das técnicas de ML aplicadas, sendo necessária a implementação de alguma técnica de balanceamento dos dados. Sendo assim, foi utilizada a técnica *Synthetic Minority Oversampling Technique* (SMOTE), a qual busca gerar novos exemplos sintéticos para a classe minoritária, ao invés de somente replicar as instâncias já existentes (Chawla *et al*., 2002). Como resultado, as classes foram equilibradas com 364 amostras para cada uma delas, resultando em valores balanceados que podem impulsionar o desempenho dos modelos treinados.

Por fim, ao se considerar a etapa de tratamento dos dados, todas as variáveis estavam preenchidas, ou seja, não havia valores nulos. Entretanto, algumas possuíam formatos inconsistentes, por exemplo, classes numéricas armazenadas como texto, o que resultou na necessidade de conversão desses valores para seus tipos corretos. Por fim, para a aplicabilidade dos modelos, os dados foram divididos aleatoriamente em conjuntos de treinamento e teste mediante o uso da função *train_test_split* da biblioteca *Scikit-learn*. Esta divisão foi realizada com uma proporção de 70% para o conjunto de treinamento e 30% para o conjunto de teste.

O parâmetro *random_state*, que controla o embaralhamento aplicado aos dados antes da divisão, foi definido como 42 para garantir a reprodutibilidade dos resultados.

### 2.2. *Machine Learning* com Python e *Scikit-learn*

De acordo com Géron (2021), ML é uma área da computação destinada a fazer com que computadores aprendam padrões a partir de dados. Ainda segundo o autor, o ML pode ser dividido em duas diferentes categorias: o aprendizado supervisionado, o aprendizado supervisionado, no qual os dados fornecidos ao modelo são rotulados, e o aprendizado não supervisionado, realizado a partir de dados não rotulados. Nesse contexto, esta pesquisa buscou aplicar diferentes algoritmos supervisionados de ML, voltados para a previsão da SOP. Para isso, foi utilizada a linguagem de programação Python, utilizando a biblioteca *Scikit-learn*.

Dentre os algoritmos treinados, tem-se o SVM, que foi treinado com a função SVC do *Scikit-learn*, a qual foi configurada com a função *Linear Kernel* (LK) – ou seja, as duas classes da variável-alvo foram separadas por meio de uma linha reta (Géron, 2021). Essa configuração permite que o algoritmo lide de maneira eficiente com dados que são linearmente separáveis, otimizando, assim, a simplicidade do modelo, já que não são necessárias transformações complexas. Também foi implementado o KNN, cujo valor escolhido para K seguiu uma abordagem iterativa, na qual foram testados diferentes valores entre 1 e 20. Sendo assim, o modelo foi treinado com cada um desses valores utilizando dados de treinamento previamente selecionados e considerando a acurácia como métrica de avaliação no conjunto de testes. A partir dessa avaliação, foi definido que o melhor valor para o parâmetro K foi 1, dessa forma o modelo foi treinado com a função *KNeighborsClassifier*, da biblioteca *Scikit-learn*, passando o K como parâmetro.

O DT foi treinado através da classe *DecisionTreeClassifie*r da biblioteca *Scikit-learn*. Por padrão, foi considerado o índice de *Gini* para a divisão de critérios, o qual realiza a medição da impureza dos nós e decide quais os pontos de divisão que devem ser considerados. Ademais, não foi especificado um limite de profundidade máxima, o que permite que o modelo cresça até que todos os nós estejam puros, ou contenham uma quantidade menor de amostras do que o mínimo permitido. E, por fim, foi implementado o RF, que foi treinado com a classe *RandomForestClassifier* da biblioteca *Scikit-learn*. Em relação ao número de árvores na floresta foi utilizado o valor padrão de 100, garantindo um equilíbrio eficiente entre desempenho e custo computacional. Assim como no treinamento com o DT, o índice de *Gini* foi utilizado para a medição da qualidade das divisões.

Quanto às métricas de desempenho utilizadas, destacam-se: acurácia, precisão, *recall* e F1-*Score*. Cada uma delas fornece uma perspectiva complementar a respeito da eficiência dos modelos.

