Received: 14 Oct 2025 | Revised: 16 Dec 2025 | Accepted: 20 Jan 2026 | Available online: 26 Mar 2026
https://doi.org/10.13057/ijas.v8i2.109931

E-ISSN: 2621-086X
https://jurnal.uns.ac.id/ijas

Indonesian Journal of Applied Statistics Vol. 8, No. 2, pp. 148-159, 2025

Deteksi *Polycystic Ovary Syndrome* (PCOS) Berbasis *Machine Learning*: Kombinasi SMOTE, *Random Forest, Gradient Boosting*, dan *Bayesian Optimization*

Naufalia Alfiryal*, Kusman Sadik, Cici Suhaeni, Agus Mohamad Soleh
Departemen Statistika dan Sains Data, IPB University, Bogor, Indonesia
*Corresponding author: naufaliaalfiryal@apps.ipb.ac.id

Abstrak

Polycystic ovary syndrome (PCOS) merupakan gangguan endokrin yang umum terjadi pada wanita usia reproduktif. Kondisi ini dapat menyebabkan gangguan ovulasi, ketidakseimbangan hormon, resistensi insulin, serta meningkatkan risiko penyakit kardiovaskular, obesitas, dan gangguan psikologis. Meskipun prevalensinya cukup tinggi, sekitar 75% kasus PCOS masih belum terdiagnosis dalam praktik klinis akibat kompleksitas gejala dan keterbatasan metode diagnosis yang digunakan saat ini. Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan pendekatan berbasis machine learning guna meningkatkan akurasi dan efisiensi deteksi PCOS. Penelitian ini membandingkan performa dua algoritma pembelajaran terawasi, yaitu random forest dan gradient boosting, dalam melakukan prediksi PCOS. Dataset yang digunakan diperoleh dari repositori publik dan memuat berbagai fitur klinis yang berkaitan dengan PCOS. Untuk menangani permasalahan ketidakseimbangan kelas, metode synthetic minority over-sampling technique (SMOTE) diterapkan pada data pelatihan. Selain itu, bayesian optimization digunakan untuk melakukan penyetelan hiperparameter pada masing-masing model agar diperoleh performa yang optimal. Evaluasi performa model dilakukan menggunakan beberapa metrik, dengan area under the curve–receiver operating characteristic (AUC-ROC) sebagai metrik utama. Hasil penelitian menunjukkan bahwa model Gradient Boosting memberikan performa terbaik dengan nilai AUC sebesar 0,8983 dan nilai recall sebesar 0,95, yang mengindikasikan sensitivitas tinggi dalam mengidentifikasi kasus PCOS. Temuan ini menunjukkan bahwa kombinasi SMOTE dan bayesian optimization efektif dalam meningkatkan akurasi prediksi, khususnya pada dataset medis yang tidak seimbang. Pendekatan yang diusulkan memiliki potensi untuk diintegrasikan ke dalam sistem pendukung keputusan klinis guna mendukung proses skrining PCOS yang lebih dini dan andal.

Polycystic ovary syndrome (PCOS) is a common endocrine disorder among reproductive-aged women. This condition can lead to ovulatory dysfunction, hormonal imbalance, insulin resistance, and an increased risk of cardiovascular disease, obesity, and psychological disorders. Despite its high prevalence, approximately 75% of PCOS cases remain undiagnosed in clinical settings due to the complexity of symptoms and limitations of current diagnostic methods. To address this issue, a machine learning-based approach is proposed to improve the accuracy and efficiency of PCOS detection. This study compares the performance of two supervised learning algorithms random forest and gradient boosting for PCOS prediction. The dataset used was obtained from a public repository and contains various clinical features associated with PCOS. To address the class imbalance problem, the synthetic minority over-sampling technique (SMOTE) was applied to the training data. Additionally, bayesian optimization was employed to fine-tune the hyperparameters of each model for optimal performance. Model performance was evaluated using several metrics, with the area under the curve–receiver operating characteristic (AUC-ROC) as the primary measure. The Gradient Boosting model achieved the best results, with an AUC of 0.8983 and a recall of 0.95, indicating high sensitivity in identifying positive PCOS cases. These findings demonstrate that the combination of SMOTE and Bayesian Optimization is effective in enhancing predictive accuracy, especially in imbalanced medical datasets. The proposed approach shows promise for integration into clinical decision-support systems to facilitate earlier and more reliable PCOS screening.

Kata Kunci: Bayesian optimization, gradient boosting, PCOS, random forest, SMOTE.
Keywords: *Bayesian optimization, gradient boosting, PCOS, random forest, SMOTE.*

How to Cite:
N. Alfiryal, K. Sadik, C. Suhaeni, and A. M. Soleh, "Deteksi polycystic ovary syndrome (PCOS) berbasis machine learning: kombinasi SMOTE, random forest, gradient boosting, dan bayesian optimization," *Indonesian Journal of Applied Statistics,* vol. 8, no. 2, pp. 148-159, 2025, doi: 10.13057/ijas.v8i2.109931.

