

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kanker payudara adalah kondisi yang terjadi ketika sel-sel abnormal pada jaringan payudara tumbuh dan berkembang secara tidak terkendali, yang dapat menyebar ke bagian tubuh lainnya dan mengancam kesehatan serta kehidupan penderitanya. Kanker payudara telah menjadi salah satu jenis kanker paling umum di dunia, dengan lebih dari 2,3 juta kasus baru dan sekitar 685.000 kematian pada tahun 2020. Penelitian terbaru dari WHO dan IARC memperkirakan bahwa angka ini akan terus meningkat hingga lebih dari 3 juta kasus baru per tahun pada 2040, terutama di negara-negara dengan ekonomi berkembang yang mengalami peningkatan signifikan dalam insiden tetapi keterbatasan dalam penanganan dini [1].

Kanker payudara merupakan penyakit yang semakin umum ditemui pada wanita di seluruh dunia, dengan insiden yang terus meningkat di berbagai negara dan kelompok usia [1]. Kanker payudara adalah suatu kondisi keganasan di mana sel-sel berkembang secara tidak terkendali. Penyakit ini dapat dimulai dari sel-sel lobulus, duktus, atau jaringan ikat di payudara, dan kemudian dapat menyebar melalui pembuluh darah dan sistem limfatik ke organ-organ lain dalam tubuh [2].

Data mining adalah proses untuk mengumpulkan informasi serta mengolah data yang diambil dari kumpulan data besar, seperti database, media sosial, dan sumber lainnya. Tujuan dari data mining adalah mengidentifikasi pola, tren, dan keunikan dalam data yang dapat dimanfaatkan sebagai alat pendukung dalam pengambilan keputusan bisnis. Proses data mining terdiri dari beberapa tahap, yaitu praproses data, eksplorasi data, pembentukan model, dan evaluasi [3].

Klasifikasi merupakan proses mengelompokkan data dengan karakteristik serupa ke dalam kategori atau kelas yang sudah ditetapkan sebelumnya. Umumnya, proses klasifikasi ini melibatkan pengelompokan data berdasarkan ciri khas yang sudah ditentukan pada setiap kelas [4]. Beberapa metode yang dapat diterapkan

untuk klasifikasi mencakup *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbors*, serta *Gradient Boosting*.

Salah satu algoritma yang sangat efektif dalam menyelesaikan masalah klasifikasi adalah *Support Vector Machine* (SVM). SVM dikenal mampu menangani data berdimensi tinggi dengan baik, yang menjadikannya unggul dalam memisahkan kelas-kelas data menggunakan *hyperplane* optimal. Akurasi SVM yang tinggi menjadikannya pilihan populer untuk berbagai aplikasi klasifikasi, terutama dalam bidang yang memerlukan tingkat ketepatan yang tinggi, seperti medis dan pengenalan pola [5].

Pemilihan fitur yang tepat serta penggunaan *classifier* yang sesuai sangat penting untuk meningkatkan akurasi dan efisiensi komputasi dalam prediksi. Salah satu algoritma yang digunakan dalam mengklasifikasikan data dalam jumlah besar adalah *Random Forest*, yang dikenal efektif untuk tugas klasifikasi [6]. Di beberapa kasus penelitian, algoritma *Random Forest* mempunyai nilai akurasi yang cukup tinggi dibanding dengan algoritma lainnya [7].

Sementara algoritma *K-Nearest Neighbors* (KNN) merupakan metode *supervised learning* yang mengklasifikasikan data baru berdasarkan kedekatannya dengan sejumlah tetangga terdekat, di mana jumlah tetangga ini ditentukan oleh nilai parameter *K*. Keunggulan utama KNN terletak pada kemampuannya memanfaatkan informasi dari tetangga-tetangga terdekat untuk menetapkan kategori yang paling sesuai, sehingga menjadikannya efektif dalam mengelompokkan objek baru berdasarkan karakteristik atau atribut yang dimilikinya [8].

Algoritma *gradient boosting* banyak digunakan karena memiliki skalabilitas dan efisiensi yang tinggi dalam menangani berbagai masalah baik pada regresi maupun klasifikasi [9]. *Gradient Boosting* merupakan metode *supervised learning* berbasis pohon keputusan, yang dimulai dengan menghasilkan pohon klasifikasi awal dan kemudian secara bertahap menambahkan pohon baru untuk meminimalkan fungsi kerugian. Keunggulan dari algoritma ini terletak pada kemampuannya dalam meningkatkan akurasi prediksi secara signifikan, dengan

hasil akurasi yang sering kali lebih dari 80% lebih baik dibandingkan model populer seperti *Random Forest* [10].

Salah satu cara efektif untuk meningkatkan kinerja algoritma adalah melalui *hyperparameter tuning*. *Hyperparameter tuning* merupakan proses untuk mengatur parameter-parameter yang secara langsung mempengaruhi performa model *machine learning*. Ada beberapa metode yang sering digunakan dalam proses ini, seperti *grid search*, *random search*, yang membantu menemukan kombinasi parameter optimal agar model bekerja lebih baik. Setiap metode memiliki pendekatan unik dalam eksplorasi nilai parameter untuk mencapai hasil yang optimal [11].

Namun, masih terdapat keterbatasan dalam penelitian sebelumnya yang hanya membandingkan performa algoritma klasifikasi tanpa mengintegrasikan proses *hyperparameter tuning* secara jelas dan menyeluruh [12]. Untuk mengatasi celah tersebut, penelitian ini secara khusus menerapkan *hyperparameter tuning* pada setiap algoritma yang dibandingkan guna memperoleh model yang optimal dan akurat. Pendekatan ini diharapkan dapat memberikan kontribusi yang lebih komprehensif dalam konteks klasifikasi kanker payudara [13].

