JAMI Jurnal Ahli Muda Indonesia

JAMI: Jurnal Ahli Muda Indonesia
Volume 5, Nomor 2, Desember 2024, Halaman 172-182

Perbandingan Skenario *Balancing Oversampling* dan *Undersampling* dalam Klasifikasi Resiko Kambuh Kanker *Tiroid* menggunakan Algoritma *SVM Linear*

Muhammad Faruqziddan[1], Ewanda Herdika Septa Aulia[2], Salsabilla Dini Azzahra[3], Prabowo Budi Utomo[4*]

[1,2,3]Program Studi Sistem Informasi, Universitas Nusantara PGRI, Kediri, Indonesia
[4] Program Studi Administrasi Server dan Jaringan Komputer, Akademi Komunitas Negeri Putra Sang Fajar, Blitar, Indonesia

faruqziddan@gmail.com[1], ewandaherdika@gmail.com[2], salsazhrra1122@gmail.com[3]
Correspondence: prabowo86@akb.ac.id [4]

ABSTRAK

Tujuan. Klasifikasi adalah proses penting dalam analisis data yang bertujuan untuk membagi objek ke dalam kategori tertentu berdasarkan karakteristik yang dimilikinya, namun salah satu tantangan utama dalam proses ini adalah ketika data yang digunakan tidak seimbang. Ketidakseimbangan dataset terjadi saat jumlah sampel dalam satu kelas jauh lebih besar dibandingkan kelas lainnya. Kondisi ini membuat model klasifikasi lebih cenderung mengenali kelas yang dominan, sementara kelas minoritas sering kali diabaikan, dalam dunia kesehatan, masalah ini menjadi sangat krusial karena akurasi prediksi bisa memengaruhi keputusan medis yang vital. Penelitian ini bertujuan untuk membandingkan tiga skenario dalam menangani ketidakseimbangan data pada klasifikasi risiko kambuhnya kanker tiroid menggunakan algoritma SVM Linear.
Material dan Metode. Penelitian ini menggunakan pendekatan metodologi *SEMMA* dalam proses eksplorasi, transformasi, pemodelan, dan evaluasi data, yang selanjutnya dilakukan penyeimbangan data menggunakan tiga skenario yaitu *non balance data, balance oversampling* dan *balance undersampling.* Hasil setiap skenario penyeimbangan data akan diklasifikasi menggunakan algoritma SVM Linear untuk diperoleh nilai akurasi risiko kambuhnya kanker tiroid.
Hasil. Hasil *classification report* menunjukkan bahwa model pada *non-balanced data* memiliki *accuracy* 88%, *recall* 85%, *precision* 86%, dan *f1-score* 86%, dengan performa yang dipengaruhi ketidakseimbangan data. Pada *balanced data* menggunakan *oversampling SMOTE*, semua metrik meningkat hingga 91%, menunjukkan bahwa *oversampling* efektif dalam menangani ketidakseimbangan. Sementara itu, *balanced data* dengan *undersampling* memberikan *accuracy* 89%, *recall* 88%, *precision* 89%, dan *f1-score* 88%, sedikit lebih rendah karena pengurangan data kelas mayoritas. *Oversampling* terbukti memberikan hasil terbaik dalam skenario ini
Kesimpulan. Hasil penelitian ini juga menunjukkan bahwa skenario dengan *oversampling* menggunakan *SMOTE* memberikan performa terbaik dibandingkan dua pendekatan lainnya. Model skenario *oversampling* mencapai akurasi hingga 91%, dengan presisi, recall, dan F1-score yang juga berada di angka 91%. Sebaliknya, model pada skenario *undersampling* meskipun memberikan hasil yang lebih baik dibandingkan *non-balanced data*, menunjukkan penurunan performa dengan akurasi sebesar 89%

Kata Kunci
Kanker Tiroid, *oversampling*, *undersampling*, SVM;

ABSTRACT

Backgrounds. *Classification is a critical process in data analysis that involves grouping objects into specific categories based on their characteristics. However, one of the primary challenges in this process arises when the data used is imbalanced. Dataset imbalance occurs when the number of samples in one class significantly exceeds that of another. This condition often causes classification models to favor the dominant class while neglecting the minority class. In the healthcare domain, this issue becomes particularly critical, as the accuracy of predictions can profoundly impact vital medical decisions. This study aims to compare three scenarios for addressing dataset imbalance in the classification of thyroid cancer recurrence risk using the Support Vector Machine (SVM) Linear algorithm.*
Methods. *This research adopts the SEMMA methodology to systematically explore, transform, model, and evaluate data. Subsequently, data balancing is applied through three scenarios: non-balanced data, balanced data with oversampling, and balanced data with undersampling. Each scenario's balanced dataset is classified using the SVM Linear algorithm to assess the accuracy of predicting thyroid cancer recurrence risk.*
Results. *The classification report reveals that the model using non-balanced data achieved an accuracy of 88%, with recall, precision, and F1-score values of 85%, 86%, and 86%, respectively. These metrics indicate that the model's performance was influenced by the dataset imbalance. In contrast, the balanced data scenario using SMOTE oversampling resulted in improved metrics across the board, with accuracy, recall, precision, and F1-score all reaching 91%. This demonstrates the effectiveness of oversampling in addressing class imbalance. Meanwhile, the balanced data scenario using undersampling achieved an accuracy of 89%, with recall, precision, and F1-score at 88%, 89%, and 88%, respectively. These results, while better than non-balanced data, were slightly lower due to the loss of information from reducing the majority class. Oversampling proved to deliver the best performance in this context.*
Conclusions. *This study highlights that the oversampling scenario using SMOTE achieved the best performance compared to the other two approaches. The oversampling model reached an accuracy of 91%, with precision, recall, and F1-score also at 91%. On the other hand, while the undersampling scenario performed better than the non-balanced data scenario, it showed a slight decline in performance, with an accuracy of 89%.*

Key Words
Tiroid Cancer, *oversampling*, *undersampling*, SVM*;*

Received: 24th November 2024
Accepted: 24th December 2024
Published: 31st December 2024 **Citation:** -