1. PENDAHULUAN

Polycystic ovary syndrome (PCOS) merupakan gangguan endokrin yang pada umumnya terjadi pada wanita usia reproduktif. Kondisi ini dapat menyebabkan gangguan ovulasi, ketidakseimbangan hormon, resistensi insulin, hingga peningkatan risiko penyakit kardiovaskular, obesitas, dan gangguan psikologis. PCOS umumnya dapat didiagnosis berdasarkan pada *Rotterdam Criteria* yang mengharuskan keberadaan minimal dua dari tiga karakteristik utama, yaitu gangguan ovulasi, hiperandrogenisme (pertumbuhan rambut berlebih dan kadar testosterone yang tinggi), dan PCOM yang ditandai dengan ≥ 20 folikel atau volume ovarium ≥ 10 cm3 [1].

Dilansir pada laman situs resmi *World Health Organization*, prevalensi terjadinya PCOS pada wanita usia reproduktif sekitar 6-13%. Namun, meskipun prevalensinya tinggi, diperkirakan sekitar 75% kasus PCOS tidak dapat terdeteksi atau terdiagnosis dengan benar dalam praktik klinis. Hal tersebut disebabkan oleh beragamnya gejala yang dialami setiap individu serta keterbatasan dalam metode diagnostik yang ada saat ini. Oleh karena itu, diperlukan metode alternatif yang dapat membantu proses diagnosis dengan lebih efisien dan akurat.

Machine learning telah banyak diterapkan dalam berbagai bidang, termasuk bidang kesehatan sebagai alat bantu dalam prediksi dan klasifikasi penyakit. Di antara algoritma yang sering digunakan dalam proses prediksi kesehatan adalah random forest dan gradient boosting. Random forest merupakan metode *ensemble* yang menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang lebih stabil dan mengurangi risiko *overfitting*. Sedangkan, gradient boosting merupakan metode yang bekerja dengan mengoptimalkan kesalahan residual secara bertahap, sehingga mampu meningkatkan akurasi prediksi secara signifikan [2]. Namun, performa kedua metode tersebut bergantung pada pemilihan hiperparameter yang optimal. Untuk meningkatkan performa model, penelitian ini menerapkan *Bayesian optimization* sebagai teknik tuning hiperparameter.

Bayesian optimization merupakan metode yang cukup popular digunakan dalam pemilihan hiperparameter dikarenakan kemampuannya untuk mendapatkan nilai optimal dengan cepat. Algoritma ini bekerja dengan cara menyesuaikan model probabilitas terhadap nilai yang diamati dari target yang akan dioptimalkan dengan pendekatan ini, diharapkan model dapat mencapai akurasi yang lebih optimal dalam mendeteksi PCOS [3].

Selain tantangan dalam pemilihan hiperparameter, salah satu kendala utama dalam prediksi PCOS menggunakan *machine learning* adalah ketidakseimbangan kelas dalam dataset. Kondisi ini terjadi ketika jumlah pasien PCOS jauh lebih sedikit dibandingkan individu sehat, sehingga model lebih cenderung mengklasifikasikan individu sebagai non-PCOS. Untuk mengatasi masalah ini, penelitian ini menggunakan *synthetic minority over-sampling technique* (SMOTE) yang merupakan teknik *oversampling* yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset klasifikasi [4].

Beberapa studi sebelumnya telah menunjukkan keberhasilan penerapan *machine learning* dalam meningkatkan akurasi prediksi, baik pada bidang teknik maupun medis. Penelitian mengenai prediksi kebisingan airfoil oleh NASA menunjukkan bahwa algoritma *random forest* dan *gradient boosting* mampu menghasilkan performa yang tinggi, khususnya ketika dikombinasikan dengan validasi silang [2]. Selain itu, penelitian terkait deteksi kanker paru-paru melaporkan bahwa penerapan tuning hiperparameter menggunakan *Bayesian optimization* dapat meningkatkan performa model secara signifikan, dengan nilai AUC pada model random forest mencapai 0,99974 [3].

Meskipun pendekatan-pendekatan tersebut telah berhasil diterapkan pada berbagai kasus, penerapanya dalam konteks PCOS dengan tantangan unik, seperti ketidakseimbangan kelas dalam data masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa *random forest*

dan *gradient boosting* dalam mendeteksi PCOS, serta mengevaluasi dampak penggunaan SMOTE dan *Bayesian optimization* terhadap peningkatan akurasi model prediksi.

2. METODE

2.1. Bahan Data

Dataset yang digunakan dalam penelitian ini diperoleh dari laman Kaggle yang berisi informasi klinis terkait dengan pasien *polycystic ovary syndrome* (PCOS). Dataset ini terdiri dari 1000 observasi dengan 5 variabel utama yang pada umumnya berkaitan dengan faktor risiko dan diagnosis PCOS, yakni *age* (umur), *body mass index* (indeks massa tubuh), *menstrual irregularity* (ketidakraturan menstruasi), *testosterone level* (tingkat testosterone), dan *antral follicle count* (jumlah folikel antral) sebagai variabel bebas, serta PCOS diagnosis (diagnosis PCOS) sebagai variable target yang bernilai kategorik biner yang terdiri dari 2 kelas, kelas 0 untuk pasien tidak terdiagnosis PCOS dan kelas 1 untuk pasien yang terdiagnosis PCOS dengan tingkat ketidakseimbangan 80,1%:19,9%.

