

BAB I

PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, manfaat dan tujuan, ruang lingkup, dan sistematika penulisan pada skripsi yang berjudul Perbandingan Performa Algoritma Logistik Regresi, *Decision Tree*, *LightGBM* Untuk Prediksi Konflik Pada Klasifikasi Varian Genetik

1.1. Latar Belakang

Perkembangan teknologi dalam bidang *genomics* telah memungkinkan pengumpulan data genetik dalam skala yang belum pernah terjadi sebelumnya. Data genetik ini sangat kaya akan informasi yang dapat digunakan untuk memahami berbagai aspek biologis, termasuk prediksi risiko penyakit dan respons terhadap terapi medis. Salah satu tantangan utama dalam analisis data genetik adalah mengidentifikasi varian genetik yang memiliki dampak signifikan terhadap kesehatan. Varian genetik ini sering kali diklasifikasikan berdasarkan potensi mereka untuk menyebabkan penyakit atau memengaruhi fungsi biologis. Namun, sering terjadi perbedaan pendapat di antara para ahli mengenai klasifikasi varian tertentu, yang dikenal sebagai klasifikasi yang bertentangan. Prediksi yang akurat terhadap varian yang berpotensi memiliki klasifikasi yang bertentangan sangat penting untuk meningkatkan kualitas penilaian klinis dan penelitian.

Dalam konteks ini, penggunaan algoritma *machine learning* untuk memprediksi klasifikasi varian genetik telah menjadi topik penelitian yang menarik. Algoritma seperti *Logistic Regression*, *Decision Tree*, dan *LightGBM* menawarkan pendekatan yang berbeda dalam menganalisis data dan membuat prediksi. *Logistic Regression*, misalnya, adalah metode statistik yang sangat baik untuk memahami hubungan antara variabel dependen biner dan satu atau lebih variabel independen. *Decision Tree*, di sisi lain, memberikan representasi grafis yang intuitif dari proses pengambilan keputusan, yang memudahkan interpretasi hasil. *LightGBM*, yang merupakan algoritma *boosting*, dikenal dengan kemampuannya dalam menangani data dalam jumlah besar dengan efisiensi tinggi dan memberikan hasil yang akurat.

Namun, meskipun berbagai algoritma *machine learning* memiliki potensi yang besar, ada kebutuhan untuk melakukan evaluasi dan perbandingan performa di antara mereka untuk memastikan pemilihan model yang paling sesuai dengan masalah yang dihadapi. Penelitian ini bertujuan untuk membandingkan performa algoritma *Logistic Regression*, *Decision Tree*, dan *LightGBM* dalam memprediksi konflik klasifikasi pada varian genetik. Melalui pendekatan ini, diharapkan dapat diperoleh wawasan yang lebih mendalam mengenai kelebihan dan kekurangan masing-masing algoritma, serta memberikan rekomendasi yang didasarkan pada hasil empiris untuk aplikasi klinis dan penelitian lebih lanjut dalam *genomics*.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah disampaikan sebelumnya, maka dapat dirumuskan suatu masalah yaitu bagaimana membandingkan performa berbagai algoritma machine learning, yaitu Logistik Regresi, *Decision Tree*, dan *Light Gradient Boosting Machine (LightGBM)*, untuk prediksi konflik pada klasifikasi varian genetik. Penelitian ini untuk menentukan model mana yang paling efektif dalam mengidentifikasi konflik klasifikasi dalam varian genetik, guna memberikan panduan dalam pemilihan algoritma yang optimal untuk penelitian genetik.

1.3. Tujuan dan Manfaat

Tujuan penelitian ini adalah untuk mengetahui kinerja dari metode *Logistic regression*, *Decision Tree*, *Light Gradient Boosting Machine* dalam Prediksi konflik pada klasifikasi varian genetic.

Manfaat yang diharapkan dari penelitian ini adalah model yang dihasilkan dapat memberikan kontribusi dalam bidang Bioinformatika

1.4. Ruang Lingkup

Ruang lingkup bertujuan untuk memberikan batasan yang jelas pada skripsi ini agar tidak menyimpang dari target yang diinginkan. Ruang lingkup dalam skripsi ini adalah sebagai berikut:

1. Data yang digunakan adalah *Genetic Variant Classifications*.
2. Model *Logistik Regresi*, *Decision Tree*, dan *Light Gradient Boosting Machine (LightGBM)* dibangun menggunakan pemrograman Python dengan memanfaatkan *library Scikit-learn*.
3. Pelatihan dan pengujian model dilakukan di lingkungan *platform web Google Colaboratory*.
4. Beberapa parameter akan dioptimasi secara manual.
5. Metrik evaluasi yang digunakan untuk menilai kinerja model mencakup akurasi, *precision*, *recall*, F1-score, dan AUC-ROC.

1.5. Sistematika Penulisan

Sistematika penulisan memberikan gambaran dari skripsi ini secara urut dan jelas. Sistematika penulisan yang digunakan terbagi dalam beberapa pokok bahasan, yaitu:

BAB I	PENDAHULUAN
	Bab ini membahas mengenai latar belakang, rumusan masalah, manfaat dan tujuan, ruang lingkup, dan sistematika penulisan dari skripsi ini
BAB II	TINJAUAN PUSTAKA
	Bab ini membahas teori-teori yang berhubungan dan mendukung topik atau masalah yang dibahas pada skripsi ini.
BAB III	METODOLOGI PENELITIAN
	Bab ini menyajikan informasi mengenai metodologi penelitian yang digunakan pada penelitian prediksi konflik klasifikasi varian genetik pada pada sumber daya public menggunakan Light Gradient Boost Machine.
BAB IV	HASIL DAN PEMBAHASAN
	Bab ini menyajikan informasi tentang lingkungan dan perangkat yang digunakan untuk penelitian ini serta menyajikan analisis hasil penelitian tentang pengaturan secara manual hyper-parameter terhadap model yang dihasilkan.

BAB V PENUTUP

Bab ini berisi tentang kesimpulan dari bab-bab yang telah dibahas dan berisi saran sebagai bahan masukan untuk pengembangan atau penelitian selanjutnya