

BAB I PENDAHULUAN

1.1 Latar Belakang

Kesehatan reproduksi merupakan pilar fundamental bagi kesejahteraan wanita dan menjadi salah satu indikator kunci Kesehatan masyarakat secara global. Salah satu gangguan yang paling umum dan berdampak signifikan Adalah Sindrom Ovarium Polikistik atau *Polycystic Ovary Syndrome* (PCOS). PCOS didefinisikan sebagai *endokrinopati* (gangguan hormon) heterogen yang paling umum, dengan prevalensi global yang diperkirakan mempengaruhi 8% hingga 13% Wanita pada usia reproduktif [1], yang berarti sekitar 1 dari 10 wanita hidup dengan kondisi ini. Dampak PCOS tidak terbatas pada aspek reproduksi semata seperti *anovulasi kronis* yang menjadi penyebab utama infertilitas, tetapi juga meluas secara sistemik ke ranah metabolik, termasuk resistensi insulin yang meningkatkan risiko obesitas, diabetes melitus tipe 2, dan penyakit kardiovaskular [2]. Selain itu, *Sindrom ini* juga melibatkan beban psikologis yang ditimbulkan, seperti peningkatan risiko depresi, kecemasan, serta harga diri yang rendah menjadi konsekuensi serius yang sering menyertai sindrom ini [3].

Permasalahan Utama terletak pada tantangan dalam proses diagnosis PCOS yang membutuhkan waktu yang lama. Diagnosis konvensional yang mengacu pada Kriteria Rotterdam bersifat kompleks, subjektif, dan memakan waktu, karena memerlukan pemenuhan setidaknya dua dari tiga kriteria yaitu: (i) *hiperandrogisme* klinis, seperti hirsutisme, atau *hiperandrogenisme* biokimia yang diukur melalui tingkat testosteron total atau bebas; (ii) disfungsi ovulasi, termasuk siklus menstruasi yang tidak teratur dan *anovulasi*; dan (iii) ovarium polikistik yang diamati melalui ultrasonografi atau kadar hormon *anti-müllerian* (AMH) yang meningkat [4], [5]. Proses diagnosis multi-tahap ini sering kali menyebabkan *keterlambatan yang signifikan* terhadap Keputusan diagnosis, karena metode ini dinilai sangat kompleks serta memakan waktu dan mahal. Hal ini dapat mempengaruhi pasien telat mendapatkan penanganan esensial dan memperburuk prognosis Kesehatan jangka Panjang pasien.

Untuk menanggapi tantangan diagnostik ini, *Machine Learning* telah turut serta menawarkan solusi potensial selama beberapa dekade terakhir. Teknologi *Machine Learning* mulai kerap dikembangkan dan diterapkan dalam berbagai sektor, termasuk di sektor kesehatan. Model *Machine Learning* memiliki kapasitas untuk menganalisis pola

kompleks dari data klinis dan biokimia multi-dimensi untuk membantu proses skrining secara objektif dan cepat [6]. Namun, penerapan *Machine Learning* dalam konteks sebagai alat bantu prediksi dini PCOS dihadapkan pada permasalahan keterbatasan data. Dataset klinis PCOS yang tersedia di repositori terbuka yang digunakan dalam penelitian ini, hanya memiliki jumlah entri yang sangat terbatas. Dataset ini juga memiliki struktur kelas yang tidak seimbang atau *imbalanced* data, di mana jumlah penderita dan non-penderita memiliki ketimpangan yang tinggi. Keterbatasan data dan data tidak seimbang ini adalah tantangan teknis serius yang membuat proses pelatihan model menjadi sulit, rentan terhadap *overfitting*, dan berisiko menghasilkan kesimpulan yang bias atau tidak dapat diandalkan.

Ditinjau dari sebuah *Systematic Literature Review* (SLR) yang dilakukan oleh Arora, et. al [7] memberikan pernyataan bahwa terdapat permasalahan, yaitu inkonsistensi perbandingan, banyak studi tidak membandingkan algoritma secara adil atau hanya terpaku pada metrik akurasi, sehingga mengabaikan sensitivitas klinis (*Recall*). Sehingga dapat disimpulkan bahwa terdapat celah yang paling fundamental yaitu bagaimana menentukan algoritma klasifikasi terbaik untuk data klinis yang ada saat ini masih belum terjawab secara tuntas.