2.2. Metode Penelitian

Adapun tahapan analisis yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Melakukan pengambilan data PCOS diagnosis pada laman kaggle.
2. Melakukan data *pre-processing*:
   - Deteksi missing value, apabila terdapat *missing value*, maka lakukan imputasi menggunakan nilai rata-rata (untuk variabel kontinu) atau nilai modul (untuk variabel diskrit), apabila tidak ada lanjutkan dengan deteksi data duplikat.
   - Deteksi data duplikat, apabila terdapat data duplikat, maka lakukan penghapusan pada salah satu data yang sama, apabila tidak ada maka lanjutkan pada tahapan berikutnya.
   - Deteksi pencilan, dilakukan pemeriksaan awal terhadap kemungkinan pencilan secara deskriptif. Namun, karena algoritma *random forest* dan *gradient boosting* relatif *robust* terhadap keberadaan pencilan, serta tidak ditemukan pencilan ekstrem yang berpotensi mendistorsi model, maka tidak dilakukan penghapusan data pencilan pada penelitian ini.
3. Membagi dataset menjadi data train dan data test dengan perbandingan 80:20.
4. Melakukan balancing data menggunakan algoritma SMOTE sehingga distribusi kelas 0 dan 1 seimbang.
5. Sebagaimana dalam penelitian yang dilakukan oleh Fulazzaky, Saefuddin, dan Soleh (2024) [5] proses *balancing* data dilakukan setelah pembagian dataset dan hanya diterapkan pada data *training* saja untuk menghindari kebocoran data.
6. Membentuk model pohon klasifikasi untuk SMOTE *random forest* dan SMOTE *gradient boosting*.
7. Melakukan penyetelan hiperparameter menggunakan *Bayesian optimization* pada data pelatihan. Proses ini bertujuan untuk menentukan kombinasi parameter terbaik berdasarkan hasil evaluasi validasi silang.
8. Menentukan model terbaik berdasarkan performa pada data pelatihan hasil proses validasi silang.
9. Evaluasi performa akhir menggunakan data pengujian dengan menerapkan model terbaik yang telah diperoleh, menggunakan metrik akurasi, presisi, *recall*, *F1-score*, AUC, dan kurva ROC. Menentukan model terbaik di antara SMOTE *random forest* dan SMOTE *gradient boosting*.

2.3. Polcystic Ovary Syndrome (PCOS)

Polycystic ovary syndrome (PCOS) merupakan gangguan endokrin yang pada umumnya terjadi pada wanita usia reproduktif [1]. PCOS disebabkan oleh berbagai faktor, baik faktor lingkungan maupun epigenetik [6]. Kondisi ini dapat menyebabkan gangguan ovulasi, ketidakseimbangan hormon, resistensi insulin, hingga peningkatan risiko penyakit kardiovaskular, obesitas, dan gangguan psikologi. PCOS umumnya dapat didiagnosis berdasarkan pada *Rotterdam Criteria* yang mengharuskan keberadaan minimal dua dari tiga karakteristik utama, yaitu gangguan ovulasi, hiperandrogenisme (pertumbuhan

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

rambut berlebih dan kadar testosterone yang tinggi), dan PCOM yang ditandai dengan ≥ 20 folikel atau volume ovarium ≥ 10 cm^3 [1].

2.4. Random Forest

Random forest merupakan metode *ensemble* yang dapat digunakan untuk klasifikasi dan prediksi. *Random forest* menggunakan teknik *bootsrap aggregating* atau *bagging* yang digunakan untuk membentuk sejumlah pohon keputusan, kemudian menggabungkan hasil prediksi dari pohon-pohon tersebut. *Random forest* memiliki keunggulan, seperti tidak terlalu rentan terhadap *overfitting* dan pencilan, efisien secara komputasi, serta cocok untuk data berdimensi tinggi [2][7].

Dalam melakukan pemilihan fitur terbaik dalam pembentukan pohon keputusan dilakukan menggunakan indeks Gini dengan persamaan sebagai berikut [3]:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \tag{1}$$

dengan P_i merupakan proporsi atribut terhadap masing-masing kelas, m sebagai jumlah kelas. Kemudian, dihitung total indeks Gini pada node internal *K* melalui persamaan di bawah ini:

$$Total\ Indeks\ Gini(K) = \frac{T_1}{T} Gini(D_1) + \frac{T_2}{T} Gini(D_2) \tag{2}$$

di mana T_1 dan T_2 secara berturut-turut merupakan jumlah data pada kelas pertama dan kedua, dan T merupakan keseluruhan jumlah data.

Penerapan indeks Gini dalam pembentukan setiap pohon keputusan dalam *random forest* sangat penting, tetapi yang paling menentukan efektivitas algoritma ini adalah tingkat kesalahannya yang bergantung pada dua faktor berikut [8].

a. Korelasi antara dua pohon keputusan, tingkat kesalahan akan meningkat jika dan hanya jika korelasi antara dua pohon apapun dari hutan acak tersebut meningkat.
b. Kekuatan pohon keputusan, semakin rendah tingkat kesalahan maka akan semakin kuat pohon yang juga akan memperkuat hutan acak. Begitupun sebaliknya.

