BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Machine learning merupakan pendekatan di dalam bidang AI yang memungkinkan komputer belajar dari data tanpa perlu program eksplisit. Data memiliki peran krusial dalam pengembangan dan peningkatan model *machine learning*. Proses pelatihan model dimulai dengan memberikan sejumlah besar data kepada model, yang memungkinkannya belajar dan mengidentifikasi pola untuk membuat prediksi yang akurat. Semakin banyak data yang diberikan, semakin baik model dapat belajar. Dengan demikian, data memiliki peran penting dalam seluruh siklus hidup model *machine learning*, dari pelatihan hingga pengujian dan peningkatan kinerja. Dalam proses pelatihan model *machine learning* sering kali data yang tersedia memiliki distribusi kelas yang tidak seimbang, dimana kelas minoritas memiliki jumlah data yang jauh lebih sedikit dibandingkan kelas mayoritas yang sering disebut dengan ketidakseimbangan data (Kurniawati, 2019). Ketidakseimbangan data dapat menimbulkan tantangan serius dalam konteks pembelajaran mesin. Salah satu masalah utamanya adalah kecenderungan model pembelajaran mesin untuk menjadi bias terhadap kelas yang lebih umum, karena jumlah sampel yang lebih banyak dapat mendominasi pengaruhnya. Sebagai akibatnya, akurasi model dapat menurun secara signifikan untuk kelas yang kurang umum (Engg & Engg, 2021).

Untuk mengatasi ini, penting untuk menemukan solusi yang efektif. Salah satu pendekatan yang dapat diambil adalah melalui *resampling* data. Dalam *resampling* data, terdapat dua metode utama, yaitu *oversampling* dan *undersampling*. Pemilihan antara *oversampling* dan *undersampling* dapat didasarkan pada jumlah data. Jika jumlah data yang tersedia terbatas, maka *oversampling* mungkin menjadi pilihan terbaik. Salah satu teknik *oversampling* yang sering digunakan untuk menangani ketidakseimbangan kelas dalam sebuah

dataset adalah SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE menciptakan sampel sintetis baru untuk kelas minoritas, membantu mencegah *overfitting* dan meningkatkan performa model. Perubahan dalam distribusi kelas setelah penerapan SMOTE menjadi subjek analisis yang penting, terutama dalam mengevaluasi pembentukan kelas menggunakan metrik seperti *silhouette score*. Klaster yang terbentuk dengan baik memiliki dampak positif terhadap akurasi *machine learning* karena memungkinkan identifikasi pola-pola yang lebih relevan dalam data.

Dalam perkembangan teknik SMOTE, terdapat beberapa modifikasi, seperti SMOTE-ENN dan *Borderline*-SMOTE, yang menitikberatkan pada fokus dan proses pembuatan sampel sintetis yang berbeda. Dengan demikian, ketiga metode, yaitu SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE, dapat menciptakan data sintetis baru yang memiliki perbedaan dalam karakteristik kelas-kelas pada data. Oleh karena itu, nilai *Silhouette Score* dapat menjadi fitur yang membedakan karakteristik hasil *resampling* dari ketiga teknik tersebut. Membandingkan nilai *Silhouette* dari ketiga metode tersebut dapat menggambarkan bagaimana kemampuan dari ketiga metode tersebut dalam membentuk data sintetis baru, apakah setiap metode membentuk kelas dengan kualitas yang sama atau berbeda menjadi topik yang menarik untuk dianalisis. Dengan menganalisis perbedaan nilai *Silhouette Score* antara ketiga metode, kita dapat menilai sejauh mana pembentukan kelas yang optimal terjadi, serta apakah ada perbedaan signifikan dalam kualitas pembentukan kelas antara ketiganya. Penulis menuangkan ide ini kedalam penelitian yang berjudul **"Analisis Karakteristik *Resampling* Smote, Smote-Enn, dan *Borderline*-Smote Berdasarkan Nilai *Silhouette Coefficient*".**

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, berikut identifikasi masalah untuk penelitian ini.

1. Masalah Ketidakseimbangan Kelas. Ketidakseimbangan kelas menciptakan tantangan serius dalam pembelajaran mesin. Dampak

utamanya adalah kecenderungan model menjadi bias terhadap kelas yang lebih umum karena jumlah sampel yang lebih besar dapat mendominasi pengaruhnya. Konsekuensinya, akurasi model dapat signifikan menurun untuk kelas yang kurang umum.