Berdasarkan permasalahan dan pembahasan yang telah disampaikan, penelitian ini membandingkan algoritma *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbors* (KNN), dan *Gradient Boosting* dengan penerapan *hyperparameter tuning* untuk mengklasifikasikan individu yang telah terdiagnosis kanker payudara berdasarkan faktor-faktor yang mempengaruhi kondisi mereka. Studi ini berfokus pada menemukan algoritma terbaik untuk mengelompokkan pasien, sehingga dapat mendukung analisis lanjutan dalam perawatan dan pengelolaan kanker payudara [14].

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah di atas, rumusan masalah yang dapat disimpulkan yaitu :

1. Bagaimana performa algoritma *Support Vector Machine* (SVM), *Random*

Forest, *K-Nearest Neighbors* (KNN), dan *Gradient Boosting* dalam mengklasifikasikan kanker payudara berdasarkan faktor risiko yang tersedia?

2. Apakah *hyperparameter tuning* dapat meningkatkan akurasi prediksi dari masing-masing algoritma dalam klasifikasi kanker payudara?
3. Algoritma manakah yang paling optimal dalam klasifikasi kanker payudara dari segi akurasi dan efisiensi komputasi?

1.3 Batasan Masalah

Setelah mengidentifikasi permasalahan dalam penelitian ini, perlu dilakukan pembatasan agar pembahasan lebih terarah dan tujuan penelitian dapat tercapai dengan efektif. Berikut ini adalah beberapa batasan masalah yang diterapkan dalam penelitian ini :

1. Data yang digunakan terbatas pada data kanker payudara yang mengandung faktor risiko seperti usia, riwayat keluarga, faktor reproduksi, gaya hidup, dan hormon.
2. Penelitian hanya membandingkan empat algoritma *machine learning*, yaitu *SVM*, *Random Forest*, *KNN*, dan *Gradient Boosting*, yang disertai *hyperparameter tuning*.
3. Metode *hyperparameter tuning* yang digunakan terbatas pada teknik *grid search*, dan *random search*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan performa empat algoritma *machine learning*, yaitu *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbors* (KNN), dan *Gradient Boosting*, dalam klasifikasi kanker payudara. Penelitian ini akan mengoptimalkan *hyperparameter* pada setiap algoritma untuk meningkatkan performa model. Fokus utama adalah mengevaluasi kemampuan algoritma dalam memprediksi risiko kanker payudara berdasarkan faktor-faktor risiko yang telah ditentukan, seperti usia, riwayat keluarga, gaya hidup, dan aspek reproduksi.

Melalui analisis mendalam, penelitian ini bertujuan untuk mengidentifikasi model yang paling akurat dan efisien dalam mendukung diagnosis kanker payudara secara dini. Hasil dari studi ini diharapkan dapat memberikan wawasan yang signifikan mengenai algoritma terbaik yang dapat diterapkan di bidang kesehatan, khususnya dalam memanfaatkan teknologi *machine learning* untuk meningkatkan akurasi prediksi dan efisiensi komputasi dalam klasifikasi penyakit.

1.5 Manfaat Penelitian

1. Pengembangan Sistem Diagnostik Medis

Hasil penelitian ini dapat digunakan oleh profesional medis dalam memilih algoritma yang paling efektif untuk analisis kanker payudara. Dengan akurasi yang lebih tinggi, model ini diharapkan dapat meningkatkan deteksi dini kanker payudara, sehingga memungkinkan pengobatan yang lebih cepat dan tepat.

2. Penerapan Hyperparameter Tuning dalam Model Kesehatan

Penelitian ini dapat memberikan panduan tentang pentingnya *hyperparameter tuning* dalam meningkatkan kinerja model *machine learning*. Ini dapat membantu dalam mengoptimalkan algoritma yang digunakan untuk pengolahan data medis, sehingga model yang dihasilkan lebih akurat dan efisien.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang masalah, rumusan masalah, tujuan, dan manfaat penelitian. Selain itu, juga dijelaskan pembatasan masalah untuk memperjelas ruang lingkup penelitian. Bab ini juga akan menyajikan sistematika penulisan untuk memberikan gambaran umum tentang alur penelitian yang dilakukan.

BAB II TINJAUAN PUSTAKA

Pada bab ini, akan dibahas teori-teori yang mendasari penelitian, termasuk konsep dasar kanker payudara dan algoritma *machine learning* yang digunakan. Penelitian sebelumnya yang relevan juga akan disajikan untuk memperkuat dasar teori yang digunakan dalam penelitian ini.

BAB III METODE PENELITIAN

Bab ini menjelaskan desain penelitian, langkah-langkah yang dilakukan untuk pengumpulan data, dan teknik analisis yang digunakan. Fokus utama adalah implementasi algoritma SVM, Random Forest, KNN, dan Gradient Boosting untuk klasifikasi kanker payudara serta penerapan *hyperparameter tuning*.

BAB IV HASIL DAN PEMBAHASAN

Di bab ini, hasil eksperimen akan dipresentasikan dan dibandingkan antar algoritma. Pembahasan mencakup analisis kinerja masing-masing model dan pengaruh dari *hyperparameter tuning* terhadap akurasi prediksi, serta perbandingan dengan penelitian sebelumnya.

BAB V KESIMPULAN

Pada Bab terakhir ini menyimpulkan hasil penelitian, memberikan rekomendasi tentang algoritma yang paling efektif, serta saran untuk penelitian lanjutan dalam pengembangan sistem deteksi kanker payudara berbasis *machine learning*.