BAB I
PENDAHULUAN

Bab pendahuluan ini membahas mengenai latar belakang, rumusan masalah, tujuan, manfaat, ruang lingkup, serta sistematika penulisan dokumen skripsi berikut.

1.1 Latar Belakang

Varian genetik manusia adalah perbedaan-perbedaan yang ada pada rangkaian *Deoxyribonucleic Acid* (DNA) dalam genom individu-individu dari sebuah populasi (Ku dkk., 2010). Varian genetik dapat berupa beberapa bentuk, seperti perubahan atau pertukaran *single nucleotide*, pengulangan *tandem*, *insertion* dan *deletion* (*indels*), dan lain sebagainya. Secara umum, varian genetik terjadi secara natural pada genom manusia serta merupakan jejak kaki dari kesalahan-kesalahan yang terjadi pada replikasi DNA saat pembelahan sel, meskipun faktor eksternal, seperti virus dan mutagen kimia juga dapat menyebabkan perubahan pada rangkaian DNA (Ku dkk., 2010).

Seiring majunya ilmu pengetahuan, pengklasifikasian terhadap perubahan-perubahan tersebut menjadi suatu hal yang penting untuk dilakukan. Sebab berdasarkan penelitian dari Ku dkk., (2010) ditemukan bahwa varian genetik berhubungan dengan berbagai penyakit manusia, termasuk *monogenic* dan penyakit-penyakit yang lebih kompleks. Para peneliti dalam mengukur hubungan suatu varian dengan penyakit, menggunakan suatu spektrum yang disebut dengan *clinical significance* yang memiliki rentang mulai dari *benign*, *likely benign*, *uncertain significance*, *likely pathogenic*, dan *pathogenic* (Richards dkk., 2015). Hasil dari pengukuran inilah yang menjadi dasar dalam penentuan klasifikasi varian genetik.

Proses pengklasifikasian varian selama ini biasanya dilakukan manual tanpa bantuan teknologi kecerdasan seperti *machine learning* di laboratorium klinis oleh para peneliti (Arvai, 2020). Hasil klasifikasi kemudian diunggah ke arsip publik seperti *ClinVar*, yakni arsip publik untuk varian genetik manusia yang dikelola oleh *National Center for Biotechnology Information* (NCBI) Amerika Serikat. *ClinVar* menyimpan laporan varian genetik manusia dan interpretasi klinisnya terhadap penyakit, yang

bersumber dari laboratorium uji klinis, laboratorium penelitian, panel ahli, dan lainnya (Landrum dkk., 2017).

Pada kenyataannya, meskipun dilakukan oleh peneliti ahli, hasil dari pengklasifikasian tidak sepenuhnya sempurna (Arvai, 2020). Menurut Arvai (2020) alasannya adalah karena hasil pengklasifikasian suatu varian genetik dari laboratorium yang berbeda yang diunggah ke arsip publik *ClinVar*, belum tentu sama. Sebagai contoh untuk varian A yang diteliti oleh laboratorium X dan Y dapat memiliki kesimpulan *clinical significance* yang berbeda. Menurut Balmana dkk., (2016) inkonsistensi hasil klasifikasi tersebut dapat menimbulkan kesalahan dalam pengambilan keputusan medis yang berkaitan dengan *clinical significance* dari varian genetik individu. Sebelumnya, belum ada penelitian yang memanfaatkan metode selain *machine learning* untuk menangani permasalahan tersebut. Penelitian yang dilakukan oleh Kiran dkk., (2020) serta penelitian yang dilakukan oleh Musin dan Gaidel (2020) menawarkan solusi yang efisien dengan memanfaatkan teknologi *machine learning*, akan tetapi model *machine learning* dari kedua penelitian tersebut belum optimal. Berdasarkan kedua penelitian tersebut, *machine learning* menunjukkan potensi solusi yang efisien untuk menangani permasalahan klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar*, tetapi model *machine learning* yang dibangun harus seoptimal mungkin. Model yang optimal dapat menjadi sebuah alat pendeteksi inkonsistensi varian genetik manusia yang baik dan efisien, serta dapat menjadi bahan pertimbangan saat pengambilan keputusan. Manfaat tersebut membuat pengguna *ClinVar* dapat mengambil keputusan yang lebih tepat terkait data *clinical significance* varian genetik manusia.