10.46510/jami.v5i2.320
ISSN 2722-4414 (p)/ 2722-4406 (e)

I. PENDAHULUAN

Kanker Tiroid merupakan salah satu jenis kanker yang mempengaruhi kelenjar tiroid yang terletak di leher, tepatnya di bagian depan leher sedikit dibawah jakun dan memiliki bentuk seperti kupu-kupu, organ ini juga memiliki dua lobus yang melilit batang tenggorokan dan terkadang dihubungkan oleh bagian tengah yang disebut *isthmus*(Shalih et al., 2023). Prevalensi kanker tiroid secara global berkisar antara 0,85% hingga 2,5%, dengan rasio kejadian 1:3 antara pria dan wanita, di mana wanita lebih rentan terkena kanker ini, terutama dalam rentang usia 20 hingga 50 tahun (Nur et al., 2023). Meskipun memiliki tingkat kematian yang tergolong rendah dibandingkan dengan jenis kanker lainnya, namun mengetahui apakah pasien memiliki risiko kambuh kanker tiroid menjadi salah satu permasalahan yang perlu dianalisa mengingat fungsi kelenjar tiroid dalam meningkatkan metabolisme kalori, mengubah makanan menjadi energi, dan mengatur detak jantung. Deteksi dini terhadap risiko kekambuhan sangat penting untuk menentukan pendekatan penanganan yang tepat bagi pasien (Borzooei & Tarokhian, 2023).

Salah satu metode yang telah digunakan untuk mengklasifikasikan risiko kambuhnya kanker tiroid adalah algoritma Random Forest dengan teknik balancing data oversampling. Penelitian sebelumnya menunjukkan hasil yang cukup baik, dengan akurasi mencapai 97,5% dalam memprediksi risiko kambuhnya kanker tiroid (Faruqziddan et al., 2024). Meskipun demikian, penelitian tersebut belum mengeksplorasi metode balancing lain, seperti undersampling, maupun algoritma klasifikasi lain seperti Support Vector Machine (SVM). Penelitian ini memberikan peluang untuk memperluas wawasan terkait bagaimana teknik balancing yang berbeda dapat memengaruhi performa model klasifikasi.

Klasifikasi sendiri adalah proses pengelompokan data ke dalam kategori tertentu berdasarkan atribut dari dataset. Tujuan utama proses ini adalah untuk membangun model atau aturan yang dapat memprediksi kelas atau label data baru berdasarkan informasi yang tersedia (Siboro et al., 2024). Dalam implementasinya terdapat beberapa algoritma klasifikasi yang telah dikembangkan seperti *Decision/classification trees*, *Naïve Bayes classifiers*, *Neural networks*, *k-nearest neighbor* dan *Support vector machines (SVM)* (Annur, 2018), yang seiring perkembangan teknologi, algoritma-algoritma ini terus disempurnakan, termasuk dalam bentuk algoritma khusus seperti Long-Short Term Memory (LSTM) yang berbasis Recurrent Neural Network, yang dirancang untuk menangani dataset berskala besar dan bersifat dinamis (Budi Utomo et al., 2024).

Namun, salah satu tantangan umum dalam klasifikasi adalah ketidakseimbangan kelas. Ketidakseimbangan kelas terjadi saat jumlah sampel dalam satu kelas jauh lebih besar dibandingkan kelas lainnya (Nurhopipah & Magnolia, 2023). Ketidakseimbangan ini menyebabkan adanya kecenderungan model untuk bias terhadap kelas mayoritas, sehingga performa klasifikasi pada kelas minoritas menjadi kurang optimal (Syahwaluddin & Alita, 2024). Dalam kasus kanker tiroid, ketidakseimbangan kelas dalam dataset menjadi masalah yang tidak bisa diabaikan. Model yang bias terhadap kelas mayoritas berpotensi mengabaikan pasien dengan risiko kambuh tinggi, padahal kelompok ini sering kali membutuhkan perhatian medis yang lebih intensif. Oleh sebab itu, penting untuk mencari solusi yang mampu memastikan model klasifikasi bekerja secara akurat, tanpa mengesampingkan pentingnya deteksi kelas minoritas (Borzooei et al., 2024).

Untuk menghadapi tantangan tersebut, berbagai teknik *balancing* data telah dikembangkan. Dua pendekatan yang paling umum adalah *oversampling* dan *undersampling*. *Oversampling* merupakan teknik *balancing data* dengan cara meningkatkan jumlah data pada kelas minoritas agar mendekati atau sama dengan jumlah data kelas mayoritas, baik dengan mereplikasi data yang ada maupun dengan menciptakan data baru secara sintetis (Hamami & Dahlan, 2022). Salah satu metode yang sering digunakan dalam teknik *Oversampling* adalah *Synthetic Minority Over-sampling Technique* (SMOTE)**,** yang menciptakan data sintetis berdasarkan pola data minoritas yang ada. Sedangkan *undersampling* merupakan teknik *balancing data* dengan cara mengurangi bagian-bagian pada kelas mayoritas dengan menghapus sebagian data, sehingga proporsi antara kedua kelas menjadi seimbang. Meskipun efektif, metode ini memiliki risiko kehilangan informasi yang penting pada kelas mayoritas (Indrawati, 2021).

