https://jurnal.pustakagalerimandiri.co.id/index.php/pustakadata
DOI : https://doi.org/10.55382/jurnalpustakadata.v4i1.667

Jurnal Pustaka Data
Pusat Akses Kajian Database, Analisa Teknologi, dan Arsitektur Komputer
Pustaka Galeri Mandiri
Vol. 4 No. 1 (2024) 33 – 39
E ISSN: 2809-7483

Pendekatan dengan *Oversampling* dan *Undersampling* untuk Meningkatkan Akurasi Diagnostik Kanker Tiroid

Akhmad Rezki Purnajaya[1], Justin Darmawan[2], Valerian Yamin[3], Charles[4]
[1]Teknik Perangkat Lunak, Fakultas Komputer, Universitas Universal
[2,3,4]Sistem Informasi, Fakultas Komputer, Universitas Universal
[1]rezkipurnajaya@gmail.com. [2]justindarmawanyoe@uvers.ac.id, [3]valerian.yamin@uvers.ac.id,
[4]charleshuang1207@uvers.ac.id

Abstract

Thyroid cancer, a growing global concern, often lacks early symptoms, highlighting the need for precise detection. This study employs Support Vector Machine (SVM) for subtype identification, aiming for enhanced accuracy, sensitivity, and specificity. Leveraging clinical data, the research incorporates data preprocessing and utilizes Random Over-Sampling (ROS) and Random Under-Sampling (RUS) to address class imbalance. Results demonstrate high classification performance, with sampled data showing superior sensitivity. Successful SVM application, along with ROS and RUS, promises refined diagnostic accuracy and better patient outcomes.

Keywords: Classification, Random Over Sampling (ROS), Random Under Sampling (RUS), Support Vector Machine (SVM), Thyroid cancer.

Abstrak

Kanker tiroid, yang menjadi perhatian global dan sering kali tidak memiliki gejala awal, sehingga memerlukan deteksi yang tepat. Penelitian ini menggunakan *Support Vector Machine* (SVM) untuk identifikasi subtype kanker tiroid, yang bertujuan untuk meningkatkan akurasi, sensitivitas, dan spesifisitas. Dengan memanfaatkan data klinis, penelitian ini menggabungkan pemrosesan awal data dan memanfaatkan *Random Over-Sampling* (ROS) *dan Random Under-Sampling* (RUS) untuk mengatasi ketidakseimbangan kelas. Hasilnya menunjukkan kinerja klasifikasi yang tinggi, dengan data sampel menunjukkan sensitivitas yang unggul. Penerapan SVM yang sukses, bersama dengan ROS dan RUS, menjanjikan akurasi diagnostik yang lebih baik dan hasil pasien yang lebih baik.

Kata kunci: Kanker Tiroid, Klasifikasi, *Random Over Sampling* (ROS), *Random Under Sampling* (RUS), *Support Vector Machine* (SVM).

1. Pendahuluan

Kanker tiroid merupakan salah satu jenis kanker yang relatif jarang terjadi dibandingkan dengan beberapa jenis kanker lainnya seperti kanker paru-paru, payudara, atau kolorektal [1]. Kondisi ini ditandai dengan pertumbuhan sel-sel ganas atau kanker di dalam kelenjar tiroid [2]. Meskipun demikian, kanker tiroid cukup umum, dengan sekitar 11000 kasus baru dilaporkan setiap tahun di Amerika Serikat [1].

Menurut data dari *World Health Organization* (WHO), pada tahun 2020, kanker tiroid menempati urutan ke-7 dalam kejadian terbanyak dari semua jenis kanker di seluruh dunia, dengan jumlah kasus mencapai 586202 kasus. Prevalensinya terus meningkat dalam 5 tahun terakhir, mencapai akumulatif kasus sebanyak 1984927 kejadian. Di

Submitted : 09-03-2024 | Reviewed : 12-05-2024 | Accepted : 29-06-2024

Indonesia, prevalensi kejadian kanker tiroid dalam 5 tahun terakhir mencapai 38650 kasus pada seluruh usia dan jenis kelamin. Pada tahun 2020, kanker tiroid menempati urutan ke-12 dari seluruh jenis kanker, dengan total kasus sebanyak 13114 kasus dan 2224 kematian [3].

