

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Salah satu penyakit yang memiliki tingkat kelangsungan hidup yang bervariasi di antara pasien adalah kanker tiroid. Untuk memprediksi tingkat kelangsungan hidup pasien berdasarkan karakteristik klinis. Kanker yang berasal dari kelenjar tiroid organ kecil berbentuk kupu-kupu di bagian depan leher terjadi ketika sel-sel tiroid tumbuh secara tidak terkendali dan membentuk tumor ganas. Karsinoma folikuler dan papiler adalah dua jenis kanker tiroid yang paling umum. Meskipun gejala kanker tiroid dapat bervariasi, beberapa yang paling umum termasuk pembengkakan atau benjolan di leher, perubahan suara, masalah menelan, dan kadang-kadang nyeri. Jenis dan stadium kanker tiroid menentukan pengobatannya, yang dapat mencakup operasi, terapi radiasi, atau terapi hormon [1]. Dalam menghadapi tantangan ini, teknologi memainkan peran penting dalam membantu analisis dan prediksi, terutama melalui perkembangan kecerdasan buatan.

Masalah utama dalam diagnosis kanker tiroid adalah kesulitan dalam membedakan antara tumor jinak dan ganas, terutama pada tipe folikular, karena biopsi aspirasi jarum halus (BAJAH) tidak selalu memberikan hasil yang pasti dan sangat bergantung pada keahlian ahli sitopatologi. Selain itu, meskipun ultrasonografi (USG) sering digunakan untuk mendeteksi nodul tiroid, metode ini tidak selalu mampu membedakan karakteristik keganasan secara akurat [2].

Segala aktivitas manusia di era modern tentu tidak lepas dari penggunaan teknologi seperti komputer, laptop, ponsel pintar, dan lainnya. Seiring berkembangnya teknologi, komputer dilatih

menggunakan bahasa pemrograman yang dapat belajar dari kumpulan data untuk melakukan tugas tertentu seperti mengenali pola, membuat prediksi, dan membuat keputusan tanpa melalui program secara eksplisit disebut *Machine learning* [3].

Machine learning terbagi menjadi tiga metode yaitu *Supervised learning*, *unSupervised learning*, dan *reinforcement learning*. Klasifikasi adalah salah satu pendekatan dari metode *Supervised learning* yang dimana pada metode ini pemberian kelas pada dataset dibutuhkan untuk melihat tingkat akurasi kinerjanya serta mempelajari hubungan dan ketergantungan antar data [4]. Algoritma seperti *LOGISTIC REGRESSION*, *RANDOM FOREST*, *GRADIENT BOOSTING* DAN *K-Nearest Neighbor (KNN)* menjadi pilihan karena kemampuan mereka untuk menangani berbagai jenis data.

Logistic Regression merupakan pembelajaran terbimbing pada masalah regresi dan klasifikasi. *Logistic Regression* menggunakan probabilitas untuk memprediksi klasifikasi data kategorikal. Nilai input dapat digabungkan secara linier menggunakan fungsi sigmoid atau logistik dan nilai koefisien untuk memprediksi hasil. Estimasi kemungkinan maksimum dianggap menggunakan fungsi sigmoid untuk mengasumsikan data yang paling mungkin, dan probabilitas diberikan antara 0 sampai 1 mengatakan apakah suatu peristiwa akan terjadi atau tidak [5].

Random Forest adalah sebuah algoritma yang digunakan dalam data mining untuk membuat model prediksi. Algoritma ini menggunakan sejumlah *Decision Tree* yang masing-masing dibangun dengan menggunakan sample data yang diambil secara acak dari data asli. Kemudian, hasil prediksi dari setiap *Decision Tree* diagregasi menjadi satu hasil akhir. *Random Forest* dapat digunakan hanya untuk klasifikasi (memprediksi kelas suatu data), tidak bisa digunakan untuk regresi (memprediksi nilai numerik suatu data) seperti yang bisa dilakukan oleh *Decision Tree* [6].

Gradient Boosting merupakan algoritma klasifikasi yang terdapat pada machine learning. Algoritma *gradient boosting* bekerja dengan membangun model secara bertahap, setiap model baru menghasilkan tahapan untuk mencoba memperbaiki kesalahan dari model sebelumnya [7].

Algoritma *K-Nearest Neighbor (KNN)* merupakan salah satu metode *Supervised learning* karena hasil *query instance* perlu diklasifikasikan berdasarkan mayoritas dari kategori KNN. Pengumpulan data menggunakan metode observasi. Algoritma KNN memiliki beberapa kelebihan diantaranya ketangguhan terhadap data training yang memiliki banyak noise dan data dalam jumlah yang besar [8]. Dalam konteks ini, metode *Logistic Regression*, *Random Forest*, *Gradient Boosting* dan *K-Nearest Neighbor (KNN)* dapat digunakan untuk prediksi penyakit kanker tiroid dengan *Machine learning*.

