BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes Mellitus Tipe 2 (DMT2) merupakan penyakit metabolik kronis yang ditandai oleh hiperglikemia akibat gangguan kerja insulin, gangguan sekresi insulin, atau kombinasi keduanya. DMT2 menjadi perhatian karena dapat berkembang secara bertahap dan berkaitan dengan berbagai komplikasi jangka panjang, termasuk gangguan kardiovaskular, ginjal, saraf, dan mata. Kondisi tersebut menunjukkan pentingnya pengembangan pendekatan analisis yang mampu mendukung pemahaman faktor risiko penyakit secara lebih komprehensif (American Diabetes Association Professional Practice Committee, 2025).

Selain faktor lingkungan dan gaya hidup, DMT2 memiliki komponen genetik yang kompleks. Studi genomik berskala besar menunjukkan bahwa kerentanan terhadap DMT2 bersifat poligenik, mencakup banyak lokus pada populasi dengan latar leluhur yang beragam, dan juga dapat dipengaruhi oleh varian langka dengan efek yang lebih besar (Mahajan et al., 2022; Suzuki et al., 2024; Huerta-Chagoya et al., 2024). Oleh karena itu, informasi yang tersimpan dalam susunan basa DNA dapat dimanfaatkan sebagai salah satu sumber data untuk mempelajari pola yang berkaitan dengan kelas DMT2 dan non-diabetes, dengan tetap membatasi interpretasi pada data yang dianalisis dan tidak langsung menyamakannya dengan diagnosis pasien.

Susunan informasi DNA berbentuk data simbolik yang tersusun dari nukleotida A, T, G, dan C. Data tersebut tidak dapat langsung digunakan oleh sebagian besar algoritma klasifikasi sehingga perlu direpresentasikan menjadi fitur numerik. Salah satu pendekatan yang banyak digunakan dalam bioinformatika adalah pemotongan susunan nukleotida menjadi token *k-mer*. Nilai k menentukan panjang pola lokal

yang direpresentasikan; perubahan nilai k dapat mengubah jumlah serta kekhususan fitur yang dihasilkan. Setelah token terbentuk, pembobotan *Term Frequency–Inverse Document Frequency* (TF-IDF) dapat digunakan untuk mengubah kemunculan token menjadi vektor numerik (Moeckel et al., 2024; Salloum et al., 2025). Perkembangan terbaru juga menunjukkan bahwa susunan DNA dapat direpresentasikan menggunakan model fondasi genomik dan pemodelan konteks panjang, seperti Nucleotide Transformer dan HyenaDNA (Nguyen et al., 2023; Dalla-Torre et al., 2025). Penelitian ini tetap menggunakan *k-mer*–TF-IDF karena representasinya lebih langsung untuk dianalisis dan sesuai dengan desain perbandingan NuSVC–XGBoost yang menjadi fokus penelitian.

Penelitian ini menggunakan dua algoritma *supervised machine learning*, yaitu *Nu-Support Vector Classification* (NuSVC) dan *Extreme Gradient Boosting* (XGBoost). NuSVC dipilih karena merupakan varian SVM yang dirancang untuk klasifikasi berbasis margin dan menyediakan parameter ν untuk mengendalikan hubungan antara kesalahan pelatihan dan jumlah *support vector*; karakteristik tersebut relevan untuk fitur berdimensi tinggi seperti representasi *k-mer*–TF-IDF (Salloum et al., 2025). XGBoost dipilih sebagai pembanding karena menggunakan *gradient-boosted decision trees*, memiliki regularisasi, serta mampu memodelkan hubungan nonlinier dan interaksi antarfitur secara bertahap. Penggunaan XGBoost pada penelitian diabetes dan penilaian risiko kesehatan terkini juga dilaporkan oleh Ganie et al. (2023), Huang et al. (2024), dan Li et al. (2024), sedangkan penerapannya secara langsung bersama NuSVC pada klasifikasi susunan DNA dilaporkan oleh Salloum et al. (2025). Dengan menggunakan dua pendekatan yang berbeda, penelitian dapat menilai apakah model berbasis margin atau model berbasis *boosting* memberikan kinerja yang lebih baik pada representasi data yang sama.

