BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes Mellitus (DM) merupakan salah satu penyakit kronis dengan tingkat prevalensi yang terus meningkat secara drastis di berbagai belahan dunia. Penyakit metabolisme ini ditandai dengan ketidakmampuan tubuh dalam mengelola kadar gula darah yang, jika tidak ditangani sejak dini, dapat memicu berbagai komplikasi fatal seperti gagal ginjal, penyakit kardiovaskular, hingga kerusakan sistem saraf [1]. Deteksi dini dan pencegahan menjadi kunci utama dalam menekan angka risiko komplikasi tersebut. Namun, proses diagnosis klinis secara konvensional seringkali membutuhkan serangkaian tes laboratorium yang memakan waktu dan biaya, serta sangat bergantung pada ketersediaan tenaga medis ahli di fasilitas kesehatan [2].

Dalam menjembatani kendala tersebut, penerapan teknologi informasi melalui sistem cerdas berbasis Machine Learning telah banyak dikembangkan di bidang informatika medis. Teknologi ini dirancang untuk membaca pola data kesehatan pasien masa lalu dan mengklasifikasikannya menjadi prediksi risiko di masa depan. Sistem semacam ini diposisikan sebagai Clinical Decision Support System (CDSS) atau sistem pendukung keputusan klinis, yang bertujuan untuk membantu dokter maupun masyarakat umum dalam melakukan deteksi atau skrining awal secara jauh lebih cepat, efisien, dan objektif [3]. Dari berbagai algoritma Machine Learning yang ada, Random Forest dan Support Vector Machine (SVM) merupakan dua metode yang terbukti sangat tangguh dalam menangani data klasifikasi medis. Random Forest unggul berkat kemampuannya memproses banyak variabel secara bersamaan melalui konsep ensemble learning, sementara SVM sangat ahli dalam menemukan garis pemisah (hyperplane) yang membedakan pasien sehat dan sakit [4].

Berbagai algoritma klasifikasi telah dikembangkan untuk mendukung prediksi penyakit Diabetes Mellitus, seperti Logistic Regression, Decision Tree, K-Nearest Neighbor, Naïve Bayes, Random Forest, dan Support Vector Machine. Masing-masing algoritma memiliki karakteristik yang berbeda dalam mengolah data dan menghasilkan prediksi. Oleh karena itu, pemilihan algoritma harus disesuaikan dengan karakteristik dataset serta tujuan penelitian. Random Forest dipilih karena merupakan algoritma berbasis ensemble yang mampu mengurangi risiko overfitting melalui pembentukan banyak pohon keputusan (decision tree) [5]. Selain itu, Random Forest relatif stabil terhadap data yang memiliki banyak variabel, tahan terhadap noise, serta mampu memberikan informasi mengenai tingkat kepentingan setiap fitur (feature importance). Karakteristik tersebut menjadikan Random Forest sesuai untuk mengolah data kesehatan yang umumnya memiliki hubungan antarfitur yang kompleks [6, 7].

Sementara itu, Support Vector Machine (SVM) dipilih karena memiliki kemampuan yang baik dalam membangun batas pemisah (hyperplane) yang optimal antara dua kelas data. Dengan menggunakan kernel Radial Basis Function (RBF), SVM mampu memodelkan hubungan nonlinier yang sering ditemukan pada data klinis, sehingga diharapkan dapat meningkatkan kemampuan model dalam membedakan individu yang berisiko Diabetes Mellitus dan yang tidak berisiko. Berdasarkan karakteristik tersebut, penelitian ini membandingkan performa Random Forest dan Support Vector Machine menggunakan dataset yang sama sehingga dapat diketahui algoritma yang memberikan kinerja terbaik dalam memprediksi Diabetes Mellitus berdasarkan indikator akurasi, presisi, recall, F1-score, dan Area Under Curve (AUC) [2].

Dalam perkembangannya, tingkat akurasi dan keandalan dari sistem prediktif klinis sangat bergantung pada kualitas serta relevansi dataset yang digunakan untuk melatih algoritma. Banyak penelitian terdahulu di ranah ini masih terpaku pada penggunaan dataset usang yang memiliki dimensi observasi sangat terbatas, sehingga memunculkan masalah kebaruan (novelty) dan bias

komputasi. Oleh karena itu, untuk memastikan model yang dibangun memiliki validitas yang tinggi, penelitian ini menggunakan dataset terbaru bertajuk Behavioral Risk Factor Surveillance System (BRFSS) Data for Diabetes Prediction yang menawarkan representasi rekam medis secara jauh lebih modern, akurat, dan relevan dengan kondisi kesehatan demografi masa kini [5].