Machine learning sendiri mulai sering digunakan untuk klasifikasi pada dunia medis, seperti pada penelitian Alharbi dan Vakanski (2023) untuk klasifikasi kanker atau seperti pada penelitian Feng dkk., (2023) untuk klasifikasi penyakit diabetes. *Machine learning* sangat membantu dalam diagnosis awal penyakit, pengembangan obat, peningkatan efisiensi operasional, serta peningkatan kualitas perawatan pasien (*European Institute of Innovation and Technology Health*, 2015). Lebih dari itu, *machine learning* juga sudah mulai dimanfaatkan agar dapat membantu dalam pertimbangan pengambilan keputusan pada dunia medis. Pada penelitian Sanchez-Martinez dkk., (2022) *machine learning* dimanfaatkan untuk pengambilan keputusan klinis pada kasus kardiologi dan pada penelitian Adlung dkk., (2021) *machine learning* dimanfaatkan untuk pengambilan

keputusan klinis pada kasus pemberian dosis insulin yang tepat untuk pasien diabetes tipe 1, serta kasus pengendalian gula darah dengan pemberian rekomendasi nutrisi yang dipersonalisasi untuk pasien diabetes tipe 2.

Penelitian-penelitian terdahulu terkait klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar*, seperti penelitian Kiran dkk., (2020) serta penelitian Musin dan Gaidel (2020) berfokus pada perbandingan antar algoritma *machine learning*. Penelitian dari Kiran dkk., (2020) adalah penelitian komparatif menggunakan algoritma, seperti *Gradient Boosting Classifier*, *Random Forest*, *Decision Trees*, *Logistic Regression*, dan *Neural Networks*, menunjukkan akurasi yang tinggi terhadap kelas 0 (konsisten) dinilai dari hasil *F1-score*. Kekurangan dari penelitian ini adalah kelas 1 (inkonsisten) sebagai kelas positif justru meraih hasil yang sangat buruk dimana tidak ada satupun model yang meraih *F1-score* di atas 9,00%, yang mengindikasikan bahwa seluruh model tidak lebih baik dari *random guessing* dalam melakukan klasifikasi. Pada penelitian tersebut juga dikatakan bahwa distribusi *kelas* condong ke kelas 0 yang berarti *dataset* tidak seimbang atau *imbalanced*. Pada kesimpulan akhir penelitian Kiran dkk., (2020) hasil terbaik ditorehkan oleh model *Random Forest* dengan skor *testing precision* 29,00%, *recall* 5,00%, dan *F1-score* 9,00% untuk kelas 1. *Dataset ClinVar* yang digunakan dalam penelitian tersebut bersumber dari *Kaggle*.

Pada penelitian komparatif lainnya yang dilakukan oleh Musin dan Gaidel (2020), mereka meneliti perbandingan algoritma *Random Forest* dan *Gradient Boosting Classifier* dengan *n_estimators* senilai 128. Hasil penelitian tersebut menunjukkan bahwa *Gradient Boosting Classifier* dapat memberikan hasil *testing F1-score* yang lebih baik dibanding *Random Forest* untuk kelas 0, yaitu sekitar 80,00%, sedangkan hasil untuk *Random Forest* sebesar 73,00%. Terkait hasil *testing F1-score* kelas 1 dari kedua algoritma adalah sama, yaitu sebesar 55,00% yang berarti algoritma sudah dapat melakukan klasifikasi berdasarkan pola pada data tetapi belum optimal. *Dataset* yang digunakan pada penelitian tersebut merupakan *dataset* yang sama dengan *dataset ClinVar* yang digunakan pada penelitian Kiran dkk., (2020). Kesimpulan akhir pada penelitian Musin dan Gaidel (2020) adalah *Gradient Boosting Classifier* lebih baik untuk klasifikasi *dataset ClinVar* dibandingkan dengan *Random Forest* karena meskipun *F1-score* sama, akan tetapi waktu latih dari *Gradient Boosting Classifier* lebih cepat dari waktu yang dibutuhkan oleh *Random Forest*. Secara keseluruhan, hasil penelitian dari Musin dan

Gaidel (2020) menghasilkan kesimpulan yang jauh lebih baik dari Kiran dkk., (2020) dinilai dari hasil *testing F1-score* kelas 1 sebagai kelas minoritas, mengingat *dataset* yang tidak seimbang.

Berdasarkan dua penelitian terhadap varian genetik manusia dengan menggunakan *dataset ClinVar* tersebut, dapat disimpulkan bahwa *dataset* yang tidak seimbang menyebabkan munculnya resiko *underfit* ataupun *overfit*. Penelitian-penelitian tersebut juga memiliki beberapa keterbatasan, seperti pada penelitian Kiran dkk., (2020) berfokus pada perbandingan algoritma, tidak adanya percobaan *hyperparameter tuning*, tidak adanya penanganan terhadap ketidakseimbangan data, dan kurangnya metrik evaluasi yang digunakan. Pada penelitian Musin dan Gaidel (2020) terdapat beberapa keterbatasan, seperti fokus pada perbandingan algoritma, *hyperparameter* yang di-*tuning* hanya satu, kurangnya penanganan terhadap ketidakseimbangan data, dan sedikitnya metrik evaluasi yang digunakan. Berbagai keterbatasan tersebut tentu dapat mempengaruhi performa dan penilaian model, serta mengingat pula proporsi data yang tidak seimbang yaitu 48754 (74,79%) untuk kelas 0 dan 16434 (25,21%) untuk kelas 1. Berdasarkan faktor-faktor tersebut, metode yang digunakan untuk membangun model harus difokuskan untuk menangani berbagai keterbatasan pada penelitian-penelitian sebelumnya, seperti *hyperparameter tuning* yang lebih beragam, penerapan metode untuk menangani ketidakseimbangan data, dan penggunaan metrik evaluasi yang lebih cocok untuk mengukur performa model terhadap *imbalanced dataset*.