### 2.3. Extraindo as melhores características da base de dados

Um conjunto de dados pode apresentar um número elevado de características que explicam a natureza do objeto em análise, resultando em alta dimensionalidade devido ao grande número de atributos. Entre essas variáveis, algumas exercem maior relevância do que outras, permitindo que o modelo alcance maiores níveis de desempenho. Nesse cenário, a seleção de características mais relevantes torna-se imprescindível para potencializar a eficácia de modelos de ML, tendo em vista a obtenção das melhores *features* (Büyükkeçeci; Okur, 2023).

No presente estudo, foi utilizado o método *SelectKBest*, disponível na biblioteca *Scikit-learn*, para selecionar as melhores características da base de dados. A escolha desse método justifica-se por sua simplicidade e eficácia na identificação de atributos com forte correlação com a variável-alvo. Para medir essa correlação, o *SelectKBest* utiliza uma função estatística que atribui uma pontuação a cada atributo com base em sua relevância para a variável-alvo, permitindo, assim, a seleção dos k atributos mais significativos.

Nesse contexto, o hiperparâmetro *score_func,* que define a função de pontuação, foi configurado como *f_classif*, que é função baseada no teste ANOVA apropriada para medir a capacidade discriminativa de atributos em problemas de classificação. O valor k foi definido como 10, após testes experimentais preliminares que apontaram este número como o melhor equilíbrio para a redução de dimensionalidade e retenção de informações cruciais. Com isso, as características com as maiores pontuações foram: *Weight* (Kg), *Cycle* (R/I), AMH (ng/mL), *Weight gain* (Y/N), *Hair growth* (Y/N), *Skin darkening* (Y/N), Pimples (Y/N), *Fast food* (Y/N), *Follicle No.* (L) e *Follicle No.* (R).

## 3. RESULTADOS E DISCUSSÃO

Nesta seção são elencados os resultados obtidos através do treinamento das diferentes técnicas de ML. Além disso, é destacada a construção da SOP ASSIST, uma API para diagnóstico da SOP.

### 3.1. Resultados com o SVM

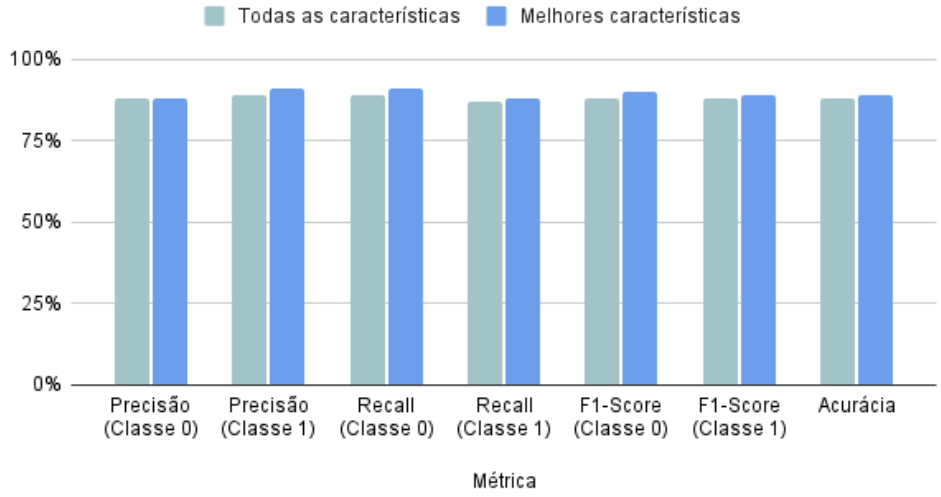
As análises relacionadas ao SVM foram realizadas em dois diferentes cenários: o primeiro considerando todas as características presentes na base de dados e o segundo após a implementação do modelo somente com as melhores *features* selecionadas.

#### 3.1.1. Resultados com todas as características

Ao se utilizar todas as características presentes na base de dados para o treinamento do modelo SVM, os resultados apresentados na Figura 2 foram obtidos a partir da avaliação no conjunto de teste previamente separado. Nesse cenário, o modelo apresentou como resultado

uma acurácia geral de 88%, o que demonstra que o SVM conseguiu se comportar de maneira eficiente no contexto de predição da SOP. No entanto, é de suma importância a análise de outras métricas para a compreensão do modelo, especialmente aquelas que buscam realizar a avaliação com base nas classes definidas pela variável-alvo do *dataset*, como precisão, *recall* e F1-*Score*.

