BAB I

PENDAHULUAN

1.1 Latar Belakang

Data mining adalah suatu pendekatan yang menggabungkan berbagai disiplin ilmu komputer untuk menemukan pola-pola baru dalam jumlah data yang besar. Hal ini melibatkan penggunaan metode yang berasal dari kecerdasan buatan, pembelajaran mesin, statistik, dan sistem basis data. Dalam konteks pengolahan data, data mining memiliki berbagai keuntungan, termasuk kemampuan untuk menghasilkan pengetahuan dan informasi baru dari data yang ada. Penggunaan data *mining* tidak hanya terbatas pada teknologi, tetapi juga dapat diterapkan dalam bidang kesehatan. Dalam konteks kesehatan, data *mining* dapat digunakan untuk memprediksi dan mendiagnosa jenis penyakit menggunakan metode yang sesuai. Salah satu contoh penerapan data mining dalam bidang kesehatan adalah prediksi dan diagnosa penyakit diabetes dengan menggunakan teknik-teknik data *mining* yang relevan (Sistem Komputer dan Sistem Informasi *et al.*, 2019).

Kasus kematian akibat diabetes di Indonesia terbesar keenam di dunia berdasarkan *International Diabetes Federation* mencatat diabetes telah menyebabkan 6,7 juta kematian di dunia pada 2021. Ini berarti ada 1 kematian setiap 5 detik. Tiongkok menjadi negara dengan jumlah kematian akibat diabetes terbesar di dunia. Kematian akibat diabetes di Tiongkok mencapai 1,39 juta orang pada 2021. Amerika Serikat berada di peringkat kedua dengan jumlah kematian sebanyak 669 ribu. Lalu, India berada di peringkat ketiga dengan jumlah sebesar 647ribu. Indonesia berada di peringkat keenam dalam daftar ini. Jumlah kematian akibat diabetes di Indonesia mencapai 236 ribu pada 2021. Pada 2021, IDF

menyebut ada 537 juta orang dewasa (usia 20 - 79 tahun) atau 1 dari 10 orang hidup dengan diabetes di seluruh dunia. 4 dari 5 orang penderita diabetes tinggal di negara berpendapatan rendah atau menengah (International Diabetes Federation, 2021). Penyakit diabetes merupakan salah satu masalah kesehatan yang berdampak luas terhadap jutaan individu di seluruh dunia. Pentingnya melakukan identifikasi dan diagnosis yang akurat untuk menentukan apakah seseorang menderita diabetes atau tidak, hal ini memegang peranan penting dalam menjalankan penanganan dan pengobatan yang sesuai (Kemenkes, 2020).

Algoritma klasifikasi merupakan suatu teknik yang digunakan untuk membuat suatu model yang mampu mengelompokkan data ke dalam kelas yang berbeda (Setio, Saputro dan Bowo Winarno, 2020). Pada konteks diagnosa penyakit diabetes, algoritma klasifikasi digunakan untuk melakukan prediksi apakah seseorang memiliki gejala diabetes atau tidak (Efendi dan Wibawa, 2018). *Support Vector Machine* dan *Extreme Gradient Boosting* merupakan dua algoritma yang memiliki kemampuan untuk mengklasifikasikan data. *SVM* adalah algoritma yang efektif dalam menangani data yang tidak linear dan mampu mengklasifikasikan data dengan jelas berdasarkan margin maksimal (Jumeilah, 2017). *XGBoost* adalah algoritma *boosting* yang memiliki kekuatan dalam menghadapi data yang rumit dan memberikan fleksibilitas dalam mengatur *hyperparameter* (Herni Yulianti, Oni Soesanto dan Yuana Sukmawaty, 2022). Meskipun *XGBoost* merupakan algoritma yang kuat, Terdapat satu kelemahan yang mendorong penggunaan kombinasi *XGBoost* dengan *Grid Search*, yaitu sensitivitas tinggi *XGBoost* terhadap nilai *default* parameter. *XGBoost* memiliki banyak parameter yang dapat disesuaikan, dan performanya sangat tergantung pada nilai parameter yang diatur dengan tepat.

Jika parameter tidak dioptimalkan dengan baik, kinerja *XGBoost* dapat mengalami penurunan (Raharjo, 2016).