Dalam penelitian ini, ketidakseimbangan kelas yang diperoleh dalam dataset *Differentiated Thyroid Cancer Recurrence* yang diterbitkan oleh *UCI Machine Learning Repositoy* digunakan sebagai studi kasus*,* dimana dataset ini memiliki ketidakseimbangan kelas yang signifikan, dengan jumlah pasien kambuh jauh lebih sedikit dibandingkan pasien yang tidak kambuh. Maka untuk mengatasi hal tersebut, dalam penelitian ini akan dibandingkan tiga skenario dalam menangani ketidakseimbangan kelas pada klasifikasi, yaitu: (1) menggunakan dataset asli tanpa *balancing*, (2) menerapkan teknik *SMOTE* untuk *oversampling*, dan (3) menggunakan *Random Under-Sampling* (RUS) untuk *undersampling.* Secara umum, tujuan pemilihan tiga skenario ini adalah untuk skenario pertama dimaksudkan dalam mengevaluasi dampak langsung dari

ketidakseimbangan data terhadap performa algoritma klasifikasi, skenario kedua bertujuan untuk melihat sejauh mana teknik SMOTE mampu meningkatkan representasi kelas minoritas melalui data sintetis serta pengaruhnya terhadap dalam peningkatan performa akurasi klasifikasi, dan skenario ketiga mengevaluasi efektivitas RUS dalam mengurangi bias kelas mayoritas dengan cara menghapus sebagian data. Pemilihan tiga skenario secara tidak langsung akan berimbas dalam kesimbangan kelas menunjang performa akurasi yang diperoleh dari proses klasifikasi yang dilakukan, disamping nanti akan dilakukan evaluasi untuk serta perbandingan untuk setiap performa akurasi yang diperoleh, sehingga dapat memberikan wawasan yang lebih mendalam mengenai cara terbaik untuk menangani ketidakseimbangan data dalam klasifikasi risiko kambuhnya kanker tiroid.

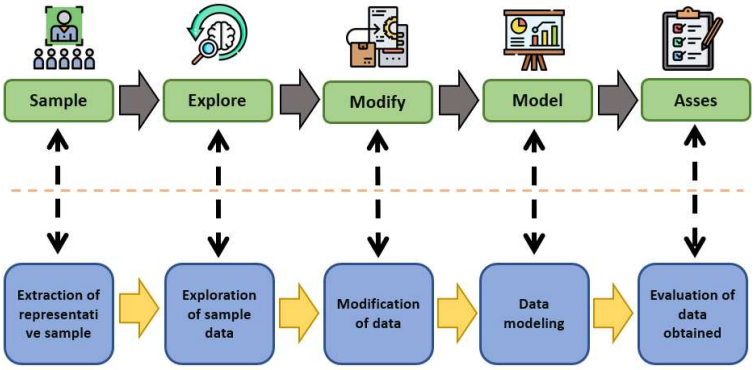
Dalam penelitian ini, algoritma *Support Vector Machine* (SVM) Linear dipilih sebagai model klasifikasi. SVM adalah algoritma yang bekerja dengan mencari *hyperplane* terbaik yang memisahkan dua kelas data, dengan memaksimalkan margin antara kelas-kelas tersebut. Untuk kasus dataset dengan pola *linear separable*, SVM Linear menawarkan solusi yang sederhana namun sangat efektif (Sabatini & Itan, 2024). Algoritma ini juga memungkinkan penerapan *kernel trick* untuk menangani data *non-linear* dalam ruang berdimensi tinggi. Pemilihan SVM Linear dalam penelitian ini didasarkan pada keunggulannya dalam menangani dataset dengan batasan kelas yang jelas dan untuk mempermudah analisis performa masing-masing skenario *balancing*.

Penelitian ini tidak hanya akan memberikan kontribusi terhadap pengembangan teknik klasifikasi pada dataset yang tidak seimbang, tetapi juga akan menghasilkan wawasan yang bermanfaat dalam mendukung pengambilan keputusan klinis terkait risiko kambuhnya kanker tiroid. Dengan evaluasi yang komprehensif, penelitian ini diharapkan dapat menentukan teknik balancing dan algoritma klasifikasi yang paling efektif, sehingga dapat digunakan sebagai referensi dalam penelitian selanjutnya di bidang yang sama.

II. MATERIAL DAN METODE

Penelitian ini menggunakan pendekatan metodologi *Data Mining Process* (SEMMA**)**, dimana SEMMA (*Sample, Explore, Modify, Model, Assess*) merupakan metode data mining yang dapat memberikan solusi terhadap masalah dan tujuan bisnis (Fadhilla Ramdhania et al., 2024). Metodologi dipandang sangat sesuai karena memungkinkan pemahaman yang mendalam terhadap data serta pengembangan model yang optimal. Penjelasan dari setiap tahapan dalam SEMMA yang dilaksanakan dalam penelitian adalah sebagai berikut:

A. SEMMA (Suwitono & Kaunang, 2022)



Gambar 1. Tahapan dalam Metode SEMMA (Andrade-Arenas et al., 2024)

1. Tahap *Sample*: Pengumpulan dan Pemilihan Dataset
   Tahap pertama adalah proses pengumpulan dan seleksi dataset yang akan digunakan dalam penelitian. Dataset *Differentiated Thyroid Cancer Recurrence* dari *UCI Machine Learning Repository* dipilih karena mencerminkan masalah ketidakseimbangan kelas yang signifikan, dengan distribusi kelas pasien kambuh sebanyak 28,2% dan pasien tidak kambuh sebanyak 71,8%. Dataset ini awalnya terdiri dari 383 data pasien dengan 16 fitur dan 1 target variabel. Luaran dari tahap ini adalah dataset mentah yang telah diperiksa untuk memastikan kelengkapan atribut dan tidak adanya data duplikat yang dapat memengaruhi kualitas analisis. Dataset ini kemudian siap untuk dianalisis lebih lanjut pada tahap berikutnya.
2. Tahap *Explore*: Analisa Karakteristik Dataset

Data yang terkumpul akan dianalisis untuk memahami karakteristik dan pola strukturnya. Dalam tahap ini bertujuan untuk memahami karakteristik dataset, termasuk pola distribusi data, hubungan antarvariabel, dan keberadaan *outlier* atau *missing values*. Analisis ini bertujuan untuk memperoleh wawasan mendalam mengenai dataset *Differentiated Thyroid Cancer Recurrence*, sehingga langkah-langkah pemrosesan selanjutnya dapat ditentukan dengan lebih tepat. Proses ini melibatkan beberapa langkah penting, yaitu:

- Analisis Distribusi Kelas: Dilakukan untuk mengukur tingkat ketidakseimbangan dan memahami proporsi data antara kelas mayoritas dan minoritas.
- Statistik Deskriptif: Menghitung nilai mean, median, standar deviasi, dan distribusi dari variabel numerik untuk memahami pola data.
- Visualisasi Data: Diagram batang dan histogram digunakan untuk memetakan distribusi data dan mengidentifikasi potensi outlier.