**Figura 2:** Resultados do desempenho com o SVM




**Fonte**: Autores.

Com relação à precisão, o modelo conseguiu atingir o valor de 88% para a classe 0, a qual corresponde a não predição da SOP, e 89% para a classe 1, a de presença da síndrome. Esses resultados demonstram que o modelo se evidenciou eficiente no contexto de minimização dos casos de falsos positivos, para ambas as categorias. Ademais, e ainda no contexto de avaliação em classes, o modelo apresentou como resultado para o *recall* os valores de 89% e 87%, para as classes 0 e 1 respectivamente, demonstrando um desempenho equilibrado em relação à detecção correta dos casos positivos da SOP. Ao se considerar o F1-*Score*, métrica resultante da combinação entre precisão e recall, o resultado foi igual para ambas as classes, com um valor de 88%, demonstrando a consistência dos resultados do modelo treinado com todas as *features*.

### 3.1.2. Resultados com as melhores características

Considerando o contexto de seleção das melhores *features* do *dataset*, conforme o que foi identificado durante o processo de extração de características, percebeu-se que o modelo SVM sofreu um leve aumento em seu desempenho geral, conforme o apresentado na Figura 2. Nessa configuração, o modelo conseguiu atingir uma acurácia média de 89% no conjunto de teste, demonstrando um pequeno aumento em relação ao desempenho obtido com todas as características.

Além disso, em relação à precisão, o modelo obteve os valores de 88% e 91%, para as classes 0 e 1, respectivamente. Dessa forma, esse resultado demonstra que o modelo, ao ser treinado somente com as melhores características, conseguiu melhorar a sua capacidade de minimização dos falsos positivos para a classe relacionada aos casos positivos da doença. O *recall* também apresentou um aumento, agora com 91% para a classe 0 e 88% para a classe 1, evidenciando uma melhora na identificação correta dos casos positivos da SOP, sem que as ocorrências negativas fossem comprometidas. Por fim, ao se analisar o F1-*Score*, foram obtidos os valores de 90% para a classe 0 e 89% para a classe 1. Tais resultados demonstram, dessa maneira, um equilíbrio ainda maior entre a precisão e o recall para o modelo ajustado com as melhores *features*, reforçando a ideia de que a seleção de características auxilia na otimização do desempenho preditivo de modelos de ML.

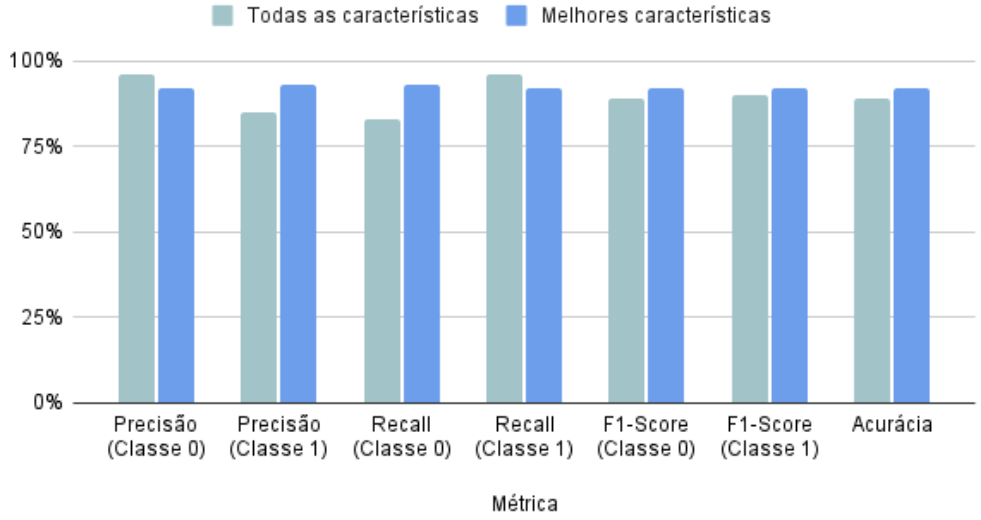
### 3.2. Resultados com o KNN

A seção a seguir apresenta os resultados obtidos ao se utilizar o KNN para o treinamento do modelo. Para isso, a análise também foi realizada considerando a abordagem de treinamento com todas as características da base de dados e com somente as melhores *features*.