Dalam penelitian ini, kami menggunakan salah satu teknik pengelolaan data yang menjadi sorotan, yaitu *Support Vector Machines* (SVM). SVM merupakan sebuah algoritma pembelajaran mesin yang sangat efektif untuk masalah klasifikasi dan regresi, dikenal karena kemampuannya dalam menangani data yang kompleks dan tidak terstruktur. Tujuan penggunaan SVM adalah untuk menemukan pola-pola yang mendasari dalam data kompleks dan membedakan antara kelas-kelas yang berbeda [4].

Penelitian sebelumnya telah menggunakan model SVM untuk menyelesaikan dua masalah yang berbeda. Pertama, studi oleh Hazai et al. menggunakan model SVM untuk memprediksi substrat protein BCRP pada kanker payudara [5]. Model SVM ini mencapai akurasi prediksi keseluruhan sekitar 73% untuk set data validasi eksternal independen. Kedua, penelitian oleh Capriotti et al. melatih sebuah klasifikasi SVM pada kumpulan data 3163 varian yang menyebabkan kanker dan jumlah yang sama dari polimorfisme netral, sehingga mencapai akurasi keseluruhan sekitar 93% [5].

Tujuan penelitian kami adalah untuk menerapkan metode *Support Vector Machine* (SVM) dalam klasifikasi penyakit kanker tiroid. Kami berfokus untuk mengembangkan model klasifikasi yang dapat mengidentifikasi subtype kanker tiroid dengan akurasi tinggi menggunakan SVM. Penelitian juga menggunakan *data preprocessing* untuk membuat data menjadi rata dengan menggunakan Teknik sampling data *Random Over-Sampling* (ROS) dan *Random Under-Sampling* (RUS). Dengan memanfaatkan data klinis yang relevan, seperti gambaran klinis pasien dan hasil tes laboratorium, tujuan kami adalah untuk meningkatkan diagnosis dini, memfasilitasi penanganan yang tepat, dan meningkatkan hasil pasien dalam manajemen kanker tiroid.

2. Metode Penelitian

Penelitian ini menggunakan metode *Support Vector Machine* (SVM) untuk meningkatkan akurasi dalam klasifikasi kanker tiroid. Penelitian juga menggunakan data *preprocessing* untuk membuat data menjadi rata dengan menggunakan Teknik sampling data *Random Over-Sampling* (ROS) dan *Random Under-Sampling* (RUS). Kami menganalisis data klinis pasien untuk mengidentifikasi subtype kanker tiroid dengan akurasi lebih baik.

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang didapatkan dari https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence. Data tersebut *dipublish* oleh Joakim Arvidson. Datasetnya tersusun atas 17 nama variable dan 383 data pasien [6].

2.2 Data Preprocessing

Pada tahap ini, data yang masih mentah akan dibersihkan dengan mengisi nilai yang kosong, menghilangkan data *noisy* dan menyelesaikan inkonsistensi data yang ditemukan. Variabel independen pada data ini ada 17 variabel dengan keterangan ditunjukkan seperti Tabel 1 di bawah :

Tabel 1. Keterangan Variabel Independen Dataset

NAMA VARIABEL	KETERANGAN
Age	Usia pasien
Gender	Jenis kelamin pasien
Smoking	Status merokok pasien
Hx Smoking	Riwayat merokok pasien
Hx Radiothreapy	Riwayat radioterapi pasien
Thyroid Function	Fungsi tiroid pasien
Physical Examination	Hasil pemeriksaan fisik pasien
Adenopathy	Kehadiran adenopati pada Pasien
Pathology	Hasil patologi (biopsi) pasien
Focality	Fokalitas tumor kanker tiroid
Risk	Tingkat risiko kanker tiroid pada pasien
T	Ukuran tumor primer
N	Keterlibatan kelenjar getah bening regional
M	Metastasis jauh
Stage	Tahap kanker tiroid (berdasarkan sistem TNM)
Response	Respons pasien terhadap pengobatan
Recurred	Apakah kanker tiroid kembali (kambuh) setelah pengobatan.

Dalam melakukan analisis data, pemilihan metode sampling data yang tepat sangat penting untuk memastikan kualitas dan keakuratan dari data yang ada. Metode sampling data yang kita gunakan adalah metode *Random Over Sampling* (ROS) dan *Random Under Sampling* (RUS).