Gradient Boosting Classifier (GBC) sendiri merupakan algoritma yang sudah sering digunakan untuk melakukan klasifikasi terhadap berbagai jenis *dataset*. Pada penelitian yang dilakukan oleh Suryana dkk., (2021) untuk klasifikasi keberhasilan *telemarketing* bank dengan *dataset* tidak seimbang, GBC yang sudah menerapkan *hyperparameter tuning* meraih skor tinggi dengan akurasi 90,39%, *Precision* 94,91%, dan AUC 0,939. Selanjutnya, terdapat penelitian komparatif yang dilakukan oleh Bentejac dkk., (2019) dengan menggunakan GBC, *Random Forest*, dan *XGBoost* untuk klasifikasi berbagai *dataset* dari UCI *repository*. Pada penelitian tersebut, GBC yang sudah melalui proses *hyperparameter tuning* terpilih sebagai *classifier* dengan akurasi terbaik dinilai dari total performa positif dari keseluruhan *dataset*. Berdasarkan pemaparan hasil dua penelitian tersebut, *hyperparameter tuning* menjadi tahap yang penting untuk diterapkan apabila ingin mendapatkan hasil yang optimal ketika

mengimplementasikan GBC. Ketelitian dalam *hyperparameter tuning* diperlukan agar *Gradient Boosting* tidak *overfit* dan mendapatkan hasil yang optimal (Bentejac dkk., 2019).

Pada kasus ketidakseimbangan *dataset*, metode paling umum yang sering digunakan adalah *balancing data*. Di sisi lain, menurut beberapa penelitian seperti penelitian Luo dkk., (2024) metode *balancing data*, seperti *over-sampling* dapat menghasilkan *redundant data*, sedangkan *under-sampling* dapat menghilangkan informasi berharga pada data. Terakhir, *hybrid-sampling* yang mengkombinasikan dua metode *sampling* sebelumnya beresiko menambah sampel yang tidak bermakna pada data. Pada penelitian lainnya yang dilakukan oleh Blagus dan Lusa (2013), menunjukkan bahwa performa dari metode *Synthetic Minority Over-sampling Technique* (SMOTE) cenderung menurun ketika diterapkan pada *dataset* berdimensi tinggi. Ini menunjukkan bahwa SMOTE justru berpotensi untuk menghasilkan *noise*, dan mengurangi performa *classifier* pada kasus *dataset* berdimensi tinggi. Tarawneh dkk., (2022) juga berargumen bahwa sampel buatan yang dibuat oleh SMOTE dapat menyebabkan *overfitting* karena sampel minoritas yang dibuat, besar kemungkinannya merupakan sampel mayoritas sehingga penggunaannya harus dihindari pada *real-world applications*. Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa metode *balancing data* tidak selalu menjadi cara terbaik dalam menangani kasus ketidakseimbangan data. Metode *balancing data* seperti SMOTE justru dapat berakibat buruk pada performa *classifier*, dan penerapannya untuk kasus dunia nyata sebaiknya dihindari (Tarawneh dkk., 2022).

Berdasarkan alasan-alasan tersebut, diperlukan metode alternatif selain *balancing data* untuk menangani kasus ketidakseimbangan data. Salah satu metode alternatif yang dapat diterapkan adalah pembobotan sampel atau *sample weighting* (Bakirarar & Elhan, 2022; Shrivastava, 2020). Pembobotan sampel merupakan metode untuk menangani ketidakseimbangan data, dengan menerapkan bobot pada setiap sampel kelas saat tahap pelatihan model. Menurut Bakirarar dan Elhan (2022) bobot ini meningkatkan kepekaan model terhadap kelas minoritas, sedangkan menurunkan kepekaan model terhadap kelas mayoritas. Ini dapat berdampak positif pada kasus distribusi sampel yang tidak seimbang pada setiap kelasnya, seperti pada penelitian Bakirarar dan Elhan (2022). Penelitian tersebut berfokus pada perbandingan berbagai metode pembobotan, seperti *Inverse of Number of Samples* (INS), *Inverse of Square Root of Number of Samples* (ISNS),

Effective Number of Samples (ENS), dan juga *Sample Based Class Weight* (SBCW) pada *dataset* medis. Berdasarkan penelitian tersebut, disimpulkan bahwa metode *sample weighting* terbaik adalah SBCW dengan skor akurasi 74,50%, *G-Mean* 71,70%, *balanced accuracy* 23,70%, dan *Matthews Correlation Coefficient* (MCC) sebesar 42,20%. Berdasarkan kesimpulan penelitian tersebut, *sample weighting* dengan metode SBCW dapat menjadi solusi alternatif untuk menangani ketidakseimbangan data pada *dataset ClinVar*.