Luaran dari tahap ini adalah wawasan mendalam mengenai karakteristik dataset termasuk identifikasi ketidakseimbangan kelas, hubungan antar fitur, dan potensi masalah seperti outlier atau missing values yang perlu diatasi pada tahap selanjutnya, yang menjadi dasar untuk langkah transformasi dan pemrosesan data selanjutnya.

3. Tahap *Modify*: Transformasi dan *Balancing* Dataset

   Tahap ini melibatkan serangkaian langkah untuk memodifikasi dataset agar siap digunakan dalam pemodelan. Tahapan yang dilakukan meliputi:

   - *Preprocessing*: Pada langkah ini, data duplikat dihapus untuk menjaga integritas dataset. Sebanyak 19 data duplikat ditemukan dan dihapus, sehingga dataset final terdiri dari 364 data.
   - *Transformation*: Variabel kategorikal dikonversi menjadi format numerik menggunakan teknik *Label Encoding*. Proses ini penting untuk memastikan kompatibilitas data dengan algoritma SVM, yang hanya dapat mengolah input dalam format numerik.
   - *Balancing* Data: data yang telah diproses kemudian diseimbangkan menggunakan dua teknik balancing utama:
     1. *Synthetic Minority Over-sampling Technique* (SMOTE), Teknik *oversampling* ini menambahkan data *sintetis* untuk kelas minoritas berdasarkan pola data yang ada. Setelah penerapan SMOTE, jumlah total data meningkat menjadi 512 dengan distribusi seimbang (256 kelas minoritas, 256 kelas mayoritas).
     2. *Random Under-Sampling* (RUS), Teknik ini mengurangi data kelas mayoritas secara acak untuk menciptakan keseimbangan. Setelah RUS, dataset berkurang menjadi 216 data, dengan masing-masing kelas terdiri dari 108 data.

   Kedua teknik ini digunakan dalam skenario yang berbeda untuk membandingkan performa model pada data yang diseimbangkan. Luaran dari tahap ini adalah tiga dataset terpisah: dataset asli tanpa *balancing*, dataset *oversampling* menggunakan SMOTE, dan dataset *undersampling* menggunakan RUS.

4. Tahap *Model*: Pengembangan Model Klasifikasi

   Pada tahap ini, dataset yang telah dimodifikasi dibagi menjadi data latih dan data uji dengan rasio 80:20. Proses pemodelan dilakukan menggunakan algoritma SVM Linear, yang dipilih karena kemampuannya menangani data dengan pola *linear separable*. Proses ini melibatkan:

   - Pembagian dataset menjadi data latih (80%) dan data uji (20%).
   - Pelatihan model menggunakan data latih untuk menghasilkan hyperplane terbaik yang memisahkan kelas minoritas dan mayoritas.
   - Pengujian model pada data uji untuk mengevaluasi performanya.

   Luaran dari tahap ini adalah tiga model klasifikasi berbasis SVM Linear, masing-masing dilatih pada dataset *non-balanced*, *oversampling*, dan *undersampling*. Model ini siap untuk dievaluasi pada tahap berikutnya.

5. Tahap *Assess*: Evaluasi Kinerja Model

   Evaluasi model dilakukan untuk menilai efektivitas teknik *balancing* data dalam meningkatkan performa klasifikasi risiko kambuhnya kanker tiroid. Evaluasi dilakukan menggunakan metrik berikut:

   - Akurasi, untuk mengukur persentase prediksi yang benar.
   - Precision, untuk mengevaluasi kemampuan model mengidentifikasi kelas minoritas dengan tepat.

- Recall, untuk mengukur sensitivitas model terhadap kelas minoritas.
- F1-Score, sebagai rata-rata harmonik antara precision dan recall.

Evaluasi dilakukan untuk setiap model dari tiga skenario *balancing* data, dan hasilnya dibandingkan untuk menentukan teknik *balancing* yang paling efektif.

B. Teknik *Balancing* Data

1. *Synthetic Minority Over-sampling Technique* (SMOTE)

SMOTE (*Synthetic Minority Over-sampling Technique)* merupakan teknik untuk menyeimbangkan distribusi data dengan meningkatkan jumlah data pada kelas minoritas hingga setara dengan kelas mayoritas (Kasanah et al., 2019). Peningkatan jumah data pada kelas minoritas dilakukan dengan cara membuat sampel sintetis baru diantara titik-titik minoritas yang ada dalam ruang fitur. Beberapa langkah yang dilakukan dalam teknik SMOTE diawali dengan menentukan jumlah tetangga terdekat yang akan digunakan untuk membuat sampel sintesis baru, setelahnya SMOTE memilih secara acak titik minoritas dari dataset dan menentukan tetangga terdekat untuk titik tersebut, kemudian baru membuat sampel sintesis baru diantara titik tersebut dengan menggabungkan atribut dari titik minoritas yang dipilih dengan atribut dari tetangga terdekat (Syahwaluddin & Alita, 2024). Namun, teknik ini memiliki risiko overfitting jika data sintetis terlalu mirip dengan data asli.

2. *Random Under Sampling (RUS)*

Undersampling merupakan metode *sampling* yang secara acak memilih sampel di kelas mayoritas dan menambahkannya ke kelas minoritas untuk membentuk sebuah dataset baru (Saputro & Rosiyadi, 2022). *Random under sampling* menggunakan pemilihan data secara acak dari kelas mayoritas untuk dihapus dari kumpulan data latih, dengan menggunakan *Random under sampling* maka data dari kelas mayoritas akan berkurang jumlahnya (Nurdin et al., 2018). Dalam pendekatan ini diterapkan dalam kumpulan data dengan kelas yang tidak seimbang dimana kelas minoritas cukup untuk pembuatan model, kekurangan datri teknik ini adalah data yang dihapus dari kelas mayoritas merupakan data acak sehingga ada kemungkinan data acak tersebut merupakan data yang berguna atau bahkan penting dalam pembangunan model klasifikasi (Imama Sabilla & Bella Vista, 2021). Luaran yang diperoleh dari teknik ini dimungkinkan dibuat jumlah data yang sama dari kelas mayoritas dan minoritas atau hanya mengurangi data mayoritas hingga jumlah tertentu.