Grid Search adalah sebuah metode yang digunakan untuk mencari parameter terbaik dalam sebuah model. Metode ini merupakan pendekatan komprehensif untuk mengeksplorasi kombinasi nilai parameter yang berbeda-beda. Dalam *Grid Search*, dapat menentukan jenis nilai prediksi yang ingin diuji terlebih dahulu, kemudian metode ini akan mencoba semua kombinasi nilai parameter untuk menemukan kombinasi yang menghasilkan kinerja terbaik (Gunawan, 2016). Dengan mengkombinasikan *XGBoost* dan *Grid Search*, dapat membantu untuk mencari kombinasi parameter yang optimal dan menghindari pengaturan yang buruk yang dapat mengurangi kinerja algoritma.

Dalam penelitian ini akan digunakan kombinasi algoritma *XGBoost* dengan *Grid Search*, dan membandingkan dengan *SVM*. Dikarenakan *SVM* memiliki kemampuan dalam mengatasi dataset dengan dimensi yang tinggi, Algoritma ini hanya membutuhkan jumlah vektor pendukung (*support vector*) yang relatif kecil untuk menentukan *hyperplane* pemisah. Hal ini membuat *SVM* menjadi lebih efisien dalam hal waktu pelatihan dan penggunaan memori jika dibandingkan dengan algoritma berbasis *tree* seperti *Random Forest* atau *Decision Tree*. (Harafani dan Wahono, 2015). Fokus utama dari penelitian ini adalah untuk mencari model yang paling efektif dalam mendiagnosa penyakit diabetes berdasarkan dataset yang ada.

1.2 Rumusan Masalah

Dalam penelitian ini, permasalahan yang diajukan adalah bagaimana membandingkan performa model yang dihasilkan oleh kombinasi algoritma *XGBoost* dengan penggunaan *hyperparameter Grid Search* dan performa model yang dihasilkan oleh algoritma *SVM* dalam mendiagnosis penyakit diabetes?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini antara lain:

- Sumber data yang digunakan adalah data sekunder yang diperoleh dari situs dan platform kompetisi Kaggle.
- Penelitian ini hanya memfokuskan pada analisis data yang tersedia dan tidak mencakup pengumpulan data baru atau penelitian lapangan.
- Dataset ini terdiri dari data pasien diabetes yang mencakup berbagai fitur klinis seperti umur, indeks massa tubuh, kadar gula darah, dan lain sebagainya.
- Algoritma *SVM* akan digunakan sebagai pembanding untuk melihat sejauh mana kombinasi *XGBoost* dengan *Grid Search* dapat meningkatkan performa prediksi dalam diagnosa penyakit diabetes.
- Pada penelitian ini, digunakan bahasa pemrograman Python dan platform Google Colab, serta aplikasi Microsoft Excel untuk memenuhi kebutuhan yang diperlukan.

1.4 Tujuan Penelitian

Adapun tujuan dalam penelitian ini untuk mengevaluasi apakah penggunaan kombinasi *XGBoost* dan *Grid Search* menghasilkan kinerja yang lebih baik dibandingkan *SVM* dalam hal akurasi dan kemampuan prediksi secara keseluruhan.

1.5 Manfaat Penelitian

Berdasarkan tujuan di atas, maka manfaat yang diharapkan dalam penelitian ini adalah sebagai berikut.

1. Dapat memperoleh pemahaman yang lebih mendalam tentang bagaimana algoritma *XGBoost* dengan *Grid Search* dan algoritma *SVM* bekerja dalam konteks diagnosa penyakit diabetes. Hal ini dapat memberikan wawasan tentang kelebihan, kelemahan, dan faktor-faktor yang mempengaruhi kinerja keduanya, sehingga memungkinkan untuk mengoptimalkan penggunaannya dalam situasi yang berbeda.
2. Penelitian ini memiliki potensi untuk berkontribusi dalam bidang penelitian ilmiah, khususnya dalam pengembangan metode diagnosa penyakit diabetes. Temuan dari penelitian ini dapat menjadi referensi bagi peneliti dan praktisi medis lainnya dalam upaya meningkatkan akurasi diagnosa penyakit diabetes dan mengembangkan algoritma klasifikasi yang lebih efektif.
3. Dengan adanya hasil diagnosa yang tepat dan pengelolaan yang efisien, pasien dapat mengendalikan penyakit diabetes dengan lebih baik. Ini akan membantu meningkatkan kualitas hidup, mengurangi risiko terjadinya komplikasi, dan memungkinkan mereka untuk menjalani kehidupan dengan lebih baik.
4. Penelitian ini memiliki potensi untuk menjadi pedoman dan referensi bagi penelitian selanjutnya yang fokus pada diagnosa penyakit diabetes. Temuan dan metodologi yang dihasilkan dari penelitian ini dapat menjadi dasar yang kuat untuk penelitian lanjutan yang berkaitan dengan penggunaan algoritma klasifikasi dalam konteks medis.