Oleh sebab itu, penelitian ini dirancang untuk fokus menyelesaikan masalah fundamental tersebut dengan pendekatan yang disesuaikan dengan tantangan data yang terbatas. Masalah spesifik yang akan diselesaikan di penelitian ini adalah bagaimana menentukan model komputasi dengan menggunakan pendekatan Random Forest sebagai metode yang mewakili *bagging ensemble*, XGBoost sebagai metode yang mewakili *boosting ensemble*, dan Support Vector Machine sebagai metode yang mewakili *margin-based classifier* dan menganalisis pendekatan yang paling optimal dan memiliki metrik yang seimbang ketika dihadapkan pada keterbatasan dataset PCOS yang sangat kecil dan tidak seimbang. Sehingga penelitian ini diharapkan dapat memberikan solusi untuk menerapkan kelas SMOTE di dalam protokol evaluasi Stratified K-fold Cross-validation. Penggunaan validasi silang berstrata ini dipilih secara spesifik untuk mengurangi risiko bias variansi dan *overfitting* yang mungkin terjadi karena ukuran dataset yang relatif kecil, sehingga memastikan setiap sampel data dapat dinilai secara adil. Penelitian ini juga menerapkan Analisis signifikansi fitur dengan *permutation sfeature importance* untuk mengidentifikasi prediktor klinis yang paling berpengaruh. Analisis fitur ini

disertakan secara spesifik untuk menjawab kebutuhan transparansi klinis, memastikan bahwa model tidak hanya akurat secara statistik, tetapi juga logis secara medis. Melalui analisis komparatif ini, penelitian ini diharapkan dapat membuktikan model mana yang paling dapat diandalkan, memberikan kontribusi sebagai alat bantu untuk mempercepat proses skrining PCOS, mengurangi beban subjektivitas diagnosis konvensional, serta mendukung pengembangan sistem pendukung keputusan klinis yang lebih akurat, sehingga berkontribusi pada pengembangan sistem skrining berbasis machine learning.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka rumusan masalah yang dapat disimpulkan adalah sebagai berikut:

1. Bagaimana penerapan teknik *preprocessing* data, khususnya *Synthetic Minority Over-Sampling Technique (SMOTE)*, dapat secara efektif mengatasi tantangan komputasi yang disebabkan oleh karakteristik dataset PCOS yang tidak seimbang?
2. Berdasarkan hasil perbandingan kinerja tersebut, algoritma manakah yang dapat diidentifikasi sebagai model paling optimal dan andal untuk skrining dini PCOS, terutama dengan mempertimbangkan sensitivitas model dalam mendeteksi kasus positif?
3. Berdasarkan model berkinerja terbaik yang teridentifikasi, fitur klinis dan biokimia apa saja yang terdeteksi sebagai prediktor paling berpengaruh dalam klasifikasi PCOS?

1.3 Batasan Masalah

Supaya penelitian ini tetap terfokus, valid, dan sesuai dengan tujuan yang telah ditetapkan, maka peneliti menetapkan Batasan-batasan masalah sebagai berikut:

1. Penelitian ini secara eksklusif menggunakan satu sumber dataset publik yang diperoleh dari repositori Kaggle. Model yang dikembangkan akan dilatih dan diuji hanya menggunakan dataset ini.
2. Variabel yang diteliti terbatas pada fitur-fitur yang telah tersedia di dalam dataset tersebut. Variabel-variabel lain seperti data genetik, data psikologis, atau data citra yang mungkin berpengaruh pada PCOS dengan sengaja diabaikan agar tetap fokus dan spesifik.
3. Studi komparatif ini difokuskan pada tiga algoritma machine learning yang telah

ditentukan, yaitu Random Forest sebagai bagging ensemble, XGBoost sebagai *boosting ensemble*, dan Support Vector Machine sebagai margin based. Adapun penanganan ketidakseimbangan data secara spesifik menerapkan metode *Synthetic Minority Over Sampling Technique* (SMOTE) sebagai teknik oversampling yang diuji efektivitasnya.

4. Untuk menjaga objektivitas perbandingan antar-algoritma, konfigurasi *hyperparameter* untuk setiap model dikontrol menggunakan parameter standar atau penyesuaian dasar. Hal ini bertujuan untuk menetapkan tolak ukur kinerja dasar algoritma.
5. Luaran penelitian ini mencakup dua hal, yaitu analisis komparatif kinerja model untuk mengidentifikasi algoritma yang paling optimal, dan implementasi model terbaik tersebut dalam bentuk prototipe antarmuka skrining berbasis web sebagai bukti konsep penerapan klinis.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, maka penelitian ini memiliki tujuan sebagai berikut:

1. Menerapkan dan menganalisis efektivitas teknik *preprocessing* data, termasuk teknik imputasi standar dan SMOTE, untuk menangani masalah nilai hilang dan struktur data yang tidak seimbang.
2. Menghitung dan menganalisis secara komparatif kinerja prediktif dengan algoritma Random Forest, XGBoost, dan SVM. Analisis komparatif ini didasarkan pada parameter metrik evaluasi standar, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *Area Under the Curve* (AUC-ROC), untuk menentukan model dengan akurasi dan reliabilitas tertinggi.
3. Mengidentifikasi dan menganalisis faktor-faktor klinis (fitur) yang paling berpengaruh dalam memprediksi PCOS menggunakan metode *Permutation Feature Importance* yang diterapkan secara seragam pada ketiga model untuk menghasilkan perbandingan yang adil dan konsisten.
4. Mengimplementasikan model algoritma terbaik yang teridentifikasi ke dalam prototipe sistem skrining PCOS berbasis web sebagai wujud nyata kontribusi penelitian yang dapat digunakan sebagai referensi untuk peneliti selanjutnya.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi yang signifikan, baik secara teoritis bagi pengembangan ilmu pengetahuan di bidang informatika medis maupun secara praktis bagi praktisi Kesehatan.