Namun, dalam implementasinya, seringkali ditemukan ketidakseimbangan data antara kelas mayoritas dan minoritas yang dapat mempengaruhi performa model prediksi. Untuk mengatasi masalah ini, dapat diterapkan teknik *SMOTE (Synthetic Minority Over-sampling Technique)*. SMOTE dapat menghasilkan data sintetik baru berdasarkan data kelas minoritas yang ada tanpa mereplikasinya untuk mengatasi tantangan Overfitting. SMOTE pertama-tama memilih setiap sampel dari sampel minoritas secara berurutan sebagai sampel akar untuk sintesis sampel tambahan untuk mensintesis sampel baru dan diulang sebanyak n kali. Terakhir, interpolasi linier dilakukan antara sampel dan setiap sampel tambahan untuk menghasilkan n sampel yang disintesis. [9].

Selanjutnya ada fitur *Recursive Feature Elimination (RFE)* adalah metode yang digunakan untuk memilih fitur-fitur penting dalam data. Bayangkan kita memiliki banyak sekali fitur, tapi tidak semuanya berguna untuk memprediksi sesuatu. RFE dapat meningkatkan akurasi prediksi serta efisiensi komputasi, terutama saat bekerja dengan dataset berdimensi besar [10].

Selain 2 fitur diatas ada satu fitur *Principal Component Analysis (PCA)* sering digunakan dalam dalam mengekstrak informasi dari suatu data. Pca tergolong dalam prosedur statistik yang bertujuan membangun dataset baru yang tidak berkorelasi antar variable tanpa menghilangkan informasi. Hasil rekonstruksi tersebut sering disebut komponen utama. Principalkomponen disusun dari komponen varian maksimum berurutan setelahnya nilai varianmenurun. Sederhananya komponen pertama memiliki varian tertinggi, komponen seterusnya memiliki varian yang lebih kecil [11]. Dengan demikian, penerapan SMOTE, *Recursive Feature Elimination (RFE)* dan *Principal Component Analysis PCA* mampu meningkatkan keseimbangan distribusi kelas pada dataset, yang pada akhirnya dapat membantu algoritma seperti KNN, Logistic Regression, Random Forest, dan Gradient Boosting dalam membangun model prediksi yang lebih akurat dan adil

Sejumlah penelitian sebelumnya telah menunjukkan bahwa algoritma *Machine learning* memiliki potensi besar dalam memprediksi berbagai penyakit, termasuk kanker tiroid. Salah satu studi yang relevan adalah penelitian oleh Putri et al, yang menerapkan algoritma *Decision Tree C4.5* untuk memprediksi tingkat kelangsungan hidup pasien kanker tiroid dan mencapai akurasi sebesar 97% dengan nilai AUC 0.95 [1].

Penelitian lain yang mendukung efektivitas metode *machine learning* dalam bidang ini dilakukan oleh Nurjanah et al., yang memanfaatkan *algoritma Random Forest* untuk mengklasifikasikan risiko kambuhnya kanker tiroid. Berdasarkan data rekam medis pasien, algoritma ini berhasil mencapai akurasi sebesar 94,25%. *Random Forest* dipilih karena kemampuannya dalam menangani variabel yang banyak serta memberikan hasil prediksi yang stabil, menjadikannya salah satu metode yang andal untuk analisis data medis [12].

Penelitian yang mendukung efektivitas metode *machine learning* dalam konteks ini dilakukan oleh Leli safitri et al, yang memanfaatkan *algoritma Decision Tree C4.5* dan *Naive Bayes* untuk memprediksi penyakit tiroid. Berdasarkan dari hasilnya, Hasil pengujian menunjukkan bahwa akurasi lebih baik diperoleh dari *Algoritma Decision Tree C4.5* sebesar 97,12% sedangkan nilai akurasi *Algoritma Naïve Bayes* sebesar 76,02%. Nilai *Area Under Curve (AUC)* pada *kurva Receiver Operating Characteristic (ROC)* menunjukkan *Algoritma Decision Tree C4.5* memiliki nilai lebih tinggi dari *Algoritma Naïve Bayes* dengan hasil klasifikasi *Good Classification* [13].

Berdasarkan latar belakang yang sudah dijelaskan maka peneliti tertarik untuk meneliti dengan topik ” **PERBANDINGAN DAN VISUALISASI DATA DALAM PREDIKSI PENYAKIT KANKER TIROID MENGGUNAKAN ALGORITMA MACHINE LEARNING**”.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang diatas, dapat dirumuskan sebagai berikut:

1. Bagaimana performa keseluruhan, sensitivitas, spesifisitas, dan akurasi model *Logistic Regression*, *Random Forest*, *Gradient Boosting* dan *K-Nearest Neighbor (KNN)* dalam memprediksi penyakit kanker tiroid ?
2. Bagaimana hasil prediksi dari setiap model berbeda, terutama dalam hal membantu memahami pola data dan membuat keputusan?
3. Apakah model yang memiliki akurasi terbaik juga memiliki kompleksitas komputasi yang paling ideal untuk digunakan dalam sistem prediksi kehidupan nyata?
4. Sejauh mana setiap model dapat mengidentifikasi pola data pada kumpulan data kanker tiroid, terutama dengan data yang tidak seimbang atau bervariasi?

