

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tiroid merupakan kelenjar endoktrin terbesar yang sangat penting bagi tubuh. Kelenjar ini berbentuk seperti kupu-kupu yang terdapat dibagian depan leher[1]. Adapun peran penting kelenjar tiroid bagi tubuh yaitu untuk menjaga metabolisme dalam tubuh dan menghasilkan banyak oksigen untuk menghasilkan protein dalam tubuh. Adanya gangguan saat kelenjar tiroid bekerja sehingga menyebabkan munculkan gangguan hipertiroid yaitu cara bekerja kelenjar tiroid yang dilakukan secara berlebihan dan hipotiroid sebaliknya yaitu kurangnya kelenjar tiroid dalam bekerja. Akibat gangguan tersebut dapat mengakibatkan seseorang terkena penyakit gondok [2]. Namun, hampir sebagian orang menganggap penyakit ini adalah penyakit yang tidak perlu dikhawatirkan. Penyakit metabolik ini menempati posisi tertinggi setelah penyakit *Diabetes Mellitus*. Jumlah data yang mengalami penyakit tiroid di dunia berdasarkan kementerian kesehatan dunia masih tergolong tinggi pada bayi baru lahir yang terkena hiperthyroid dengan angka 1600 tiap tahun [3] dan banyak didominasi oleh perempuan dengan prevalensi terjadi 4 hingga 10 kali dibandingkan pria dengan angka persentase 14,7% berdasarkan data dari Kemenkes RI tahun 2015 [4].

Klasifikasi merupakan teknik pendekatan pada *data mining* dengan cara mengelompokkan data berdasarkan kelas yang sesuai. Klasifikasi termasuk pada teknik pembelajaran terbimbing (*supervised learning*) dengan mempelajari data yang sudah diberikan label sebelumnya [5]. Dalam penerapan klasifikasi dibantu dengan beberapa algoritma yang digunakan dan salah satu metode sering digunakan adalah algoritma *Support Vector Machine*. Algoritma *Support Vector Machine* merupakan algoritma terbaru yang dapat memprediksi atau mengklasifikasi suatu data dengan membagi data berdasarkan kelompoknya. Untuk membagi dua kelas tersebut terdapat margin dan juga menggunakan fungsi *hyperplane* terbaik untuk memisahkan kelas tersebut secara maksimal [6]. Pada ruang kerja dimensi tinggi, ditambahkan *kerner trick* untuk menemukan *hyperplane* terbaik dengan pada fungsi pertidaksamaan linear. Meskipun algoritma ini mempunyai nilai keakuratan yang tinggi, akan tetapi algoritma SVM dalam melakukan pengolahan data terdapat

adanya kompleksitas komputasi atau *Quadratic Programming* sehingga akan membutuhkan waktu perhitungan yang lama dan juga terdapat parameter seperti kelas dan jenis kernel yang digunakan dan mempengaruhi performansi dari algoritma SVM sendiri[7]. Selain itu, penggunaan data yang digunakan dalam pengoperasian dengan jumlah kelas lebih dari dua atau *multiclass* dapat memberikan pengaruh dalam keakuratan yang dihasilkan dan juga data mudah sensitif terhadap noise dan outlier pada data yang mempengaruhi hasil klasifikasi[8]. Metode *Least-Squares Support Vector Machine* (LS-SVM) dikembangkan sebagai solusi permasalahan dengan metode SVM yang telah dikembangkan kembali oleh para ahli untuk mempermudah perhitungan data. Metode ini dapat digunakan pada data kelas yang memiliki jumlah kelas yang banyak atau multi kelas dan dapat diterapkan pada data *real time* atau diproses dalam waktu yang singkat. Metode LS-SVM ini memiliki beberapa keunggulan dibandingkan dengan metode SVM sendiri adalah perhitungan komputasi dengan mudah dengan menggunakan fungsi persamaan linear. LS-SVM ini hampir sama seperti metode SVM yaitu menggunakan persamaan non-linear yang lebih disederhanakan pada persamaan linear untuk mengatasi adanya permasalahan kuadratik dan juga dapat menghasilkan tingkat akurasi yang lebih baik [9]. Tingkat efisiensi yang dimiliki metode LS-SVM ini mampu menyelesaikan permasalahan dalam komputasi dengan skala yang lebih besar dari SVM [10]. Dalam penerapan pada kedua metode tersebut, adapun penelitian mengenai prediksi penyakit hepatitis dengan menggunakan algoritma LS-SVM menghasilkan tingkat keakuratan yang tinggi sebesar 93,7% [11].