2.5. Gradient Boosting

Gradient boosting adalah sebuah algoritma *boosting* yang kuat dengan jalan menggunakan beberapa pembelajar yang lemah sehingga memperoleh pembelajar yang kuat. Ide utama dari metode ini adalah meningkatkan kemampuan prediksi *ensemble* secara bertahap dengan mengatasi kesalahan yang dibuat oleh model-model sebelumnya [2].

Secara matematis, *gradient boosting* dapat dirumuskan sebagai berikut [9]:

$$\tilde{F}(x) = argmin\ L(y, F(x)) \tag{3}$$

dengan $\tilde{F}(x)$ merupakan fungsi aproksimasi yang memetakan himpunan variable input $x = \{x_1, x_2, \dots, x_n\}$ pada variabel respon y dengan cara meminimalkan fungsi kerugian $L(y, F(x))$. Pada tahap awal, akan dibentuk sebuah model dasar $F_0(x)$ yang pada umumnya merupakan fungsi konstanta. Kemudian, dilakukan proses iterasi di mana pada setiap iterasi ke-*m*, residual akan dihitung sebagai *negative gradient* dari fungsi kerugian dengan menggunakan persamaan di bawah ini:

$$\tilde{y}_i = -\left[\frac{\delta L\left(y_i, F(x_i)\right)}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \dots, N \tag{4}$$

Model akan diperbarui dengan menambahkan hasil prediksi pembelajar lemah terhadap model sebelumnya. Melalui prinsip inilah, *gradient boosting* dapat memperbaiki akurasi secara bertahap hingga memperoleh performa yang optimal.

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

Gradient boosting unggul dalam daya prediksi yang tinggi. Akan tetapi, kemungkinan besar akan mengalami *overfitting* apabila dataset memiliki *noise*. Maka dari itu, untuk mengurangi permasalahan tersebut perlu dilakukan pengoptimalan hiperparameter, seperti *learning rate* yang lebih rendah dapat menghindari *overfitting* dan *subsample* yang kurang dari 1 dapat mengurangin *varians* dan meningkatkan bias [10].

2.6. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE adalah teknik *oversampling* yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset klasifikasi. Algoritma ini bekerja dengan membuat sampel sintetik berdasarkan titik data dari kelas minoritas menggunakan interpolasi dari titik-titik terdekat (*nearest neighbors*). Hal ini membantu dalam meningkatkan performa model klasifikasi dengan memberikan keseimbangan dalam distribusi data kelas. SMOTE membangkitkan data dengan menggunakan persamaan berikut [4]:

$$x_{syn} = x_i + (x_{knn} - x_i)\gamma \tag{5}$$

dengan x_{syn} sebagai pengamatan baru hasil pembangkitan, x_i sebagai pengamatan ke-i, x_{knn} sebagai x dengan jarak terdekat dari x_i, dan γ sebagai bilangan acak yang ada pada rentang 0 dan 1 [4][11].

Teknik SMOTE memiliki keunggulan, yakni mampu mengurangi kemungkinan *overfitting* yang merupakan salah satu kelemahan dari teknik *oversampling*, tidak menyebabkan pengurangan data sehingga tidak ada informasi yang hilang, dan meningkatkan akurasi pengklasifikasian pada kelas minoritas [12][13]. SMOTE juga telah terbukti efektif dalam berbagai bidang, seperti deteksi penipuan, diagnosis medis, dan manajemen risiko. Selain itu, SMOTE memiliki sifat independent dari algoritma yang menjadikannya langkah *pre-processing* yang serbaguna untuk algoritma klasifikasi apapun [14].

2.7. Bayesian Optimization

Bayesian optimization merupakan metode yang cukup popular digunakan dalam tuning hiperparameter dikarenakan kemampuannya untuk mendapatkan nilai optimal dengan cepat [3][15]. Metode ini menggunakan *Gaussian process* (GP) untuk dapat memahami distribusi hasil secara komprehensif. Algoritma ini bekerja dengan cara menyesuaikan model probabilitas terhadap nilai yang diamati dari target yang akan dioptimalkan. *Bayesian optimization* menggunakan rantai Bayesian untuk memperbarui nilai berdasarkan parameter sebelumnya. Formula Bayes digunakan untuk menghitung distribusi probabilitas posterior, serta varians dan rata-rata akurasi untuk setiap nilai hiperparameter. Akurasi rata-rata yang lebih tinggi menunjukkan model yang lebih kuat [3].

Dalam proses optimasinya, *Bayesian optimization* mengacu pada Teorema Bayes yang diuraikan pada persamaan berikut:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{6}$$

dengan $P(A)$ merupakan probabilitas prior dari A, $P(B)$ merupakan probabilitas dari variabel B yang telah diamati, dan $P(A|B)$ serta $P(B|A)$ merupakan probabilitas posterior di mana kemungkinan dari variabel A yang diberikan B, begitupun sebaliknya [3].

Adapun proses penentuan hiperparameter optimal dilakukan melalui langkah-langkah berikut [3]:

1. Membentuk model probabilitas sebagai pendekatan fungsi objektif.
2. Menemukan dan menerapkan hiperparameter terbaik.
3. Memperbarui model berdasarkan hasil terbaru.
4. Mengulang proses hingga batas iterasi atau waktu tercapai.