Berlandaskan pertimbangan seluruh keunggulan yang telah dijelaskan sebelumnya, penelitian ini mengimplementasikan algoritma *Gradient Boosting Classifier* untuk klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar*. Selain itu, saat proses pelatihan model juga akan diterapkan *hyperparameter tuning* serta pembobotan sampel. Hal ini penting untuk diterapkan sebagai upaya meningkatkan kinerja model pada *dataset* yang tidak seimbang, agar hasil klasifikasi menjadi optimal.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, maka dapat dirumuskan rumusan masalah yaitu bagaimana implementasi dan kinerja dari algoritma *Gradient Boosting Classifier* untuk klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar* dengan penerapan *hyperparameter tuning* serta pembobotan sampel.

1.3 Tujuan

Tujuan dari penelitian ini adalah untuk mengoptimalkan implementasi dan mengevaluasi kinerja algoritma *Gradient Boosting Classifier* dalam membangun model terbaik untuk klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar* dengan menerapkan *hyperparamter tuning* dan pembobotan sampel.

1.4 Manfaat

Manfaat yang diharapkan dari penelitian ini adalah dapat dibangunnya suatu model berbasis algoritma *Gradient Boosting Classifier* yang dapat melakukan klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar* dengan baik.

1.5 Ruang Lingkup

Ruang lingkup bertujuan untuk memastikan bahwa skripsi ini sesuai dan tidak melenceng dari tujuan yang sudah ditetapkan. Ruang lingkup dari skripsi ini, yaitu:

1. *Dataset* yang digunakan dalam penelitian ini adalah *ClinVar* yang diperoleh dari *Kaggle*. *Clinvar* merupakan *public archive* yang dikelola oleh *National Center for Biotechnology Information*.
2. Model *Gradient Boosting Classifier* dibangun dengan bahasa pemrograman *Python* dengan memanfaatkan *library Scikit-learn*.
3. Pelatihan dan pengujian model dilakukan pada lingkungan *Jupyter Notebook* dari *Anacoda*.
4. Evaluasi model dilakukan dengan menggunakan *confusion matrix*.
5. Validasi model dilakukan dengan menggunakan *Stratified K-Fold Cross-Validation*.
6. *Hyperparameter* yang akan dicari kombinasi terbaiknya meliputi *n_estimators*, *max_depth*, *learning_rate*, dan *max_features*.

1.6 Sistematika Penulisan

Sistematika penulisan bertujuan untuk memberikan gambaran laporan yang urut dan jelas dari skripsi. Berikut adalah sistematika penulisan skripsi ini.

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, manfaat dan tujuan, ruang lingkup, dan sistematika penulisan skripsi mengenai Implementasi Algoritma *Gradient Boosting Classifier* untuk klasifikasi inkonsistensi varian genetik pada arsip publik *ClinVar*.

BAB II LANDASAN TEORI

Bab ini membahas teori-teori yang berkaitan dengan topik atau masalah yang dibahas pada skripsi ini. Bab ini terdiri dari *state-of-the-art*, varian genetik manusia, *ClinVar dataset*, klasifikasi, *gradient boosting classifier*, *decision tree regression*, *spearman correlation coefficient*, metode imputasi, *interquartile range*, *winsorization*, metode *encoding* data, *robust*

scaling, pembobotan sampel, *cross-validation*, *grid search*, evaluasi model, juga *tools* dan *library*.

BAB III METODOLOGI PENELITIAN

Bab ini menyajikan informasi mengenai metodologi penelitian yang digunakan pada penelitian Implementasi Algoritma *Gradient Boosting Classifier* untuk klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar*. Tahapan yang dilakukan dalam penyelesaian masalah di antaranya adalah garis besar penyelesaian masalah, pengumpulan data, *data preprocessing*, optimasi *hyperparameter*, pelatihan model, pengujian model, dan evaluasi kinerja model. Pada bab ini juga terdapat perhitungan manual dari metode yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan informasi tentang lingkungan dan perangkat yang digunakan dalam penelitian, skenario pelatihan, serta menyajikan pembahasan dari analisis hasil penelitian dalam Implementasi Algoritma *Gradient Boosting Classifier* untuk klasifikasi inkonsistensi varian genetik manusia pada arsip publik *ClinVar*.

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.