C. *Support Vector Machine* (SVM)

Algoritma *Support Vector Machine* (SVM) merupakan metode klasifikasi dalam ranah *supervised learning* di data mining yang memiliki kemampuan untuk melakukan pembagian *linear* pada data input yang bersifat *nonlinear* dan mempunyai dimensi yang besar dengan memanfaatkan fungsi *kernel* (Azzahra et al., 2023). SVM menggunakan *hyperplane* secara optimal dalam mengklasifikasikan data menjadi kelompok data dalam ruang dimensi yang lebih tinggi dengan menggunakan jarak antara *hyperplane* dan data terdekat dari setiap kelas (Putri & Wijayanto, 2022). Permasalahan *nonlinear* yang sering dihadapai dalam penggunaan SVM dapat diatasi dengan memodifikasi *trick kernel* ke dalam SVM yang akan menjadi pemisah kelas atau *hyperplane* menjadi dua kelas didalam ruang vektor, dalam penelitian ini kernel yang akan digunakan adalah kernel *linear*, seperti yang ditunjukkan dalam persamaan berikut (Rahman Isnain et al., 2021):

Jenis Kernel	Model
Linear	$K(x \,.\, x') = x \,.\, x'$

Dengan ketentuan:

- $K(x \,.\, x')$ untuk Kernel fungsi yang digunakan untuk menghitung kesamaan antara dua vektor x dan x' dan ruang fitur.
- $x \cdot x$ untuk produk skalar atau dot product antara dua vektor x dan x'. Dalam bentuk ini, kernel linier hanya menghitung hasil kali elemen-elemen yang sesuai dari kedua vektor dan menjumlahkannya.

III. HASIL

A. *Sample* dan *Explore*

Tahap awal adalah pengambilan data yang akan digunakan untuk proses analisis. Data yang digunakan adalah *Differentiated Thyroid Cancer Recurrence* yang berasal dari *UCI repository*. Berikut ini adalah informasi terkait dengan dataset.:

Tabel 1. Informasi Dataset

Dataset Characteristics	*Tabular*
Subject Area	*Health and Medicine*
Associated Tasks	*Classification*
Feature Type	*Real, Categorical, Integer*
Instances	*383*
Features	*16*
Has missing value ?	*No*

Dataset terdiri dari 383 pasien, dengan rincian 108 pasien (28,2%) mengalami kambuh dan 275 pasien (71,8%) tidak mengalami kambuh. Dengan memanfaatkan ketidak seimbangan kelas inilah penelitian eksperimental ini dilakukan. Dataset memiliki 17 atribut, dengan 16 *feature* dan 1 target.

B. *Modify*

Pada tahap ini, dilakukan tiga tahapan utama, yaitu *preprocessing, transformation,* dan *balancing*. Tujuan utamanya adalah memastikan bahwa dataset tidak mengandung duplikat, *missing value*, *outlier*, atau data kotor, serta memastikan dataset dapat menjalankan tiga skenario yang telah direncanakan dengan penerapan *balancing* data.

1. *Preprocessing*

Pada tahap *preprocessing* hanya dilakukan penghapusan data duplikat, mengingat informasi dari *UCI Repository* menunjukkan bahwa dataset tersebut tidak memiliki *missing value*. *Feature* dari data berjumlah 15 merupakan kategorikal sehingga tidak memungkinkan adanya *outlier*, dan 1 *feature* berbentuk numerik yang adalah umur sehingga tidak diperlukan pengecekan *outlier*.

a. Data duplikat

Tahap ini dilakukan untuk memastikan tidak ada duplikasi dalam dataset, guna mempermudah proses pengolahan data. Setelah pengecekan, ditemukan 19 data duplikat yang kemudian dihapus. Dengan demikian, jumlah dataset yang tersisa adalah 364 dengan proses yang ditunjukan pada gambar 2.

```
duplicate_rows = dataset[dataset.duplicated()]
print("Jumlah baris duplikat:", len(duplicate_rows))

data = dataset.drop_duplicates()
print("Jumlah baris setelah menghapus duplikat:", len(data))
```

```
Jumlah baris duplikat: 19
Jumlah baris setelah menghapus duplikat: 364
```

Gambar 2. Penghapusan Duplikat

2. *Transformation*

Transformation yang dilakukan mencakup pengubahan 15 variabel kategorikal menjadi format numerik menggunakan fungsi *label encoder*. *Label encoder* digunakan untuk mengonversi data kategorikal ke dalam bentuk numerik, di mana setiap kategori dalam suatu fitur diberikan representasi berupa angka. Transformasi ini penting untuk memastikan data dapat digunakan dalam model *SVM* yang membutuhkan *input* berupa nilai numerik dengan hasil ditunjukan pada gambar 3.