#### 3.2.1. Resultados com todas as características

Após a utilização de todas as características no treinamento do modelo KNN, os resultados apresentados na Figura 3 referem-se à sua avaliação no conjunto de teste. Nesse contexto, o modelo obteve uma acurácia geral de 89%, o que indica que o KNN conseguiu desempenhar de forma satisfatória a tarefa de predição da SOP. Entretanto, assim como no caso anterior, é importante investigar métricas complementares para uma melhor compreensão do desempenho do modelo.

**Figura 3:** Resultados do desempenho com o KNN




**Fonte**: Autores.

Com relação à precisão, o KNN conseguiu alcançar um valor de 96% para a classe 0, enquanto, para a classe 1, o modelo atingiu 85%. Tais valores, indicam, assim, que o modelo apresentou um alto desempenho no contexto de redução de falsos positivos para a classe 0, porém, apresentou uma leve queda nesse índice para a classe 1, demonstrando uma precisão maior para a detecção de casos negativos da síndrome. Além disso, ao se observar os valores para o *recall*, o modelo obteve os valores de 83% e 96%, para as classes 0 e 1, respectivamente. Essa análise demonstra que o KNN foi mais eficiente na detecção correta dos casos positivos da SOP, embora apresente uma taxa maior de omissão para a classe 0. Já no que se refere ao F1-*Score*, os valores obtidos se mostram equilibrados, com 89% para a classe 0 e 90% para a classe 1, demonstrando a consistência do modelo treinado com todas as *features*.

3.2.2. Resultados com as melhores características

No contexto de seleção das melhores *features* do conjunto de dados, o KNN obteve um leve aumento em seu desempenho geral, conforme apresentado na Figura 2. Considerando essa abordagem, o modelo conseguiu atingir uma acurácia de 92% no conjunto de teste, demonstrando uma melhoria em relação ao treinamento com todas as características.

Já em relação à precisão, o modelo obteve os valores de 92% e 93% para as classes 0 e 1, respectivamente. Assim, tal observação evidencia um melhor equilíbrio entre a minimização de falsos positivos para ambas as classes. O *recall* também apresentou melhora em seus resultados, com 93% para a classe 0 e 92% para a 1, destacando uma melhora na identificação correta dos casos positivos de SOP, sem que seja comprometida a identificação dos casos negativos. Por fim,

a análise do F1-*Score*, o qual obteve o mesmo valor de 92% para ambas as classes, reflete o equilíbrio mais refinado entre a precisão e o *recall*.

### 3.3. Resultados com o DT

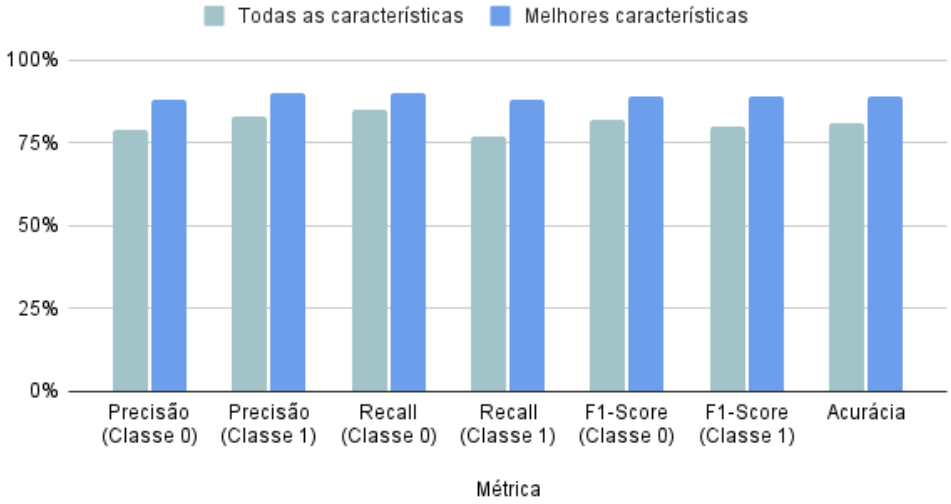
Nesta seção, são apresentados os resultados obtidos a partir do treinamento realizado com o modelo DT. Assim como nas técnicas anteriores, a análise foi conduzida em dois cenários: o primeiro utilizando todas as características disponíveis no *dataset* e o segundo considerando somente as melhores features selecionadas.