2.8. Teknik Evaluasi

Dalam penelitian ini, akan digunakan beberapa teknik evaluasi untuk mengevaluasi performa dari kedua metode yang digunakan, yakni [3]:

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

Akurasi

Akurasi merupakan presentasi prediksi yang benar dari keseluruhan data, baik untuk kelas positif maupun negatif. Adapun persamaan untuk menghitung akurasi sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

dimana,
TP (*True Positive*) = jumlah prediksi positif yang benar
TN (*True Negative*) = jumlah prediksi negatif yang benar
FP (*False Positive*) = jumlah prediksi positif yang salah
FN (*False Negative*) = jumlah prediksi negatif yang salah

Presisi

Presisi mengukur seberapa banyak prediksi positif yang benar dibandingkan dengan seluruh prediksi positif. Adapun persamaan untuk presisi sebagai berikut:

$$Presisi = \frac{TP}{TP + FP} \tag{8}$$

Recall

Recall mengukur kemampuan model dalam menemukan semua data positif yang benar. Semakin tinggi recall, semakin sedikit kasus positif yang terlewat. Adapun persamaan dari recall sebagai berikut:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

F1-score

F1-score adalah rata-rata harmonik antara presisi dan recall, digunakan saat keseimbangan antara keduanya diperlukan. Adapun persamaan untuk F1-score sebagai berikut:

$$F1 - score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \tag{10}$$

F1-score berguna saat ada ketidakseimbangan kelas dalam data. Hal ini dikarena mempertimbangkan baik kesalahan dalam mendeteksi kelas positif (*false negative*) maupun memprediksi positif secara salah (*false positive*).

Kurva ROC (Receiver Operating Characteristing)

Kurva ROC (*receiver operating characteristing*) atau disebut juga sebagai nilai AUC (*area under the curve*) merupakan visualisasi yang digunakan untuk mengevaluasi performa prediksi hasil dan membandingkan hasil klasifikasi. Kurva ROC terbagi menjadi dua sumbu dengan sumbu x sebagai tingkat positif yang salah (*false positive*) dan sumbu y sebagai tingkat positif sebenarnya (*true positive*). AUC dapat dianggap sebagai probabilitas area di bawa kurva ROC. Semakin tinggi nilai AUC, maka semakin efektif teknik klasifikasi yang digunakan. Nilai AUC selalu berada pada rentang 0 sampai 1. Adapun kategori dari nilai AUC dapat dilihat pada Tabel 1.

Tabel 1. Kategori nilai AUC

Nilai AUC	Kategori
0,90 – 1,00	Klasifikasi Sangat Baik

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

Nilai AUC	Kategori
0,80 – 0,90	Klasifikasi Baik
0,70 – 0,80	Klasifikasi Cukup
0,60 – 0,70	Klasifikasi Buruk
0,50 – 0,60	Tidak Baik

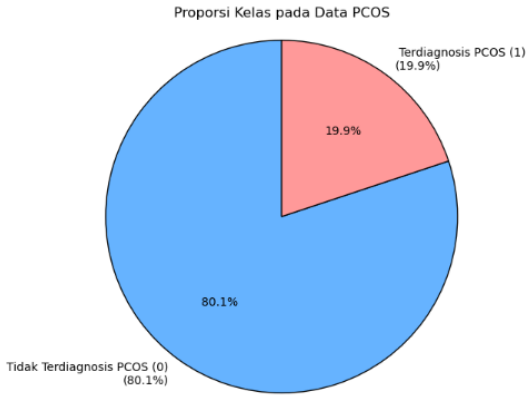
3. HASIL DAN PEMBAHASAN

3.1. Analisis Deskriptif

Gambar 1. Matriks korelasi

Berdasarakan matriks korelasi pada Gambar 1, fitur *menstrual irregularity* menunjukkan hubungan linier tingkat sedang terhadap diagnosis PCOS dengan nilai korelasi sebesar 0,47. Hal ini mengindikasikan adanya asosiasi yang cukup jelas antara ketidakteraturan menstruasi dan diagnosis PCOS. Selain itu, indeks massa tubuh (BMI) juga menunjukkan korelasi positif sebesar 0,38, yang mengindikasikan adanya kecenderungan hubungan linier positif antara BMI dan diagnosis PCOS.

Sementara itu, fitur *testosterone level* (ng/dL) dan antral follicle count memiliki nilai korelasi positif yang relatif kecil, masing-masing sebesar 0,20 dan 0,19, sehingga hubungannya tergolong lemah dan bersifat deskriptif. Adapun fitur usia (*age*) menunjukkan korelasi negatif yang sangat lemah sebesar −0,06, yang mengindikasikan hampir tidak adanya hubungan linier. Oleh karena itu, interpretasi korelasi dalam penelitian ini tidak dimaksudkan untuk menyimpulkan hubungan sebab–akibat, melainkan sebagai analisis eksploratif awal terhadap karakteristik data.



Gambar 2. Proporsi kelas data PCOS

E-ISSN: 2621-086X
https://doi.org/10.13057/ijas.v8i2.109931

Berdasarkan Gambar 2 dapat dilihat bahwa terjadi ketidakseimbangan distribusi kelas. Kelas 0 (Tidak Terdiagnosis PCOS) jauh lebih banyak sebesar 80,1% dibandingkan dengan kelas 1 (Terdiagnosis PCOS) sebesar 19,9%. Hal ini menunjukkan bahwa dari 1000 observasi yang ada mayoritas pasien tidak terdiagnosis PCOS.