Dataset Behavioral Risk Factor Surveillance System (BRFSS) ini memiliki keunggulan komparatif yang signifikan karena tidak hanya bertumpu pada indikator klinis konvensional, tetapi juga mengintegrasikan secara mendalam berbagai metrik gaya hidup pasien tingkat individu. Parameter objektif seperti riwayat aktivitas fisik, metrik obesitas yang diukur melalui Indeks Massa Tubuh (BMI), tingkat glukosa, riwayat kolesterol, hingga rekam jejak konsumsi rokok tersedia secara komprehensif. Kelengkapan fitur rekam medis riil inilah yang memungkinkan algoritma untuk menangkap pola risiko diabetes secara lebih holistik dan menghindari risiko kebocoran data (data leakage) yang kerap terjadi akibat manipulasi data sintesis pada dataset berskala kecil [6]. Kompleksitas fitur medis dan parameter gaya hidup dalam dataset ini memberikan ruang komputasi yang sangat ideal untuk membandingkan kecerdasan kedua algoritma yang diajukan. Melalui arsitektur dataset ini, kapabilitas Random Forest dapat dievaluasi dalam memberikan wawasan mengenai tingkat kepentingan fitur (feature importance) untuk memetakan indikator gaya hidup mana yang paling memicu sindrom metabolik. Di sisi lain, ketangguhan algoritma SVM menggunakan fungsi kernel Radial Basis Function (RBF) akan diuji secara maksimal untuk memetakan dan memisahkan sebaran riwayat data pasien yang memiliki korelasi non-linear tinggi antar-parameternya [7].

Lebih dari sekadar membandingkan akurasi, evaluasi kinerja algoritma di ranah medis menuntut fokus analitis yang berbeda. Kesalahan sistem dalam memprediksi seorang pasien diabetes menjadi pasien sehat (False Negative) memiliki konsekuensi dan risiko fatalitas yang jauh lebih tinggi dibandingkan

jika sistem memprediksi pasien sehat menjadi diabetes (False Positive). Oleh karena itu, penelitian ini tidak hanya mencari algoritma dengan tingkat kebenaran umum tertinggi, melainkan memfokuskan tolok ukur pada metrik Recall kelas positif, guna memastikan sistem memiliki tingkat kepekaan (sensitivitas) yang maksimal dalam mengenali individu yang benar-benar berisiko tinggi. Pada akhirnya, sebuah model algoritma prediktif yang akurat tidak akan memberikan manfaat nyata secara praktis jika implementasinya hanya berhenti pada script kode di lingkungan pengujian komputasi (Jupyter Notebook). Banyak penelitian terdahulu yang luput dalam menerjemahkan model matematis tersebut menjadi perangkat lunak yang bisa diakses oleh pengguna akhir. Menjawab tantangan tersebut, penelitian ini tidak hanya berfokus pada optimasi algoritma di belakang layar, tetapi juga merancang purwarupa (prototype) aplikasi antarmuka berbasis web sederhana. Melalui antarmuka ini, hasil pelatihan model diintegrasikan secara utuh sehingga dapat dioperasikan langsung sebagai alat bantu skrining kesehatan mandiri yang aplikatif, interaktif, dan berlandaskan pada metodologi data yang solid [8].

1.2 Rumusan Masalah

Berdasarkan permasalahan yang dipaparkan pada latar belakang maka dirumuskan beberapa masalah yang dihadapi yaitu :

1. Bagaimana algoritma machine learning Random Forest dan Support Vector Machine dapat digunakan untuk memprediksi penyakit diabetes mellitus ?
2. Bagaimana perbandingan performa algoritma Random Forest dan Support Vector Machine dalam memprediksi penyakit diabetes mellitus ?
3. Berdasarkan hasil evaluasi metrik akurasi, presisi, *recall*, dan AUC *score*, algoritma manakah yang memberikan performa paling optimal dan stabil dalam memprediksi penyakit Diabetes Mellitus pada dataset pasien wanita dengan volume data yang telah dikembangkan?