#### 3.3.1. Resultados com todas as características

Utilizando todas as características do conjunto de dados para o treinamento da DT, os desempenhos apresentados na Figura 4 foram aferidos no conjunto de teste. Nesse caso, o modelo alcançou uma acurácia de 81%, indicando um desempenho satisfatório na predição da SOP, com uma taxa de acertos consistente para o cenário analisado. De maneira semelhante à técnica previamente analisada, a avaliação de métricas complementares é fundamental para uma compreensão mais aprofundada do desempenho do modelo.

No que diz respeito à precisão, o modelo alcançou 79% para a classe 0 e 83% para a classe 1. Isso evidencia que a técnica alcançou uma taxa de acerto satisfatória, com um desempenho superior na identificação de casos positivos da doença (classe 1) em comparação aos casos negativos (classe 0). Por sua vez, em relação à métrica *recall*, o modelo atingiu um desempenho de 85% para o rótulo 0 e 77% para o rótulo 1. Esses resultados indicam um desempenho intermediário no que diz respeito ao reconhecimento de casos positivos da doença, com destaque para o rótulo 1, que obteve uma avaliação mais detalhada. Com relação à métrica F1-*Score*, o classificador atingiu 82% e 80%, respectivamente, para os rótulos 0 e 1, demonstrando um desempenho consistente para ambos os rótulos.

**Figura 4:** Resultados do desempenho com o DT




**Fonte**: Autores.

3.3.2. Resultados com as melhores características

Para o cenário de treinamento do modelo considerando a seleção de *features* mais relevantes, foi possível identificar um aumento geral de desempenho das métricas utilizadas no estudo, conforme ilustrado na Figura 4. Nesse contexto, o classificador atingiu 89% de acurácia no conjunto de teste, evidenciando resultados superiores ao treinamento com todos os atributos como entrada.

Em relação ao desempenho na métrica de precisão, o modelo atingiu 88% para o rótulo 0 e 90% para o rótulo 1. Esses resultados indicam que, ao ser treinado somente com as características mais relevantes, o modelo conseguiu aumentar a precisão das previsões para os casos positivos de SOP. Da mesma forma, quanto à métrica *recall*, o modelo apresentou resultados ainda mais sólidos, atingindo 90% para a classe 0 e 88% para a classe 1, evidenciando sua eficácia em identificar corretamente a maioria dos exemplos positivos em ambas as classes. Por fim, na métrica de avaliação F1-*Score*, o classificador também obteve um desempenho superior, alcançando 89% para ambos os rótulos. Esse resultado indica que o modelo teve um desempenho significativamente melhor quando comparado ao treinamento com todas as *features*, reforçando a importância da seleção adequada de características para otimizar o desempenho do classificador.

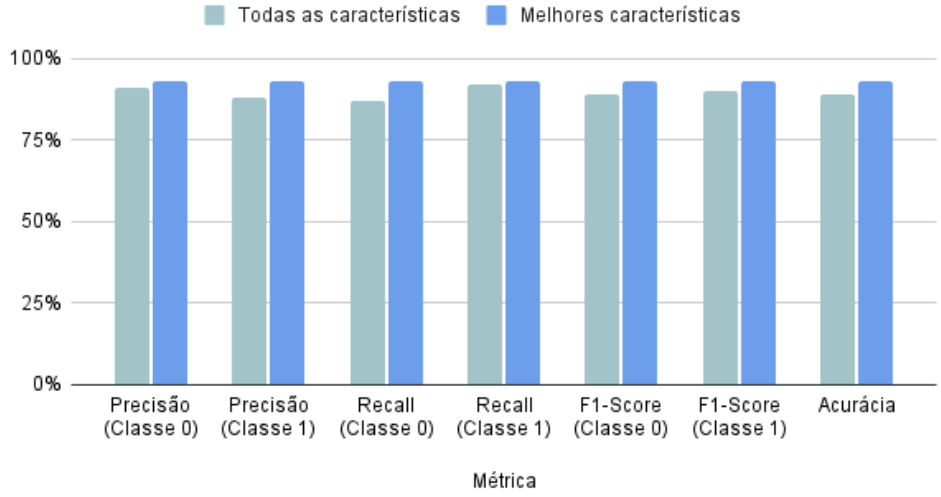
### 3.4. Resultados com o RF

A seção a seguir apresenta os resultados obtidos a partir do treinamento do modelo RF, considerando todas as características da base de dados e somente as melhores *features*.