3.2. Pre-Processing Data

Sebelum dilakukan pemodelan, tahap pre-processing data dilakukan untuk memastikan kualitas dan kesiapan data yang akan digunakan. Tahapan awal meliputi pemeriksaan data hilang (*missing values*) dan data duplikat. Hasil pemeriksaan menunjukkan bahwa seluruh variabel dalam dataset tidak mengandung data hilang maupun data duplikat, sehingga tidak diperlukan tindakan imputasi maupun penghapusan observasi.

Selain itu, dilakukan pemeriksaan pencilan (*outlier*) secara deskriptif pada variabel numerik. Namun, karena algoritma *random forest* dan *gradient boosting* relatif *robust* terhadap pencilan serta tidak ditemukan nilai ekstrem yang berpotensi memengaruhi hasil pemodelan, maka data pencilan tidak dihapus pada penelitian ini. Seluruh variabel dalam dataset telah tersedia dalam bentuk numerik, sehingga tidak diperlukan proses *encoding* tambahan sebelum data diproses lebih lanjut.

Dataset kemudian dibagi menunjukkan menjadi data pelatihan (*training set*) dan data pengujian (*testing set*) dengan rasio 80:20. Pembagian ini dipilih untuk memberikan proporsi data pelatihan yang cukup dalam proses pembelajaran model, sekaligus mempertahankan data pengujian yang representatif untuk evaluasi performa akhir.

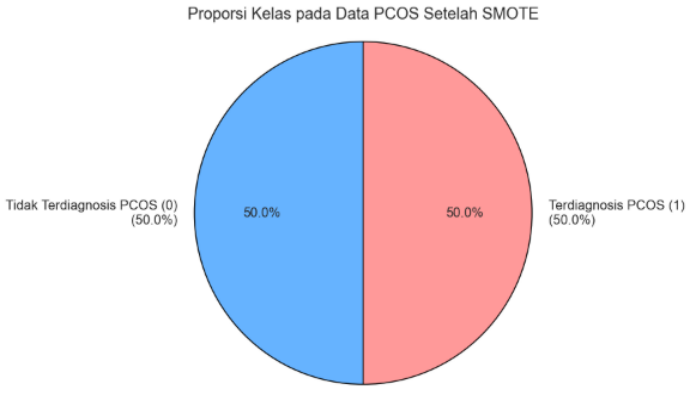
3.3. Pemilihan Fitur

Sebelum dilakukan pemodelan, maka perlu dilakukan pemilihan fitur untuk memastikan bahwa hanya fitur-fitur yang relevan dan bebas dari risiko data leakage (kebocoran data) yang akan digunakan dalam pelatihan model. Pada proses ini, fitur *menstrual_irregularity* tidak diikutsertakan dalam pemodelan karena kemungkinan fitur tersebut berkaitan langsung dengan kriteria diagnosis PCOS itu sendiri.

Sebelumnya, fitur *menstrual_irregularity* telah dimasukkan dalam pemodelan dan menghasilkan performa model yang sangat tinggi, nilai akurasi, dan nilai AUC mendekati atau melebih 90%. Akan tetapi, performa tersebut terindikasi tidak wajar karena model cenderung “menghafal” label dari fitur tersebut.

3.4. *Balancing Data* Menggunakan SMOTE

Berdasarkan analisis deskriptif sebelumnya, diketahui bahwa distribusi kelas pada data tidak seimbang, maka akan dilakukan penanganan dengan menggunakan SMOTE. SMOTE diterapkan setelah dilakukan *splitting data* dengan proporsi 80:20. Hasil SMOTE pada Gambar 3 menghasilkan distribusi data yang seimbang untuk kedua kelas sehingga model dapat belajar dari representasi kelas minoritas secara lebih optimal.



Gambar 3. Proporsi kelas data PCOS setelah SMOTE

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

3.5. Parameter Tuning dengan Bayesian Optimization

Parameter tuning dilakukan menggunakan *Bayesian optimization* setelah tahapan pre-processing dilakukan. Optimasi dilakukan secara terpisah untuk kedua model, yakni *random forest* dan *gradient boosting*. Melalui proses tuning parameter ini, diperoleh kombinasi parameter terbaik yang dapat meningkatkan performa masing-masing model.

Pada tahap ini, terlebih dahulu ditentukan ruang pencarian (*search space*) untuk setiap hiperparameter utama. Untuk *random forest*, ruang pencarian meliputi *n_estimators* {100, 300, 400, 500, 1000}, *max_depth* {5, 10, 20, 30, 40, 50, 100} (dengan nilai 100 merepresentasikan kedalaman tak terbatas), *min_samples_split* {2, 5, 10, 20, 30, 40}, *min_samples_leaf* {1, 2, 5, 10}, *max_features* dalam rentang kontinu 0,1–1,0, serta parameter bootstrap {0, 1}.