Penelitian Salloum et al. (2025) secara khusus membandingkan NuSVC dan XGBoost untuk klasifikasi susunan DNA yang berkaitan dengan Diabetes Mellitus. Penelitian tersebut menggunakan representasi berbasis *k-mer* dan TF-IDF serta

melaporkan bahwa XGBoost memberikan kinerja lebih tinggi daripada NuSVC pada data yang mereka gunakan. Hasil tersebut menjadi dasar metodologis penelitian ini, namun penelitian sekarang tidak hanya membandingkan dua algoritma. Nilai *k-mer* divariasikan menjadi $k = 3$, $k = 4$, dan $k = 5$ untuk menilai pengaruh tingkat granularitas representasi terhadap kinerja kedua model.

Perbandingan model dilakukan menggunakan pemisahan data latih dan data uji, 10-*fold cross-validation* pada data latih, serta beberapa metrik evaluasi, yaitu *accuracy*, *precision*, sensitivitas, spesifisitas, *F1-score*, ROC-AUC, dan *log loss*. TF-IDF disesuaikan hanya pada bagian pelatihan dan SMOTE diterapkan hanya pada data pelatihan di setiap *fold*. Rancangan ini digunakan agar informasi dari data validasi atau data uji tidak ikut memengaruhi pembentukan fitur, penyeimbangan kelas, maupun pemilihan model (Moscovich & Rosset, 2022; Demircioglu, 2024).

Keterkaitan penelitian dengan Fisika Medis terletak pada penggunaan pendekatan kuantitatif dan komputasional untuk menganalisis informasi biologis yang berkaitan dengan kesehatan. Susunan informasi DNA terlebih dahulu ditransformasikan ke ruang fitur matematis, kemudian hubungan antarfitur dianalisis menggunakan algoritma klasifikasi, optimasi, probabilitas, serta metrik statistik. Pendekatan komputasional yang mengintegrasikan informasi genetik dengan model *machine learning* untuk penilaian risiko DMT2 juga telah digunakan pada penelitian kesehatan berbasis data modern (Huang et al., 2024). Dengan demikian, kontribusi fisika dalam penelitian ini berada pada aspek pemodelan, pengolahan data, dan evaluasi kuantitatif sistem biologis, bukan pada eksperimen laboratorium atau penetapan diagnosis klinis.

Berdasarkan uraian tersebut, penelitian ini berfokus pada klasifikasi susunan informasi DNA pada penyakit Diabetes Mellitus Tipe 2 menggunakan variasi *k-mer* untuk membandingkan metode NuSVC dan XGBoost. Variasi k digunakan untuk membentuk representasi fitur dengan tingkat detail yang berbeda, sedangkan perbandingan kedua metode dilakukan untuk menentukan konfigurasi yang memberikan

kinerja klasifikasi terbaik pada kumpulan data penelitian.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana pengaruh variasi *k-mer* dengan nilai $k = 3$, $k = 4$, dan $k = 5$ terhadap pembentukan representasi fitur TF-IDF dari susunan informasi DNA kelas DMT2 dan NONDM?
2. Bagaimana kinerja metode NuSVC dan XGBoost dalam mengklasifikasikan susunan informasi DNA kelas DMT2 dan NONDM pada setiap variasi nilai k berdasarkan metrik evaluasi yang digunakan?
3. Metode dan nilai k manakah yang memberikan kinerja klasifikasi terbaik berdasarkan 10-*fold cross-validation* pada data latih dan evaluasi akhir pada data uji?