3.4.1. Resultados com todas as características

Ao utilizar todas as características disponíveis no conjunto de dados para o treinamento do RF, os resultados apresentados na Figura 5 corresponderam à *performance* no conjunto de teste. Nessa situação, o modelo obteve uma acurácia geral de 89%, demonstrando um bom desempenho para a predição da SOP. Entretanto, assim como nas situações anteriores, é importante a avaliação de outras métricas para um melhor entendimento do modelo.

**Figura 5:** Resultados do desempenho com o RF




**Fonte**: Autores.

Com relação à precisão, o RF atingiu os valores de 91% e 88% para as classes 0 e 1, respectivamente, indicando uma boa capacidade de minimização dos falsos positivos, especialmente para a classe de não predição da síndrome. Em termos de *recall*, o modelo atingiu 87% para a classe 0 e 92% para a classe 1, demonstrando um desempenho equilibrado na identificação correta dos casos positivos e negativos. Por fim, ao se analisar o F1-*Score*, foram obtidos os valores de 89% e 90% para as classes 0 e 1, respectivamente, reforçando a consistência desse modelo.

3.4.2. Resultados com as melhores características

Após a extração das melhores *features*, o RF apresentou uma otimização geral nas métricas de desempenho quando avaliado sobre o conjunto de teste, conforme ilustrado na Figura 5. Com isso, o modelo atingiu uma acurácia de 93%, evidenciando um desempenho superior após a utilização das melhores características.

Em termos de precisão, o modelo apresentou valores de 93% para ambas as classes, indicando um alto grau de confiabilidade na minimização de falsos positivos. Além disso, o *recall*

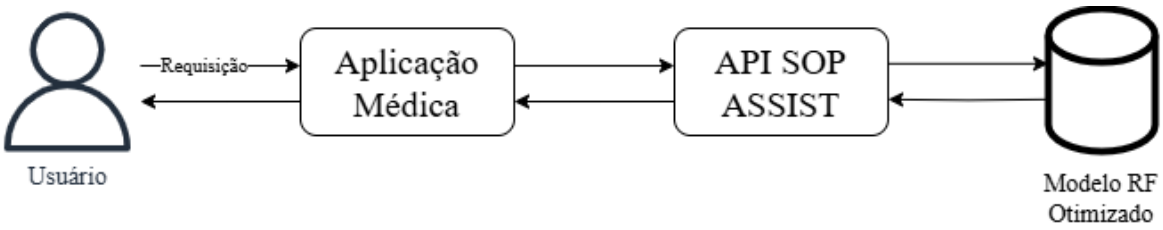
também apresentou uma melhora, com 93% para ambas as classes, reforçando a capacidade do modelo identificar corretamente os casos positivos e negativos da SOP. Por fim, o F1-*Score* obteve os valores de 93% para ambas as classes, destacando o equilíbrio obtido entre a precisão e *recall*.

### 3.5. Resultados da SOP ASSIST

A SOP ASSIST corresponde a uma API, implementada utilizando o *framework* Flask, que permite a disponibilização de um modelo de inteligência artificial para diagnóstico da SOP. Sendo assim, foram usados os melhores resultados alcançados, ou seja, o RF otimizado com as melhores *features*, baseado nos testes já realizados anteriormente.

Em relação a seu funcionamento, mostrado na Figura 6, a SOP ASSIST consegue receber dados do usuário, os quais podem ser tratados e armazenados por uma aplicação médica, no formato JSON, contendo as variáveis clínicas requeridas pelo modelo RF otimizado, e retorna uma predição binária indicando a presença ou ausência da SOP, acompanhada de um *score* de confiança, o qual representa a precisão da predição estabelecida pelo modelo.

**Figura 6:** Funcionamento da API SOP ASSIST




**Fonte**: Autores.

Como forma de validar a robustez e a confiabilidade da API, foram conduzidos testes funcionais e preditivos especificamente na rota de inferência. Os testes seguiram uma abordagem sistemática, com o envio de requisições contendo diferentes conjuntos de dados sintéticos, abrangendo amostras positivas e negativas para a SOP. Desse modo, cada requisição foi avaliada quanto à consistência da predição, verificando se os resultados estavam conforme os rótulos disponíveis nos dados de teste; e a estabilidade do modelo, observando o comportamento do score de confiança para casos ambíguos.