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response
0	27	0	0	0	0	2	3	3	2	1	2	0	0	0	0	2
1	34	0	0	1	0	2	1	3	2	1	2	0	0	0	0	1
2	30	0	0	0	0	2	4	3	2	1	2	0	0	0	0	1
3	62	0	0	0	0	2	4	3	2	1	2	0	0	0	0	1
4	62	0	0	0	0	2	1	3	2	0	2	0	0	0	0	1

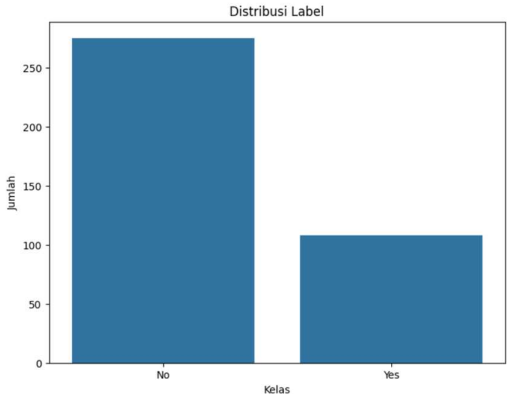
Gambar 3. Hasil Transformation

3. *Balancing*

Kategori dataset dibagi menjadi tiga kategori, yaitu *Non-Balanced Data, Balanced Data Oversampling,* dan *Balanced Data Undersampling*. Mekanisme *balancing* dilakukan melalui dua pendekatan yaitu *oversampling* menggunakan metode *SMOTE*, dan *undersampling* menggunakan metode *Random Undersampler*.

a. Non balance data

Pada skenario pertama dataset yang digunakan adalah dataset asli tanpa ada penyesuaian. Pada tahap ini, kelas *No* (tidak kambuh) jauh lebih banyak dibandingkan dengan kelas *Yes* (kambuh). Dengan distribusi 28,2% mengalami kambuh dan 71,8% tidak kambuh dengan visualisasi ditunjukan pada gambar 4.



Gambar 4. Distribusi label *non balance data*

b. Balance data Oversampling

Pada skenario kedua, dilakukan *oversampling* menggunakan teknik *SMOTE*. Langkah-langkah dalam penerapan *SMOTE* dijelaskan pada gambar 5.

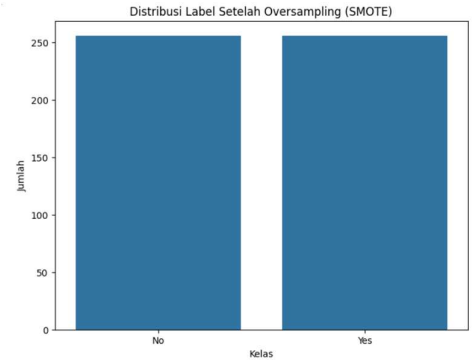
```
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_over_resampled, y_over_resampled = smote.fit_resample(X, y)
print("Jumlah data setelah oversampling:")
print(y_over_resampled.value_counts())
```

```
Jumlah data setelah oversampling:
Recurred
No     256
Yes    256
Name: count, dtype: int64
```

Gambar 5. *Script balancing data* dengan *SMOTE*

- Baris pertama mengimpor modul *SMOTE* dari pustaka *imbalanced-learn*.
- Membuat objek *SMOTE* dengan parameter *random_state=42*
- Baris ketiga menghasilkan data fitur dan *label* yang seimbang dengan membuat data sintetis pada kelas minoritas menggunakan *SMOTE*.
- Baris ke 4 dan 5 menampilkan jumlah tiap kelas yang sudah *dibalancing*.

Cara kerja *SMOTE* adalah dengan mengidentifikasi sampel dari kelas minoritas, kemudian menghitung jarak dengan kelas mayoritas. Berdasarkan jarak ini, *SMOTE* menciptakan sampel sintetis baru dengan menggabungkan sampel minoritas dan tetangga terdekatnya. Setelah penerapan *SMOTE* jumlah total data meningkat menjadi 512, dengan masing-masing kelas berjumlah 256. Visualisasi *label SMOTE* ditunjukan pada gambar 6.



Gambar 6. Distribusi *label balance data SMOTE*

c. *Balance data Undersampling*

Pada skenario ketiga, dilakukan *undersampling* dengan teknik *Random Under-sampling*. Langkah-langkah dalam penerapan *Random Under-sampling* dijelaskan pada gambar 7.

```
from imblearn.under_sampling import RandomUnderSampler
rus = RandomUnderSampler(random_state=42)
X_under_resampled, y_under_resampled = rus.fit_resample(X, y)
print("Jumlah data setelah undersampling:")
print(y_under_resampled.value_counts())
```

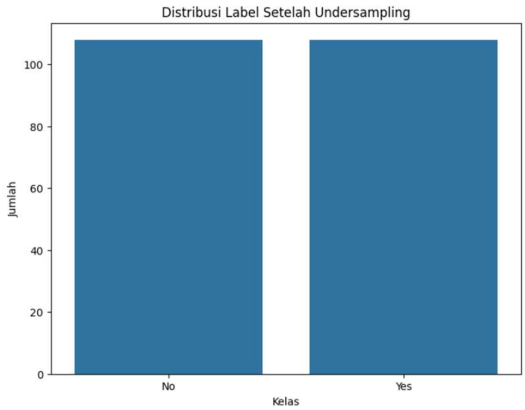
```
Jumlah data setelah undersampling:
Recurred
No     108
Yes    108
Name: count, dtype: int64
```

Loading

Gambar 7. *Script RandomUnderSampling*

- Baris pertama mengimpor modul *RandomUnderSampler* dari pustaka *imblearn.under_sampling.*
- Membuat objek *RandomUnderSampler* dengan parameter *random_state=42*
- Baris ketiga menghasilkan data fitur dan *label* yang seimbang dengan menghaous data dari kelas mayoritas menggunakan *RandomUnderSampler.*
- Baris ke 4 dan 5 menampilkan jumlah tiap kelas yang sudah *dibalancing.*

Proses ini dimulai dengan mengidentifikasi kelas mayoritas, kemudian secara acak memilih sampel dari kelas mayoritas untuk dihapus, sehingga proporsi antara kelas mayoritas dan minoritas menjadi seimbang. Setelah penerapan *undersampling*, jumlah data berkurang menjadi 216 dengan masing-masing kelas memiliki jumlah 108. Visualisasi label *Random Under-Sampling* ditunjukan pada gambar 3.7.