2. Masalah Perubahan Karakteristik Kelas. Penambahan data sintetis pada kelas-kelas minoritas melalui teknik SMOTE dapat mengakibatkan perubahan pada karakteristik kelas-kelas tersebut. Perubahan distribusi kelas ini berpotensi mempengaruhi kinerja model pembelajaran mesin yang dilatih pada dataset tersebut.
3. Masalah Perbandingan Metode *Resampling*. Tersedia berbagai metode *resampling* untuk mengatasi ketidakseimbangan kelas, masing-masing dengan kelebihan dan kekurangan. Oleh karena itu, penting untuk membandingkan kinerja metode *resampling* guna menentukan pendekatan yang paling efektif dalam suatu konteks tertentu.

1.3 Pembatasan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, berikut pembatasan masalah untuk penelitian ini.

1. Ruang Lingkup Metode *Resampling*. Penelitian ini akan membatasi analisis pada tiga metode *resampling* utama, yakni SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE. Metode *resampling* lainnya tidak dibahas secara mendalam.
2. Dataset Numerik Sintetis. Fokus utama analisis akan berpusat pada penggunaan dataset numerik yang merepresentasikan kondisi ketidakseimbangan kelas. Hasil penelitian mungkin tidak dapat secara langsung diterapkan pada dataset dunia nyata dengan karakteristik yang berbeda.
3. Evaluasi dengan *Silhouette Coefficient*. Evaluasi kualitas metode *resampling* akan difokuskan pada *Silhouette Coefficient*.
4. Tanpa Pembanding Langsung dengan Metode Lain. Penelitian ini tidak akan memasukkan pembanding langsung dengan metode lain di luar

ketiga metode *resampling* yang dipilih. Fokus tetap pada pemahaman karakteristik dan kontribusi ketiga metode *resampling* tersebut.

1.4 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, berikut Rumusan masalah untuk penelitian ini.

1. Bagaimana karakteristik data hasil *resampling* menggunakan metode SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE?
2. Bagaimana perbandingan nilai *Silhouette Coefficient* hasil *resampling* dari metode SMOTE-ENN, SMOTE-ENN, dan *Borderline*-SMOTE?
3. Bagaimana hubungan antara kualitas hasil *resampling* menggunakan metode SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE dengan akurasi *machine learning*?

1.5 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, berikut tujuan dari penelitian ini.

1. Menganalisis karakteristik data yang dihasilkan dari proses *resampling* menggunakan metode SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE untuk menyeimbangkan ketidakseimbangan kelas pada dataset.
2. Membandingkan nilai *Silhouette Coefficient* dari hasil *resampling* antara metode SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE untuk mengevaluasi kualitas pembentukan kelas baru.
3. Menginvestigasi hubungan antara kualitas hasil *resampling* menggunakan metode SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE dengan akurasi *machine learning*, untuk memahami sejauh mana kualitas *resampling* mempengaruhi kinerja model.

1.6 Manfaat Penelitian

Berdasarkan tujuan penelitian yang telah diuraikan sebelumnya, berikut manfaat dari penelitian ini.

1. Pengembangan Pemahaman.
    a. Menambah pemahaman dalam karakteristik data hasil *resampling* dengan menerapkan metode SMOTE, SMOTE-ENN, dan *Borderline*-SMOTE.
    b. Memberikan wawasan mendalam terhadap perbandingan nilai *Silhouette Coefficient* antara ketiga metode *resampling*.
2. Optimasi Model Pembelajaran Mesin.
    a. Membantu peneliti dan praktisi dalam memilih metode *resampling* yang paling sesuai untuk mengatasi ketidakseimbangan data.
    b. Mendukung optimalisasi kinerja model pembelajaran mesin khususnya dalam konteks ketidakseimbangan distribusi kelas.
3. Kontribusi pada Pengembangan Metode *Resampling*.
    a. Memberikan kontribusi pada pengembangan metode *resampling* yang efisien dan dapat diaplikasikan secara luas.
    b. Mengidentifikasi kelebihan dan kelemahan dari masing-masing metode *resampling*, memandu pengembangan teknik yang lebih canggih.
4. Pengembangan Ilmu Pengetahuan.
    a. Menambah literatur dalam bidang pengembangan model pembelajaran mesin dengan fokus pada ketidakseimbangan data.
    b. Memberikan dasar bagi penelitian lanjutan yang dapat mengeksplorasi lebih jauh aspek-aspek khusus dari metode *resampling*.
5. Peningkatan Implementasi Teknologi AI.

    Meningkatkan keberhasilan implementasi teknologi AI pada berbagai aplikasi dengan memastikan seimbangnya distribusi kelas dalam dataset. Penelitian ini diharapkan dapat memberikan manfaat praktis dan konseptual bagi perkembangan ilmu pengetahuan serta aplikasi teknologi kecerdasan buatan.