Como resultado, foi possível verificar que a SOP ASSIST processa corretamente as requisições, fornecendo respostas rápidas e coerentes com as expectativas. Desse modo, observou-se que o modelo integrado a API é capaz de lidar com variações nas entradas e ainda assim manter um bom desempenho preditivo. Com isso, reforça-se o potencial da API para

integração em sistemas médicos, servindo como uma ferramenta eficiente de apoio ao diagnóstico da SOP com base em modelos de ML.

### 3.6. Discussão

Os resultados deste trabalho evidenciam a eficácia das técnicas de ML no diagnóstico precoce da SOP, com destaque para a importância da seleção de características. A análise comparativa dos modelos treinados com todas as características presentes no conjunto de dados e com somente as características selecionadas, mostrou um aumento significativo no desempenho geral dos classificadores. Desse modo, tal resultado reforça a relevância dessa estratégia para aprimorar a eficácia dos modelos e obter resultados mais consistentes. Nesse contexto, o modelo SVM apresentou um desempenho consistente em termos de acurácia, alcançando 88% com todas as *features* e 89% ao utilizar somente as mais relevantes. Esse resultado está em conformidade com estudos anteriores, como o de Suha e Islam (2023), que reportaram uma taxa de acerto de 91% na predição da SOP; além de Swamy e KS (2022), que registraram 83%.

O classificador KNN obteve um desempenho superior, com acurácia de 89% ao utilizar todas as variáveis e 92% ao considerar somente as características mais relevantes. Esses valores superam, então, os relatados de investigações anteriores, como os apontados por Suha e Islam (2023), que atingiram 86%, Swamy e KS (2022), com 85%, e Tiwari *et al*., (2022), que registraram um desempenho inferior, com 67%. Já a técnica DT obteve 81% de acurácia ao empregar todas as *features* disponíveis e 89% ao utilizar um subconjunto otimizado. Esse desempenho se mostra superior a trabalhos anteriores, como os de Suha e Islam (2023), que relataram 88%, Swamy e KS (2022), com 75%, e Tiwari *et al*., (2022), que registraram 87%.

Entre os modelos avaliados, o RF destacou-se como o de melhor desempenho, tanto no conjunto completo de variáveis quanto no subconjunto de características mais relevantes. O modelo atingiu 89% de acurácia ao utilizar todas as *features* e 93% ao aplicar seleção de atributos, superando diversos estudos sobre predição da SOP. Por exemplo, Sreejith, Khanna Nehemiah e Kannan (2022) relataram um desempenho de 90%, enquanto Swamy e KS (2022) obtiveram 85%. Além disso, os valores obtidos se alinham com pesquisas como as de Suha e Islam (2023), que registraram 91% e Tiwari *et al*., (2022), com 92% de acurácia.

A seleção de características mais relevantes, realizada por meio do método *SelectKBest*, identificou atributos clinicamente relevantes, como *Weight* (Kg) e AMH (ng/mL), reforçando a importância desses fatores no diagnóstico precoce da SOP. No entanto, o estudo apresenta limitações, como o uso de um conjunto de dados desbalanceado e a validação em um único *dataset*, o que pode limitar a generalização dos resultados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



Dessa forma, os resultados obtidos não apenas confirmam a eficácia das técnicas de ML no diagnóstico da SOP, como também evidenciam o valor da seleção de características na construção de modelos mais precisos e eficientes, reforçando o potencial prático da abordagem proposta como ferramenta de apoio à decisão clínica na área da saúde da mulher.

## 4. CONCLUSÃO

Os resultados obtidos neste estudo demonstram o potencial das técnicas de ML no suporte ao diagnóstico precoce da SOP. Entre os modelos avaliados, o RF apresentou o melhor desempenho, alcançando uma acurácia de 93% com a seleção das melhores características, superando os demais classificadores. O SVM, KNN e DT também obtiveram desempenhos satisfatórios, evidenciando a eficácia desses algoritmos para a classificação da SOP. A seleção de características mostrou-se uma etapa essencial para o aprimoramento dos modelos, permitindo a identificação dos atributos mais relevantes para a predição da síndrome. Além das análises comparativas dos modelos, destaca-se a criação da SOP ASSIST, um modelo de inteligência artificial construído no formato de uma API, visando auxiliar no diagnóstico da SOP.

Apesar dos avanços, algumas limitações devem ser consideradas, como o uso de um único conjunto de dados e a necessidade de validação com novos dados clínicos reais. Sendo assim, novos conjuntos de dados, que tenham impacto clínico significativo para o processo de tomada de decisões, podem ser considerados como também a implementação de uma ferramenta para apoiar tanto a população quanto a equipe médica no diagnóstico e monitoramento da SOP.