Sementara itu, ruang pencarian hiperparameter pada *gradient boosting* meliputi *learning rate* 0,001–0,5, *n_estimators* {10, 50, 100, 200, 500, 1000}, *max_depth* {1, 2, 3, 5, 7, 10, 15}, *min_samples_split* {2, 5, 10, 20, 50, 100}, *min_samples_leaf* {1, 2, 5, 10, 20, 50}, serta parameter *subsample*, *max_features*, dan *ccp_alpha* yang dieksplorasi dalam rentang kontinu masing-masing 0,1–1,0, 0,01–1,0, dan 0–0,5.

Proses *Bayesian optimization* diawali dengan sejumlah titik awal (initial points) untuk membangun model probabilistik awal, kemudian dilanjutkan dengan proses iterasi optimasi hingga batas jumlah iterasi yang telah ditentukan. Pada setiap iterasi, kombinasi hiperparameter dievaluasi menggunakan *k-fold cross-validation* dengan nilai k = 5 pada data pelatihan. Metrik AUC digunakan sebagai fungsi objektif yang dimaksimalkan dalam proses optimasi.

Kombinasi hiperparameter terbaik yang diperoleh dari proses ini selanjutnya digunakan untuk membangun model final. Nilai hiperparameter optimal untuk masing-masing model ditampilkan pada Tabel 3 dan Tabel 4.

Tabel 3. Parameter optimal untuk *random forest*

Parameter	Nilai Parameter Optimal
N estimators	300
Max depth	50
Min sample split	30
Min_samples leaf	5
Max features	0,9372

Tabel 4. Parameter optimal untuk *gradient boosting*

Parameter	Nilai Parameter Optimal
Learning rate	0,0922
N estimators	200
Max depth	10
Min sample split	2
Min_samples leaf	2
Subsample	0,9494
Max features	0,6256
Ccp alpha	0,0143

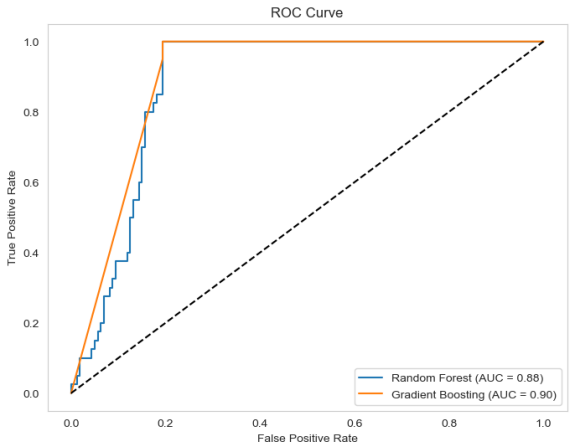
3.6. Perbandingan Performa Model Random Forest dan Gradient Boosting

Berdasarkan hasil evaluasi untuk kedua model yang disajikan pada Tabel 5, diketahui bahwa *gradient boosting* dengan SMOTE dan tuning menggunakan *Bayesian optimization* menghasilkan performa yang lebih baik dilihat dari keseluruhan metrik dibandingkan dengan model *random forest*.

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

Tabel 5. Perbandingan performa *random forest* dan *gradient boosting*

Model	Metrik Evaluasi	Nilai
Random Forest	Akurasi	0,82
	Presisi	0,53
	Recall	0,85
	F1-Score	0,65
	AUC	0,88
Gradient Boosting	Akurasi	0,83
	Presisi	0,55
	Recall	0,95
	F1-Score	0,70
	AUC	0,89



Gambar 4. Plot kurva ROC

Berdasarkan plot kurva ROC pada Gambar 4, kedua model menunjukkan performa klasifikasi yang cukup baik dalam membedakan kelas PCOS dan non-PCOS. Model *gradient boosting* memiliki nilai *recall* yang lebih tinggi, yang penting dalam konteks medis karena dapat mengurangi risiko kesalahan mengklasifikasikan pasien PCOS sebagai non-PCOS (*false negative*). Namun demikian, peningkatan *recall* juga perlu dipertimbangkan bersama presisi, karena peningkatan *false positive* dapat berdampak pada efisiensi proses *screaning*. Oleh karena itu, keseimbangan antara *recall* dan presisi menjadi aspek penting dalam penerapan model sebagai sistem pendukung keputusan klinis.

4. SIMPULAN

Penelitian ini dilakukan untuk mengevaluasi performa algoritma *random forest* dan *gradient boosting* dalam diagnosis PCOS pada data yang tidak seimbang. Sebagai pembanding, model awal (*baseline*) dibangun tanpa penyeimbangan data dan tanpa optimasi hiperparameter. Hasil penelitian menunjukkan bahwa penerapan SMOTE dan *Bayesian optimization* mampu meningkatkan performa model dibandingkan pendekatan baseline tersebut, khususnya pada metrik AUC dan *recall*.

Berdasarkan penelitian yang telah dilakukan, diperoleh hasil bahwa algoritma *gradient boosting* dengan penyeimbangan data menggunakan SMOTE dan optimasi hiperparameter melalui *Bayesian optimization* menghasilkan performa terbaik dengan nilai AUC sebesar 0,8983 dan recall sebesar 0,95. Hasil ini menunjukkan bahwa pendekatan yang digunakan mampu memberikan performa yang konsisten dalam mendeteksi kasus PCOS, khususnya pada kondisi data yang tidak seimbang dan dievaluasi menggunakan berbagai metrik performa.