1.5.1 Manfaat Teoretis

Secara teoritis, penelitian ini akan memperkaya literatur dan Khazanah keilmuan dalam beberapa aspek:

1. Penelitian ini menunjukkan bagaimana kinerja Random Forest, XGBoost, dan Support Vector Machine ketika dihadapkan pada kondisi data yang terbatas dan tidak seimbang.
2. Penelitian ini memberikan kontribusi metodologis berupa penerapan protokol SMOTE yang ketat di dalam setiap lipatan *Stratified K-Fold Cross-Validation*, sehingga proses evaluasi model bebas dari bias kebocoran data dan menghasilkan estimasi kinerja yang lebih objektif pada dataset medis berukuran terbatas.
3. Pengembangan konsep *Feature Importance* Dengan menganalisis dan memvalidasi fitur klinis paling berpengaruh secara komputasional, penelitian ini memberikan kontribusi konsep berupa validasi silang (*cross-validation*) antara pengetahuan medis dan temuan berbasis data. Hal ini bertujuan untuk membantu memperkuat teori tentang faktor-faktor klinis mana yang paling prediktif untuk PCOS dari perspektif ilmu data.

1.5.2 Manfaat Praktis

Secara praktis, temuan dari penelitian ini dapat diterapkan di lapangan untuk memberikan dampak nyata, seperti:

1. Wawasan dari analisis *permutation feature importance* dapat memberikan masukan berbasis data bagi praktisi medis untuk mengidentifikasi fitur klinis mana yang paling prediktif, yang berpotensi membantu mengoptimalkan fokus dalam proses skrining awal.
2. Penelitian ini menghasilkan prototipe system skrining PCOS berbasis web yang dapat digunakan langsung sebagai referensi oleh pengembang perangkat lunak dalam membangun *Clinical Decision Support System* (CDSS) yang lebih komprehensif di masa mendatang.

1.6 Sistematika Penulisan

Penelitian ini menerapkan sistematika penulisan supaya penyusunan penulisan menjadi lebih teratur dan dapat dipahami dengan mudah, sistematika penulisan yang terdapat pada penelitian adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi pengantar mengenai penelitian yang dilakukan. Isi bab ini mencakup latar belakang masalah yang menjelaskan urgensi diagnosis dini PCOS dan potensi *machine learning* untuk mengisinya, Rumusan Masalah, Batasan Masalah yang mendefinisikan ruang lingkup data dan metode, tujuan penelitian secara rinci, Manfaat penelitian, dan diakhiri dengan sistematika penulisan ini.

BAB II LANDASAN TEORI

Bab ini menyajikan dua bagian utama, yaitu tinjauan Pustaka dan dasar teori. Tinjauan Pustaka akan memaparkan *state-of-the-art* atau penelitian terdahulu yang relevan terkait prediksi PCOS menggunakan *machine learning*. Dasar teori akan menjabarkan konsep-konsep kunci yang digunakan, meliputi teori akan menjabarkan konsep-konsep kunci yang digunakan, meliputi teori mengenai Sindrom Ovarium Polikistik (PCOS), prinsip kerja machine learning, serta arsitektur teoretis dari algoritma Random Forest, XGBoost, dan Support Vector Machine (SVM).

BAB III METODE PENELITIAN

Bab ini memaparkan secara rinci Langkah-langkah teknis penelitian. Di dalamnya terdapat penjelasan mengenai objek penelitian, alur kerja penelitian, serta metodologi rinci. Metodologi ini mencakup tahapan *data selection*, *preprocessing*, dan *evaluation*.

BAB IV HASIL DAN PEMBAHASAN

Bab ini merupakan inti dari penelitian yang menyajikan seluruh temuan. Bagian pertama berfokus pada hasil analisis komparatif model, yang disajikan melalui tabel evaluasi kinerja dari ketiga algoritma. Bagian ini juga akan menyajikan hasil analisis fitur terpenting. Seluruh temuan tersebut akan dibahas dan diinterpretasikan secara mendalam untuk menjawab mengapa hasil tersebut didapat.

BAB V PENUTUP

Bab terakhir ini berisi Kesimpulan yang dirangkum dari hasil pembahasan pada bab sebelumnya dan secara lugas menjawab poin-poin pada rumusan masalah. Selain itu, bab

ini juga akan memuat saran yang berisi rekomendasi konstruktif untuk pengembangan penelitian ini di masa mendatang, baik dari sisi akurasi model maupun fungsionalitas aplikasi.