Dessa forma, esta pesquisa contribui não apenas para o avanço científico e tecnológico no campo da saúde digital, mas também para a promoção de diagnósticos mais equitativos, acessíveis e eficazes, com potencial de impacto social positivo para a saúde feminina.

## REFERÊNCIAS

BÜYÜKKEÇECI, M.; OKUR, M. C. A comprehensive review of feature selection and feature selection stability in machine learning. **Gazi University Journal of Science**, [*s. l.*], v. 36, n. 4, p. 1506-1520, Dec. 2023. Disponível em: https://doi.org/10.35378/gujs.993763. Acesso em: 23 jan. 2025.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, [*s. l.*], v. 16, p. 321–357, June. 2002. Disponível em: https://doi.org/10.1613/jair.953. Acesso em: 20 jan. 2025.

CHE, Y.; YU, J.; LI, Y. S.; ZHU, Y.; TAO, T. Polycystic Ovary Syndrome: Challenges and Possible Solutions. **Journal of Clinical Medicine**, [*s. l.*], v. 12, n. 4, p. 1500, Feb. 2023. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC9967025/. Acesso em: 8 fev. 2025.

GÉRON, A. **Mãos à Obra**: Aprendizado de Máquina com Scikit-Learn, Keras e TensorFlow. 2. ed. atual. Rio de Janeiro: Alta Books, 2021. 614 p. ISBN 9788550815480.

NASIM, S.; ALMUTAIRI, M. S.; MUNIR, K.; RAZA, A.; YOUNAS, F. A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics. **IEEE Access**, [*s. l.*], v. 10, p. 97610-97624, 2022. Disponível em: https://ieeexplore.ieee.org/document/9885199. Acesso em: 25 jan. 2025.

RANGEL, F. R.; LOPES, C. C. A.; REZENDE, M. C. B.; SALES, C. B.; MAGALHÃES, A. C. T. Síndrome dos Ovários Policísticos: Revisão Sistemática da Etiologia, Fisiopatologia, Diagnóstico e Tratamento. **Brazilian Journal of Implantology and Health Sciences**, [*s. l.*], v. 6, n. 8, p. 5403–541, ago. 2024. Disponível em: https://doi.org/10.36557/2674-8169.2024v6n8p5403-5412. Acesso em: 8 fev. 2025.

SILVA, T. dos S.; OLIVEIRA, M. D. P. de; BRASIL, L. G. O impacto da Síndrome do Ovário Policístico na vida das mulheres. **Brazilian Journal of Health Review**, [*s. l.*], v. 7, n. 5, p. e72576, maio. 2024. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BJHR/article/view/72576. Acesso em: 8 fev. 2025.

SREEJITH, S.; KHANNA NEHEMIAH, H.; KANNAN, A. A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier. **Healthcare Analytics**, [*s. l.*], v. 2, p. 100102, 2022. Disponível em: https://doi.org/10.1016/j.health.2022.100102. Acesso em: 8 fev. 2025.

SUHA, S. A.; ISLAM, M. N. Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique. **Heliyon**, [*s. l.*], v. 9, n. 3, p. e14518, Mar. 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S2405844023017255. Acesso em: 25 jan. 2025.

SWAMY, S. R.; KS, N. P. Hybrid Machine Learning Model for Early Discovery and Prediction of Polycystic Ovary Syndrome. *In:* INTERNATIONAL CONFERENCE ON ADVANCED TECHNOLOGIES IN INTELLIGENT CONTROL, ENVIRONMENT, COMPUTING & COMMUNICATION ENGINEERING (ICATIECE), 2., 2022, Bangalore, India. **Anais eletrônicos** [...]. Bangalore: IEEE, 2022. p. 1-8. Disponível em: https://www.proceedings.com/67995.html. Acesso em: 27 jan. 2025

TIWARI, S.; KANE, L.; KOUNDAL, D.; JAIN, A.; ALHUDHAIF, A.; POLAT, K.; ZAGUIA, A.; ALENEZI, F.; ALTHUBITI, S. A. SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. **Expert Systems with Applications**, [*s. l.*], v. 203, p. 117592, 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417422009046. Acesso em: 30 jan. 2025.