Penerapan metode LS-SVM ini juga terdapat masalah dalam fitur yang akan digunakan. Penerapan seleksi fitur ini bertujuan agar dapat mempercepat proses klasifikasi dan kinerja dari metode yang akan digunakan [12]. Salah satu seleksi fitur yang digunakan adalah *Information Gain*. Dalam memaksimalkan hasil klasifikasi yang dilakukan dengan cara memilih nilai tertinggi, seleksi fitur *Information Gain* dapat memeriksa karakteristik dan mengantisipasi label kelas dengan tepat. Kelebihan lain dalam seleksi fitur ini dapat digunakan pada data yang berdimensi tinggi sehingga dapat mengurangi terjadinya resiko *overfitting*. Pada penelitian yang dilakukan oleh [13] penggunaan seleksi fitur *Information Gain*

yang diterapkan pada klasifikasi *machine learning* (J48, Random Forest, MLP, SVM (SMO), dan Naïve Bayes) dalam memprediksi tampilan akademik siswa menghasilkan nilai akurasi yang cukup tinggi karena adanya pengurangan fitur yang tidak diperlukan. Penelitian selanjutnya dilakukan oleh [14] menghasilkan akurasi lebih tinggi menggunakan seleksi fitur *information gain* yakni 84% dan implementasi tanpa seleksi fitur sebesar 74,33% dalam analisis sentimen pada aplikasi Peduli Lindungi.

Klasifikasi penyakit tiroid menjadi salah satu tantangan penting dalam bidang kesehatan. penggunaan metode LS-SVM banyak digunakan oleh peneliti dalam melakukan klasifikasi penyakit. Namun, belum banyak peneliti yang menggunakan metode *hybrid* LS-SVM dengan seleksi fitur *Information Gain* pada jumlah dataset yang besar. Selanjutnya untuk mengatasi ketidakseimbangan data dilakukan teknik SMOTE untuk menyeimbangkan kelas minoritas. Penggabungan kedua metode pada penelitian ini bertujuan agar dapat digunakan untuk mengklasifikasikan penyakit tiroid secara lebih efektif.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang ada, rumusan masalah dapat diuraikan sebagai berikut:

1.2.1 Permasalahan

Dalam menghadapi klasifikasi penyakit tiroid terdapat jumlah data pada setiap kelas penyakit tiroid yang tidak seimbang. Kemudian permasalahan selanjutnya adalah data tiroid dengan dimensi dataset yang tinggi, maka pada penelitian ini dilakukan dengan tujuan untuk mengetahui keefektifan penggunaan seleksi fitur dalam mengatasi jumlah dimensi dataset yang tinggi menggunakan seleksi fitur *Information Gain* yang digabungkan dengan metode Least Square Support Vector Machine (LS-SVM).

1.2.2 Solusi Permasalahan

Dalam penelitian ini mengusulkan penerapan seleksi fitur *Information Gain* dan metode LS-SVM pada klasifikasi penyakit tiroid dan juga menggunakan teknik SMOTE sebagai penyeimbang dengan jumlah kelas data minoritas.