Gambar 8. Script RandomUnderSampling

C. *Model*

Model dibangun menggunakan algoritma *SVM* dengan *kernel linear* dengan rasio pembagian data training dan testing sebesar 80:20. Langkah-langkah Pembangunan model ditunjukan pada gambar 9.

```
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)
y_pred = svm_model.predict(X_test)
```

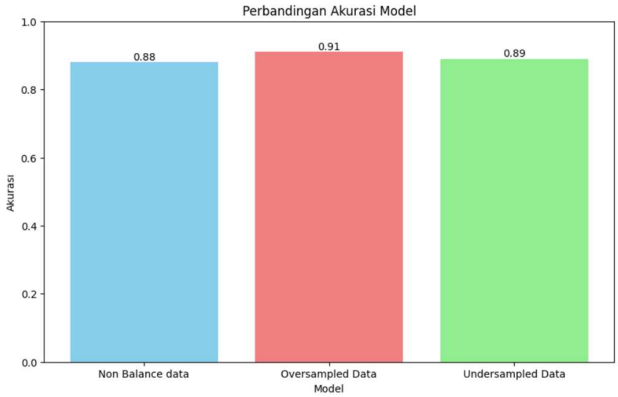
Gambar 9. *Script model SVM*

- Baris 1 - 3 mengimpor *library* dan modul yang diperlukan dalam pembuatan model.
- Baris 4 pembuatan data *training* dan *testing* dengan perbandingan 80:20.
- Baris 5 mendefinisikan model *SVM dengan kernel linear*.
- Baris 6 pelatihan model *SVM* dengan data *training*.
- Baris 7 prediksi data *testing* dengan model yang sudah di latih.

Pembuatan model tersebut dilakukan 3 kali sesuai dengan jumlah sekenario yang telah disiapkan.

D. *Assess*

Pada tahap ini, dilakukan evaluasi terhadap model *SVM linear* yang telah dibangun untuk menilai keakuratan dan kegunaannya dalam menyelesaikan tugas klasifikasi. Proses evaluasi melibatkan pengujian model pada data *testing* dan mengukur seberapa baik model dapat mengklasifikasikan data, Gambar 10 menunjukan visualisasi dari perbandingan 3 sekenario yang telah diterapkan.



Gambar 10. Visualisasi perbandingan *accuracy* model

Tabel 2. *Classification report*

Skenario	Jenis data	*Accuracy*	*Recall*	*Precision*	*F1 Score*
1	*Non balance data*	88%	85%	86%	86%
2	*Balance data oversampling*	**91%**	**91%**	**91%**	**91%**
3	*Balance data undersampling*	89%	88%	89%	88%

Hasil *classification report* menunjukkan bahwa model pada *non-balanced data* memiliki *accuracy* 88%, *recall* 85%, *precision* 86%, dan *f1-score* 86%, dengan performa yang dipengaruhi ketidakseimbangan data. Pada *balanced data* menggunakan *oversampling SMOTE*, semua metrik meningkat hingga 91%, menunjukkan bahwa *oversampling* efektif dalam menangani ketidakseimbangan. Sementara itu, *balanced data* dengan *undersampling* memberikan *accuracy* 89%, *recall* 88%, *precision* 89%, dan *f1-score* 88%, sedikit lebih rendah karena pengurangan data kelas mayoritas. *Oversampling* terbukti memberikan hasil terbaik dalam skenario ini.

IV. KESIMPULAN

Preprocessing data dan *transformation data* yang tepat sangat penting dalam meningkatkan kinerja model klasifikasi resiko kambung kanker tiroid. Dalam penelitian ini telah berhasil dibandingkan tiga skenario penanganan ketidakseimbangan data dalam klasifikassi risiko kambunhnya kanker Tiroid menggunakan Algoritma Support Vector Machine (SVM) Linear,

dengan menggunakan dataset **Differentiated Thyroid Cancer Recurrence** dari UCI Machine Learning Repository digunakan sebagai studi kasus dikarenakan nilai distribusi data yang tidak seimbang (28,2% kambuh dan 71,8% tidak kambuh). Melalui penerapan tiga skenario penanganan ketidakseimbangan data – *non-balanced data, balanced data* menggunakan *oversampling SMOTE* dan *balanced data* dengan *undersampling* serta mekanisme *Preprocessing data* dan *transformation data* dapat dilakukan evaluasi pengaruh teknik balancing terhadap performa model klasifikasi.

Hasil penelitian ini juga menunjukkan bahwa skenario dengan *oversampling* menggunakan *SMOTE* memberikan performa terbaik dibandingkan dua pendekatan lainnya. Model skenario *oversampling* mencapai akurasi hingga 91%, dengan presisi, recall, dan F1-score yang juga berada di angka 91%. Hal ini secara tidak langsung mempertegas bahwa penambahan representasi kelas minoritas melalui data *sintetis* cukup efektif untuk mengatasi bias terhadap kelas mayoritas tanpa perlu kehilangan informasi penting. Sebaliknya, model pada skenario *undersampling* meskipun memberikan hasil yang lebih baik dibandingkan *non-balanced data*, menunjukkan penurunan performa dengan akurasi sebesar 89%. Penurunan ini disebabkan oleh penghapusan data kelas mayoritas yang berpotensi mengurangi informasi signifikan.

Saran penelitian selanjutnya dapat dilakukan adalah dengan memperluas cakupan menggunakan algoritma *machine learning* lain dan membandingkan performanya dengan *SVM*. Selain itu, penggunaan teknik penyeimbangan data yang lebih kompleks, dapat dieksplorasi untuk hasil yang lebih optimal.

V. DAFTAR PUSTAKA

Andrade-Arenas, L., Rubio-Paucar, I., & Yactayo-Arias, C. (2024). Predictive models in Alzheimer's disease: an evaluation based on data mining techniques. *International Journal of Electrical and Computer Engineering*, *14*(3), 2988–3002. https://doi.org/10.11591/ijece.v14i3.pp2988-3002

Annur, H. (2018). KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES. *ILKOM*, *10*(2). https://doi.org/10.33096/ilkom.v10i2.303.160-165

Azzahra, S. P., Apriyanto, Y. A., & Wijaya, A. (2023). ANALISIS SENTIMEN ULASAN APLIKASI DEEPL PADA GOOGLE PLAY DENGAN METODE SUPPORT VECTOR MACHINE (SVM). *Jurnal Sistem Informasi (JUSIN)*, *4*(2), 59–66. https://doi.org/https://doi.org/10.32546/jusin.v4i2.2368

Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhian, A. (2024). Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2095–2104. https://doi.org/10.1007/s00405-023-08299-w

Borzooei, S., & Tarokhian, A. (2023). *Differentiated Thyroid Cancer Recurrence [Dataset]*. UCI Machine Learning Repository. https://doi.org/10.24432/C5632J

Budi Utomo, P., Aswi Ramadhani, R., & Kurniawan, H. (2024). Deteksi Gerak Tangan sebagai Pengenal Bahasa Isyarat menggunakan Mediapipe dan Long-Short Term Memory. *Jurnal SIMETRIS*, *15*(1).