1.3 Batasan Masalah

Adapun batasan masalah yang digunakan dalam penelitian ini Adalah

1. Dataset yang digunakan merupakan data sekunder yang diperoleh melalui platform Kaggle dengan nama Health & Lifestyle Data for Diabetes Prediction: A Preprocessed Dataset for EDA, Regression, and Predictive Modeling. Dataset ini merupakan dataset sintetis (synthetic dataset) yang dikembangkan berdasarkan distribusi statistik dan literatur kesehatan masyarakat sehingga menyerupai karakteristik data klinis nyata tanpa menggunakan identitas pasien asli. Dataset terdiri atas sekitar 98.000 record dengan 31 atribut yang meliputi karakteristik demografi, gaya hidup, riwayat kesehatan, pengukuran klinis, serta variabel target diagnosis diabetes. Dataset tersedia dalam format CSV dan digunakan sebagai sumber data sekunder dalam penelitian ini. h**ttps://www.kaggle.com/datasets/alamshihab075/health-and-lifestyle-data-for-diabetes-prediction** berikut Adalah link datasetnya.
2. Batasan Atribut (Fitur): Indikator yang dianalisis dibatasi pada 10 fitur klinis dan gaya hidup utama yang paling relevan dengan status diagnosis diabetes, yaitu: Hipertensi (hypertension_history), Kolesterol Total (cholesterol_total), Indeks Massa Tubuh (bmi), Status Merokok (smoking_status), Riwayat Penyakit Kardiovaskular (cardiovascular_history), Aktivitas Fisik (physical_activity_minutes_per_week), Kadar HbA1c (hba1c), Kadar Glukosa Puasa (glucose_fasting), Jenis Kelamin (gender), dan Kategori Usia (Age).
3. Algoritma Klasifikasi: Pemodelan Machine Learning hanya dibatasi pada komparasi dua algoritma, yaitu Random Forest Classifier dan Support Vector Machine (SVM). Khusus untuk algoritma SVM, fungsi kernel yang digunakan dibatasi pada Radial Basis Function (RBF) untuk menangani pola data klinis yang bersifat non-linear.
4. Metode Penyeimbangan Data: Untuk mengatasi ketidakseimbangan proporsi kelas target tanpa memicu kebocoran data (data leakage), teknik penyeimbangan kelas dilakukan menggunakan metode Undersampling secara eksklusif pada data latih (training set). Penelitian ini tidak menggunakan metode augmentasi data sintesis (oversampling).

5. Metrik Evaluasi: Tolok ukur perbandingan performa model dibatasi pada analisis Confusion Matrix (meliputi Accuracy, Precision, Recall, dan F1-Score) serta skor kurva ROC-AUC. Evaluasi secara khusus menitikberatkan pada optimasi nilai Recall guna meminimalkan kesalahan prediksi False Negative pada kelas penderita diabetes.
6. Implementasi Sistem: Luaran (output) akhir dari penelitian ini dibatasi pada perancangan purwarupa (prototype) aplikasi Clinical Decision Support System (CDSS) berbasis web sederhana menggunakan framework Streamlit, dan bukan berupa sistem rekam medis elektronik rumah sakit berskala penuh.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah menjawab berbagai masalah yang telah penulis uraikan pada perumusan masalah, yaitu :

1. Untuk mengetahui bagaimana algoritma machine learning Random Forest dan Support Vector Machine dapat digunakan untuk memprediksi penyakit diabetes mellitus.
2. Untuk mengetahui performa terbaik antara algoritma Random Forest dan Support Vector Machine dalam memprediksi penyakit diabetes mellitus.
3. Melakukan evaluasi mendalam dan membandingkan efektivitas algoritma Random Forest dan Support Vector Machine (SVM) dalam menghasilkan prediksi yang akurat pada dataset Diabetes Mellitus yang telah dikembangkan menjadi 10.000 *records* guna menjamin stabilitas model.

1.4.2 Manfaat Penelitian

Manfaat yang diperoleh dalam penelitian ini adalah :

1. Memberikan manfaat dalam dunia kesehatan untuk memprediksi penyakit diabetes mellitus khususnya pasien perempuan.

2. Memberikan manfaat dalam dunia pendidikan khususnya pada penelitian terkait dengan prediksi penyakit diabetes mellitus pada pasien dengan menggunakan teknik machine learning.

3. Memberikan rekomendasi algoritma klasifikasi terbaik antara Random Forest dan Support Vector Machine (SVM) melalui analisis komparatif performa pada dataset guna mendapatkan model prediksi yang memiliki tingkat reliabilitas dan stabilitas paling optimal.

1.5 Sistematika Penulisan

Sistematika penulisan pada penelitian ini dibagi menjadi 5 bagian, yaitu :

BAB I : PENDAHULUAN

Pada bab ini membahas secara jelas mengenai latar belakang penelitian, identifikasi masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II : LANDASAN TEORI

Pada bab ini berisikan berbagai teori yang berhubungan dengan algoritma machine learning random forest dan support vector machine dalam memprediksi penyakit diabetes mellitus.

BAB III : METODOLOGI PENELITIAN

Pada bab ini pembahasan mengenai implementasi metode algoritma random forest dan support vector machine yang dilakukan dalam memprediksi penyakit diabetes mellitus.

BAB IV : HASIL DAN PEMBAHASAN

Pada bab ini berisikan uraian pada hasil dan pembahasan dari penelitian yaitu analisis, pengolahan dataset, implementasi algoritma random forest dan support vector machine, dan melakukan perbandingan performa algoritma.

BAB V : PENUTUP

Pada bab ini berisikan kesimpulan dari hasil penelitian yang dilakukan dan saran terhadap kekurangan dari penelitian ini.