1.2.3 Pertanyaan Penelitian

1. Berapa nilai akurasi, *precision*, *recall*, dan *f-1 score* dengan metode LS-SVM dan seleksi fitur information gain untuk melakukan klasifikasi data penyakit tiroid?
2. Berapa hasil perbandingan kinerja metode LS-SVM dengan seleksi fitur Information Gain dan tanpa seleksi fitur Information Gain dalam melakukan klasifikasi data penyakit tiroid?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

1. Mengetahui kinerja algoritma LS-SVM dengan seleksi fitur untuk klasifikasi penyakit tiroid.
2. Mengetahui pengaruh SMOTE pada klasifikasi algoritma LS-SVM untuk mengatasi permasalahan data yang tidak seimbang

1.3.2 Manfaat

Manfaat yang ingin dicapai pada penelitian ini adalah:

1. Bagi Peneliti
Penelitian ini dapat digunakan oleh peneliti sebagai dasar penelitian lanjutan yang akan diteliti.
2. Bagi Pengguna
Penelitian ini dapat menjadi salah satu cara untuk mengetahui penyakit tiroid sedari awal.

1.4 Batasan Masalah

Batasan masalah ini dimaksudkan agar masalah yang diteliti jelas, terarah, dan tidak terlalu luas sehingga tidak terjadi salah pengertian terhadap masalah penelitian. Beberapa hal perlu diperhatikan, yaitu :

1. Studi kasus yang dilakukan untuk klasifikasi penyakit tiroid yang diperoleh dari *website resmi* UC Irvine Machine Learning pada link <https://archive.ics.uci.edu/dataset/102/thyroid+disease> pada tahun 2024 sejumlah 7530 data dengan jumlah fitur sebanyak 31.
2. Fitur-fitur yang digunakan untuk klasifikasi penyakit thyroid adalah *age*, *sex*, *query_on_thyroxine*, *on_thyroxine*, *on_antithyroid_meds*, *sick*, *pregnant*, *thyroid_surgery*, *I131_treatment*, *query_hypothyroid*,

query_hyperthyroid, lithium, goitre, tumor, hypopituitary, psych, TSH_measured, TSH, T3_measured, T3, FT4_measured, FT4, T4_measured, T4, FTI_measured, FTI, TBG_measured, TBG, referral_source, , patient_id dan *target*

3. Dalam dataset *target* yang digunakan terdapat tiga kelas yaitu normal, hyperthyroid (adanya kekurangan kerja hormon tiroid pada tubuh), dan hypothyroid (adanya kelebihan kerja hormon tiroid pada tubuh).
4. Data pada kelas normal sebanyak 6755 data, kelas *hyperthyroid* sebanyak 190 data, dan *hypothyroid* sebanyak 585 data.
5. Terdapat jumlah kelas yang tidak seimbang sehingga dilakukan tahapan imbalanced data (SMOTE)
6. Metode yang digunakan untuk tahap klasifikasi penyakit tiroid adalah LS-SVM dengan parameter kernel RBF dan kernel Linear.
7. Dalam menentukan parameter terbaik menggunakan metode grid search pada algoritma LS-SVM
8. Evaluasi kinerja yang digunakan menggunakan confusion matrix yaitu akurasi, *precision*, *recall*, dan f1-score. Indikator parameter yang digunakan menggunakan parameter kernel Linear $C=0.1, 10, 100$ dan parameter kernel RBF dengan $\text{Gamma} = 1, 100, 1000$ dan $C=0.1, 10, 100$

1.5 Sistematika Pembahasan

Sistematika penelitian proposal tugas akhir ini terdiri dari tiga bab yaitu

Bab I Pendahuluan

Berisikan latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, dan sistematika penulisan.

Bab II Kajian Pustaka

Berisikan teori-teori serta definisi-definisi dari seleksi fitur *Information Gain* metode sampling SMOTE, *Support Vector Machine*, dan *Least square- Support Vector Machine* sebagai dasar acuan dalam penelitian yang menjadi dasar perhitungan untuk mengkaji bab pembahasan

Bab III Metode Usulan

Berisikan tentang alur penelitian, sumber data, langkah-langkah pengolahan data, dan skenario pengujian.

Bab IV Hasil dan Pembahasan

Berisikan hasil implementasi dari arsitektur sistem dan hasil pengujian yang sesuai dengan skenario pengujian pada Bab III

Bab V Penutup

Berisikan mengenai kesimpulan dan saran yang didapatkan dari hasil pengujian yang telah dilakukan.