1.3 BATASAN MASALAH

Untuk menghindari berbicara tentang hal-hal di luar tema dan judul penelitian, masalah yang akan dijadikan acuan akan dibatasi oleh batas-batas berikut:

1. Penelitian ini akan menggunakan *dataset* yang telah tersedia dari *kaggle*.
“<https://www.kaggle.com/code/ashikshahriar/thyroid-cancer-risk-prediction/input>”
2. Dataset yang digunakan dalam penelitian ini dibatasi pada dataset kanker tiroid yang mencakup variabel-variabel standar seperti *Patient_ID*, *Age*, *Gender*, *Country*, *Ethnicity*, *Family_History*, *Radiation_Exposure*, *Iodine_Deficiency*, *Smoking*, *Obesity*, *Diabetes*, *TSH_Level*, *T3_Level*, *T4_Level*, *Nodule_Size*, *Thyroid_Cancer_Risk*, *Diagnosis*.
3. Model *Machine learning* yang dibandingkan terbatas pada empat algoritma, yaitu:
 - a. *Logistic Regression*
 - b. *Random Forest*
 - c. *Gradient Boosting*
 - d. *K-Nearest Neighbor (KNN)*
4. Penelitian ini akan menggunakan 3 fitur yaitu:
 - a. Synthetic Minority Oversampling Technique (SMOTE)
 - b. *Recursive Feature Elimination (RFE)*
 - c. *Principal Component Analysis (PCA)*

1.4 TUJUAN PENELITIAN

Adapun tujuan penelitian adalah:

1. Menganalisis perbandingan efektivitas prediksi data dari empat model *Machine learning* (*Logistic Regression*, *Random Forest*, *Gradient Boosting* dan *K-Nearest Neighbor (KNN)*) dalam memprediksi penyakit kanker tiroid.

2. Mengidentifikasi model *Machine learning* yang memberikan representasi visual paling informatif dan mudah diinterpretasi untuk prediksi kanker tiroid.
3. Mengevaluasi kekuatan dan kelemahan masing-masing model dalam hal kemampuan prediksi performa dan interpretabilitas hasil prediksi.

1.5 MANFAAT PENELITIAN

Adapun manfaat penelitian adalah:

1. Penelitian ini dapat memberikan kontribusi pada pengembangan ilmu pengetahuan terkait penggunaan metode *Logistic Regression*, *Random Forest*, *Gradient Boosting* dan *K-Nearest Neighbor (KNN)* dalam prediksi penyakit kanker tiroid. Hasil penelitian dapat menjadi dasar penelitian selanjutnya dan pengembangan metode prediksi penyakit kanker tiroid.
2. Membantu tenaga medis dalam memilih model *Machine learning* yang lebih efektif dan interpretatif untuk mendukung diagnosis kanker tiroid secara lebih akurat dan efisien.
3. Mendorong pemanfaatan teknologi *Machine learning* dalam dunia medis untuk meningkatkan akurasi dan efisiensi dalam pengambilan keputusan klinis
4. Memberikan referensi bagi pengembang sistem kesehatan berbasis *AI* dalam memilih algoritma yang sesuai dengan kebutuhan prediksi kanker tiroid.

1.6 SISTEMATIKA PENULISAN

Supaya penyusunan laporan dapat mudah dan teratur, penulis telah menyusun rencana yang terbagi dari 5 bab, secara spesifik sebagai berikut:

BAB I : PENDAHULUAN

Pada Bab Ini memaparkan bagian latar belakang masalah, perumusan permasalahan, pembatasan permasalahan, tujuan dan manfaat dari penelitian, serta sistematika dari penulisan.

BAB II : LANDASAN TEORI

Bagian ini akan menjelaskan beberapa teori dasar berkaitan dengan penelitian dan teori pendukung lain, berkaitan dengan penelitian ini.

BAB III : METODOLOGI PENELITIAN

Bagian ini akan memaparkan struktur penelitian serta menguraikan langkah-langkah melaksanakan penyelesaian penelitian.

BAB IV : ANALISIS

Bab keempat membahas berbagai analisa proses pengumpulan data klasifikasi yang sesuai dengan topik yang diambil.

BAB V : KESIMPULAN

Bagian terakhir akan menguraikan sebuah kesimpulan yang telah diambil dari seluruh data penelitian ini serta beberapa saran untuk meningkatkan mutu kualitas penelitian selanjutnya.