

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penyakit kardiovaskular (PKV) merupakan penyebab kematian utama di dunia, diperkirakan merenggut 17,9 juta jiwa setiap tahunnya. PKV adalah sekelompok gangguan jantung dan pembuluh darah, termasuk penyakit jantung koroner, penyakit serebrovaskular, penyakit jantung rematik, dan kondisi lainnya. Lebih dari empat dari lima kematian akibat PKV disebabkan oleh serangan jantung dan stroke, dan sepertiganya terjadi prematur pada orang di bawah usia 70 tahun[1].

Oleh karena itu prediksi untuk klasifikasi penyakit jantung sangat penting. Faktor risiko yang dapat dikendalikan dan yang tidak dapat dikendalikan dijelaskan dalam makalah ini. Prediksi penyakit jantung dilakukan dengan algoritma random forest machine learning[2]. Namun, performa model Random Forest sangat bergantung pada konfigurasi hyperparameter, seperti jumlah pohon, kedalaman maksimum, dan jumlah fitur yang digunakan.[3].

Untuk mengoptimalkan parameter tersebut, metode konvensional seperti Teknik Grid Search digunakan dalam mengeksplorasi kombinasi hiperparameter secara sistematis. Metode ini melakukan pencarian menyeluruh melalui subset hiperparameter yang telah ditentukan, memungkinkan identifikasi konfigurasi optimal yang memaksimalkan kinerja model[4].

Belakangan ini telah dikembangkan Auto Machine Learning (AutoML) untuk memudahkan praktisi dalam mengembangkan algoritma secara otomatis. AutoML merujuk pada konsep pelatihan machine learning dan tuning parameter secara otomatis dalam batas waktu yang ditentukan. Salah satu fitur AutoML disediakan oleh H2O dengan library open source yang dikembangkan oleh perusahaan H2O.ai bernama H2OAutoML[5].

Oleh karena itu, penting dilakukan penelitian komparatif antara H2O AutoML dengan fokus pada model Random Forest dan Random Forest yang dioptimasi menggunakan Grid Search, untuk menilai perbedaan kinerja keduanya dalam klasifikasi penyakit jantung[6]. Penelitian ini bertujuan untuk melakukan analisis terhadap performa kedua metode berdasarkan metrik akurasi, F1-score, dan AUC, serta melakukan uji statistik guna memastikan apakah ada perbedaan hasil yang diperoleh signifikan secara statistik.

1.2 Rumusan Masalah

1. Membandingkan hasil kinerja terbaik model Random Forest yang dioptimasi menggunakan Grid Search dan hasil kinerja dari model H2O AUTOML dalam klasifikasi penyakit jantung berdasarkan metrik evaluasi seperti Akurasi, F1-score, dan AUC

1.3 Batasan Masalah

Beberapa batasan masalah yang di gunakan pada penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya berfokus pada model Random Forest sebagai algoritma klasifikasi utama, baik pada implementasi H2O AutoML maupun Grid Search, sehingga hasil tidak mewakili algoritma lain seperti XGBoost atau SVM.
2. Proses optimasi hyperparameter pada metode Grid Search dilakukan secara manual dengan parameter terbatas (*n_estimators*, *max_depth*, dan *min_samples_split*), sedangkan H2O AutoML difokuskan hanya pada model Distributed Random Forest (DRF) untuk memastikan perbandingan yang seimbang.

1.4 Tujuan Penelitian

Tujuan penelitian merupakan pernyataan mengenai apa yang akan di hasilkan atau dicapai oleh peneliti

1. Apakah terdapat perbedaan kinerja yang signifikan secara statistik antara model Random Forest dengan Grid Search dan model H2O AutoML berdasarkan hasil evaluasi tersebut
2. Sebagai referensi dasar bagi pengembang sistem pendukung keputusan medis berbasis machine learning yang terkhususkan pada penerapan optimasi hyperparameter dengan menggunakan pendekatan H2O AutoML dan Grid Search pada algoritma random forest

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara langsung maupun tidak langsung. Adapun manfaat penelitian sebagai berikut:

a. Manfaat Teoritis

Penelitian ini berkontribusi pada pengembangan ilmu pengetahuan di bidang machine learning dan data mining yang terkhususkan pada penerapan optimasi hyperparameter dengan menggunakan pendekatan H2O AutoML dan Grid Search pada algoritma random forest dan hasil dari penelitian ini dapat menjadi acuan bagi peneliti selanjutnya untuk lebih memahami efektifitas masing masing metode optimasi pada dataset yang bersifat kompleks dan tidak seimbang

b. Manfaat Praktis

Penelitian ini dapat membantu meningkatkan akurasi klasifikasi resiko penyakit jantung, mengurangi waktu komputasi, serta mempercepat proses pengambilan keputusan berbasis data sehingga penelitian ini dapat menjadi dasar bagi pengembang sistem pendukung keputusan medis berbasis machine learning yang dapat di terapkan secara luas di institusi kesehatan

1.6 Sistematika Penulisan

Untuk memperoleh gambaran yang jelas dalam penulisan penelitian dan menjadi pedoman penulisan penelitian. Sistematika penelitian sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini membahas tentang : latar belakang masalah, rumusan masalah, Batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian dan sistematika

BAB II TINJAUAN PUSTAKA

Pada bab ini berisi tentang teori teori yang mendasar mengenai: tinjauan Pustaka, penelitian terdahulu, dasar teori,

BAB III METODE PENELITIAN

Bab ini akan membahas tentang metode apa saja yang digunakan dalam penelitian ini

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil dari masing masing model dan pembahasan perbandingan hasil kedua model

BAB V PENUTUP

Bab ini berisi tentang keseluruhan kesimpulan yang di peroleh dari uraian uraian bab sebelumnya dan menjawab rumusan masalah