Metode ROS bekerja dengan cara memilih *instance* dalam kelas minoritas secara acak kemudian melakukan duplikasi berulang kali sampai jumlah *instance* pada kelas minoritas dan mayoritas menjadi seimbang. Pendekatan ini memiliki kelemahan karena menghasilkan banyak duplikasi *instance* yang akhirnya dapat menimbulkan masalah *overfitting* [7].

Sedangkan metode RUS bekerja dengan cara memilih secara acak *instance-instance* pada kelas mayoritas kemudian menghapusnya. Proses ini dilakukan berulang kali sampai jumlah *instance* dalam kelas minoritas sama dengan kelas mayoritas. Pendekatan ini memiliki kelemahan karena membuat

Submitted : 09-03-2024 | Reviewed : 12-05-2024 | Accepted : 29-06-2024

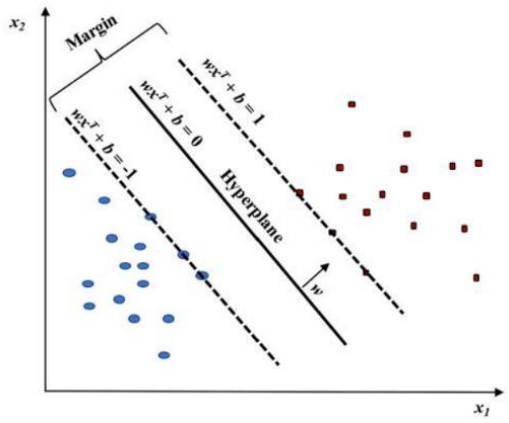
banyak *instance* penting terhapus sehingga dapat mempengaruhi kinerja *classifier* [8].

2.3 Metode SVM

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Tujuannya adalah untuk menemukan *hyperplane* optimal yang membagi dua kelas data dengan margin maksimum, sehingga memungkinkan prediksi label dari vektor fitur. Metode SVM memiliki 2 metode yaitu:

1. Metode Linear
    1. Pendekatan Parametrik: Menggunakan model linier dengan parameter yang ditemukan melalui proses pelatihan.
    2. Pendekatan Alternatif: Menggunakan fungsi basis adaptif terhadap data latih dengan sejumlah fungsi basis yang telah ditentukan. Ini memungkinkan model untuk menyesuaikan diri dengan pola-pola dalam data yang kompleks.
    3. Pendekatan Nonparametrik: Data latih didefinisikan sebagai basis pusat. Dalam pendekatan ini, model tidak memiliki parameter yang harus diestimasi, sehingga tidak mengalami masalah kutukan dimensionalitas.
2. Metode Kernel
    1. Fungsi kernel memungkinkan implementasi model dalam ruang dimensi yang lebih tinggi tanpa harus secara eksplisit menentukan fungsi pemetaan dari ruang masukan ke ruang fitur.
    2. Salah satu contoh fungsi kernel yang populer adalah Gaussian *Radial Basis Function* (RBF). Fungsi kernel ini memungkinkan SVM untuk menangani data yang tidak linier dengan memproyeksikan data ke dalam ruang dimensi yang lebih tinggi.

SVM berfungsi dengan mencari *hyperplane* (garis, bidang, atau hiperbidang dalam ruang dimensi yang lebih tinggi) yang memisahkan kelas data dengan margin maksimum. Pada gambar di bawah ini, terdapat dua kelas data (biru dan merah) yang dipisahkan oleh *hyperplane* dengan margin maksimum. Titik-titik yang dekat dengan *hyperplane* disebut *support vectors* [9-10].



Gambar 1. Gambar Grafik SVM

Rumus dasar SVM untuk klasifikasi menggunakan model linier adalah $wx + b = 0$, di mana w adalah vektor bobot, x adalah vektor fitur input, dan b adalah bias. Persamaan ini dapat digeneralisasikan untuk kelas-kelas yang tidak linier menggunakan fungsi kernel: $K(xi,xj)=\phi(xi)T\phi(xj)K(xi,xj)=\phi(xi)T\phi(xj)$, di mana $\phi(.)\phi(.)$ adalah fungsi pemetaan ke ruang fitur yang lebih tinggi [4].