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X

Untuk penelitian selanjutnya, terdapat beberapa hal yang dapat dikembangkan. Pertama, model yang sudah dibangun sebaiknya diuju pada dataset lain yang berasal dari institusi medis berbeda sehingga dapat diketahui seberapa baik model dapat bekerja pada data yang lebih beragam. Kedua, agar prediksi yang dihasilkan dapat lebih mudah dipahami, sebaiknya digunakan metode interpretasi, seperti SHAP atau LIME. Selain itu, dapat dicobakan kombinasi dari beberapa model sekaligus (*ensembling*) atau menambahkan fitur-fitur klinis lain. Terakhir, sebelum model ini diterapkan secara langsung di lingkungan klinis, perlu dilakukan studi terkait bagaimana sistem ini diterapkan agar tidak hanya menghasilkan hasil yang akurat secara teknis, tetapi juga efisien dan benar-benar dapat membantu di lapangan.

DAFTAR PUSTAKA

[1] J. P. Christ and M. I. Cedars, "Current guidelines for diagnosing PCOS," *diagnostics,* vol. 13, no. 6, pp. 1-11, 2023, doi: https://doi.org/10.3390/diagnostics13061113.

[2] S. B. Nadkarni, G. S. Vijay, and R. C. Kamath, "Comparative study of random forest and gradient boosting algorithms to predict airfoil self-noise," *Engineering Proceedings*, vol. 59, no. 24, 2023, doi: https://doi.org/10.3390/engproc2023059024.

[3] Y. F. Zamzam, T. H. Saragih, R. Herteno, Muliadi, D. T. Nugrahadi, and P.-H. Huynh, "Comparison of CatBoost and random forest methods for lung cancer classification using hyperparameter tuning bayesian optimization-based," *j. electron. electromedical. eng. med. inform*, vol. 6, no. 2, pp. 125–136, 2024, doi: https://doi.org/10.35882/jeeemi.v6i2.382.

[4] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan metode SMOTE random forest dan SMOTE XGBoost untuk klasifikasi tingkat penyakit hepatitis C pada imbalance class data," *Jurnal Gaussian*, vol. 9, no. 3, pp. 227–236, 2020, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/article/view/28915.

[5] T. Fulazzaky, A. Saefuddin, and A. M. Soleh, "Evaluating ensemble learning techniques for class imbalance in machine learning: A comparative analysis of balanced random forest, SMOTE-RF, SMOTEBoost, and RUSBoost," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 969–980, 2024, doi: https://doi.org/10.15294/sji.v11i4.15937.

[6] M. Dapas and A. Dunaif, "Deconstructing a syndrome: genomic insights into pcos causal mechanisms and classification," *Endocrine Reviews.*, vol. 43, no. 6, pp. 927–965, 2022, doi: https://doi.org/10.1210/endrev/bnac001.

[7] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272–278, 2012, [Online]. Available: https://www.ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf.

[8] J. K. Jaiswal and R. Samikannu, "Application of random forest algorithm on feature subset selection and classification and regression," in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, Hyderabad, India, pp. 65–70, 2016, doi: https://doi.org/10.1109/WCCCT.2016.25.

[9] Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, L. F. A. Bernardo, and H. M. Hussein, "Prediction of ecofriendly concrete compressive strength using Gradient Boosting Regression Tree combined with GridSearchCV hyperparameter-optimization techniques," *Materials*, vol. 15, no. 21, p. 7432, 2022, doi: https://doi.org/10.3390/ma15217432.

[10] M. Jun, "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area," *International Journal of Geographical Information Science.*, vol. 35, no. 11, pp. 1533–1552, 2021, doi: https://doi.org/10.1080/13658816.2021.1887490.

[11] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk mengatasi imbalance class dalam klasifikasi kepribadian MBTI menggunakan naive bayes classifier," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 5, pp. 1033–1042, 2024, doi: https://doi.org/10.25126/jtiik.2024117989.

[12] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *European Conference on Computer Vision (ECCV 2016)*, Cham, Switzerland, pp. 467–482, 2016, doi: https://doi.org/10.1007/978-3-319-46478-7_29.

[13] N. P. Y. T. Wijayanti, E. N Kencana, dan I. W. Sumarjaya, "SMOTE: potensi dan kekurangannya pada survei," *e-Jurnal Matematika*, vol. 10, no. 4, pp. 235–240, Nov. 2021, doi: https://doi.org/10.24843/MTK.2021.v10.i04.p348.

[14] G. Husain et al., "SMOTE vs. SMOTEENN: a study on the performance of resampling algorithms for addressing class imbalance in regression models," *Algorithms*, vol. 18, no. 1, pp. 37, 2025, doi: https://doi.org/10.3390/a18010037.

[15] M. Diessner, K. J. Wilson, dan R. D. Whalley, "On the development of a practical Bayesian optimization algorithm for expensive experiments and simulations with changing environmental conditions," *Data-Centric Eng.*, vol. 5, pp. e45, 2024, doi: https://doi.org/10.1017/dce.2024.40.

https://doi.org/10.13057/ijas.v8i2.109931
E-ISSN: 2621-086X