1.3 Batasan Masalah

1. Data yang dianalisis berupa entri susunan informasi DNA dalam format FASTA yang dikelompokkan ke dalam kelas DMT2 dan NONDM berdasarkan label pada kumpulan data penelitian.
2. Variasi representasi dibatasi pada *k-mer* dengan $k = 3$, $k = 4$, dan $k = 5$, kemudian dibobot menggunakan TF-IDF.
3. Algoritma klasifikasi yang dibandingkan dibatasi pada NuSVC dan XGBoost dengan konfigurasi parameter sebagaimana dijelaskan pada metodologi penelitian.

4. Penanganan ketidakseimbangan kelas menggunakan SMOTE hanya pada bagian pelatihan, sedangkan data validasi dan data uji tidak mengalami proses *oversampling*.

5. Kinerja model dievaluasi menggunakan *accuracy*, *precision*, sensitivitas, spesifisitas, *F1-score*, ROC-AUC, *log loss*, dan *confusion matrix*.

6. Hasil penelitian dibatasi pada klasifikasi entri susunan informasi DNA pada kumpulan data yang digunakan dan tidak dimaksudkan sebagai diagnosis klinis pasien maupun bukti biomarker genetik tanpa validasi biologis lebih lanjut.

1.4 Tujuan Penelitian

Tujuan penelitian ini disusun secara langsung untuk menjawab rumusan masalah, yaitu:

1. Menganalisis pengaruh variasi *k-mer* dengan nilai $k = 3$, $k = 4$, dan $k = 5$ terhadap pembentukan representasi fitur TF-IDF dari susunan informasi DNA kelas DMT2 dan NONDM.

2. Membangun dan mengevaluasi model klasifikasi NuSVC dan XGBoost pada setiap variasi nilai k menggunakan metrik evaluasi yang telah ditetapkan.

3. Membandingkan kinerja NuSVC dan XGBoost serta menentukan metode dan nilai k yang memberikan kinerja klasifikasi terbaik berdasarkan validasi silang pada data latih dan evaluasi akhir pada data uji.

1.5 Manfaat Penelitian

1. Manfaat akademik menambah referensi penelitian di bidang Fisika Medis komputasional dan bioinformatika.

2. Manfaat praktis memberikan dasar metodologis bagi pengembangan analisis klasifikasi data genetik terkait Diabetes Mellitus, yang pada tahap selanjutnya masih memerlukan validasi biologis dan klinis sebelum dapat digunakan sebagai sistem pendukung keputusan kesehatan.

3. Manfaat keilmuan mengintegrasikan konsep fisika, dan *machine learning* dalam analisis data medis.

1.6 Sistematika Penelitian

Skripsi ini disusun dalam lima bab dengan sistematika sebagai berikut:

1. **BAB I PENDAHULUAN**

   Bab ini memuat latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

2. **BAB II TINJAUAN PUSTAKA**

   Bab ini membahas Diabetes Melitus, peran faktor genetik, susunan informasi DNA, sumber data biologis, representasi *k-mer* dan TF-IDF, pembelajaran mesin, NuSVC, XGBoost, metrik evaluasi, serta penelitian terdahulu yang berkaitan dengan topik penelitian.

3. **BAB III METODOLOGI PENELITIAN**

   Bab ini menjelaskan desain penelitian, alat dan bahan, sumber data, pemeriksaan kualitas data, pembagian data, pembentukan fitur, penyeimbangan kelas, konfigurasi model, validasi silang, evaluasi data uji, dan penyimpanan hasil.

4. **BAB IV HASIL DAN PEMBAHASAN**

   Bab ini menyajikan hasil audit data, perubahan komposisi data, hasil validasi silang, pemilihan nilai k dan model, evaluasi pada data uji, analisis *confusion matrix*, ROC-AUC, *log loss*, serta tingkat kepentingan fitur XGBoost.

5. **BAB V KESIMPULAN DAN SARAN**

   Bab ini memuat kesimpulan yang menjawab rumusan masalah serta saran untuk pengembangan penelitian selanjutnya.