Fadhilla Ramdhania, K., Fitrianto Hidayat, D., & Salkiawati, R. (2024). Implementasi Metode Naïve Bayes dan Support Vector Machine (SVM) untuk Menganalisis Sentimen Pengguna Twitter terhadap Transjakarta. *JMPM: Jurnal Matematika Dan Pendidikan Matematika*, *9*(1), 1–14. https://doi.org/https://dx.doi.org/10.26594/jmpm.v9i1.4494.

Faruqziddan, M., Aulia, E. H. S., Azzahra, S. D., Ristyawan, A., & Daniati, E. (2024). Klasifikasi Risiko Kambuhnya Kanker Tiroid Menggunakan Algoritma Random Forest. *In Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, *8*(1), 63–74.

Hamami, F., & Dahlan, I. A. (2022). Klasifikasi Cuaca Provinsi Dki Jakarta Menggunakan Algoritma Random Forest Dengan Teknik Oversampling. *Jurnal Teknoinfo*, *16*(1), 87. https://doi.org/10.33365/jti.v16i1.1533

Imama Sabilla, W., & Bella Vista, C. (2021). Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan. *Jurnal Politeknik Caltex Riau*, *7*(2), 329–339. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjZh8GyzL-KAxWb1DgGHXKxLycQFnoECEQQAQ&url=https%3A%2F%2Fjurnal.pcr.ac.id%2Findex.php%2Fjkt%2Farticle%2Fview%2F5027%2F1747&usg=AOvVaw1mJCFXq3AeYh4QB1uOR0H-&opi=89978449

Indrawati, A. (2021). Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset. *JIKO (Jurnal Informatika Dan Komputer)*, *4*(1), 38–43. https://doi.org/10.33387/jiko.v4i1.2561

Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 196–201. https://doi.org/10.29207/resti.v3i2.945

Nur, A., Santosa, A., & Siti Komariyah, A. (2023). KARAKTERISTIK KANKER TIROID DI MALUKU UTARA TAHUN 2017-2020. *Jurnal Endurance : Kajian Ilmiah Problema Kesehatan*, *8*(2), 246–252. https://doi.org/10.22216/jen.v8i2.2161

Nurdin, Hamdhana, D., & Iqbal, M. (2018). Aplikasi Quick Count Pilkada Dengan Menggunakan Metode Random Sampling Berbasis Android. *E-Journal Techsi Teknik Informasi*, *10*(1), 141–154.

Nurhopipah, A., & Magnolia, C. (2023). PERBANDINGAN METODE RESAMPLING PADA IMBALANCED DATASET UNTUK KLASIFIKASI KOMENTAR PROGRAM MBKM. *Jurnal Publikasi Ilmu Komputer Dan Multimedia* , *2*(1), 9–22. https://doi.org/https://doi.org/10.55606/jupikom.v2i1.862

Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika : Jurnal Sistem Komputer*, *11*(1), 59–66. https://doi.org/10.34010/komputika.v11i1.4350

Rahman Isnain, A., Indra Sakti, A., Alita, D., & Satya Marga, N. (2021). SENTIMEN ANALISIS PUBLIK TERHADAP KEBIJAKAN LOCKDOWN PEMERINTAH JAKARTA MENGGUNAKAN ALGORITMA SVM. *JDMSI*, *2*(1), 31–37. https://t.co/NfhnfMjtXw

Sabatini, T., & Itan, V. (2024). Implementasi Support Vector Machine untuk Klasifikasi Kasus Monkeypox: Pendekatan Oversampling dan Undersampling untuk Mengatasi Ketidakseimbangan Kelas. *Journal of Digital Ecosystem for Natural Sustainability (JoDENS)*, *4*(1), 38–43. https://www.kaggle.com/datasets/muhammad4hmed/monke

Saputro, E., & Rosiyadi, D. (2022). Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes. *Bianglala Informatika*, *10*(1), 42–47. https://doi.org/https://doi.org/10.31294/bi.v10i1.11739

Shalih, M. G., Utami, M. R., Adam, M. I., & Shadrina, J. A. (2023). Edukasi Hormon Tiroid dan Antitiroid Terhadap Penyakit Gondok di SMK Wirasaba Karawang. *Jurnal Dorkes (Dedikasi Olahraga Dan Kesehatan)*, *1*(2), 50–57.

Siboro, O., Pricilia Banjarnahor, Y., Gultom, A., Antonius Siagian, N., & Silitonga, P. D. (2024). Penanganan Data Ketidakseimbangan dalam Pendekatan SMOTE Guna Meningkatkan akurasi Algoritma K-NN. *SNISTIK : Seminar Nasional Inovasi Sains Teknologi Informasi Komputer*, *1*(Mei), 473–478. https://doi.org/https://doi.org/10.54367

Suwitono, Y. A., & Kaunang, F. J. (2022). Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras. *Jurnal Komtika (Komputasi Dan Informatika)*, *6*(2), 109–121. https://doi.org/10.31603/komtika.v6i2.8054

Syahwaluddin, R., & Alita, D. (2024). Penerapan Oversampling Pada Klasifikasi Ujaran Kebencian Menggunakan Bidirectional Encoder Representations from Transformers. *The Indonesian Journal of Computer Science*, *13*(4). https://doi.org/10.33022/ijcs.v13i4.4295