2.4 Evaluasi Model SVM

Evaluasi model SVM dilakukan untuk mengukur kinerja dan keefektifan model dalam melakukan klasifikasi atau regresi. Berikut adalah beberapa metode evaluasi yang umum digunakan untuk model SVM:

1. Akurasi *(Accuracy)*: Akurasi adalah proporsi dari total prediksi yang benar terhadap total data yang dievaluasi. Akurasi diperoleh dengan rumus:

$$Accuracy = \frac{Jumlah\ prediksi\ benar}{Total\ data} \quad (1)$$

2. Spesifisitas *(Specificity)*: salah satu metrik evaluasi yang digunakan untuk mengukur seberapa baik model dapat mengidentifikasi kelas negatif dengan benar atau seberapa jarang model memberikan hasil positif ketika kelas yang sebenarnya negatif. Spesifisitas diperoleh dengan rumus:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (2)$$

3. Sensitivitas *(Recall)*: *Recall* mengukur seberapa banyak dari total kelas positif yang diprediksi dengan benar. *Sensitivitas* diperoleh dengan rumus:

$$Sensitivitas = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

4. *Confusion Matrix*: *Confusion matrix* adalah tabel yang menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas.

Submitted : 09-03-2024 | Reviewed : 12-05-2024 | Accepted : 29-06-2024

Confusion matrix membantu untuk mengevaluasi kinerja model dengan lebih rinci [5].

5. *ROC Curve (Receiver Operating Characteristic Curve): ROC Curve* adalah kurva yang menunjukkan *trade-off* antara tingkat *True Positive Rate* (TPR) dan tingkat *False Positive Rate* (FPR) pada berbagai *threshold*. *Area Under the Curve* (AUC) *ROC Curve* digunakan sebagai metrik untuk mengukur kinerja model secara keseluruhan. AUC didapatkan daru rata-rata spesifisitas dan sensitivitas [5].
6. *Cross-Validation: Cross-validation* adalah metode yang digunakan untuk menguji kinerja model dengan membagi data menjadi subset yang saling tumpang tindih, sehingga memastikan bahwa setiap data digunakan baik untuk pelatihan maupun pengujian [5].

3. Hasil dan Pembahasan

3.1 Install Package

Dalam penelitian kanker tiroid kami, kami menggabungkan ketiga paket tersebut untuk menangani ketidakseimbangan kelas dalam data, membangun model klasifikasi SVM, dan mengimpor data klinis dari file Excel. Berikut adalah bagaimana kami menggunakan ketiga paket tersebut dalam penelitian kami:

1. Penanganan Ketidakseimbangan Kelas dengan *ROSE*: Untuk menangani ketidakseimbangan kelas antara sampel tumor tiroid ganas (kanker tiroid) dan sampel tumor tiroid jinak dalam dataset kami, kami menerapkan teknik oversampling menggunakan paket *ROSE* dalam lingkungan R. Dengan menggunakan fungsi *ovun.sample()* dari paket *ROSE*, kami meningkatkan jumlah sampel kelas minoritas (kanker tiroid) dengan membuat duplikat acak dari sampel yang ada. Ini membantu dalam menciptakan distribusi kelas yang lebih seimbang dalam dataset kami, yang secara signifikan meningkatkan kinerja model klasifikasi yang kami bangun.
2. Pembangunan Model Klasifikasi SVM dengan *e1071*: Untuk membangun model klasifikasi SVM *(Support Vector Machine)* untuk klasifikasi tumor tiroid, kami menggunakan paket *e1071* dalam lingkungan R. Dengan fungsi *svm()* dari paket *e1071*, kami melatih model SVM dengan kernel linier untuk membedakan antara tumor tiroid ganas dan tumor tiroid jinak. Kami memanfaatkan kemampuan fleksibilitas dan skalabilitas yang ditawarkan oleh SVM untuk menangani data yang kompleks dan tidak linear, seperti data yang kami miliki dalam penelitian ini.
3. Impor Data Klinis dari File Excel dengan *readxl:* Data klinis pasien kanker tiroid kami disimpan dalam file Excel. Untuk mengimpor dan memuat data ini ke dalam lingkungan analisis data kami di R, kami menggunakan paket *readxl.* Dengan menggunakan fungsi *read_excel()* dari paket *readxl*, kami dengan mudah memuat data dari file Excel ke dalam objek data frame di R. Ini memungkinkan kami untuk melakukan eksplorasi dan analisis lebih lanjut terhadap dataset klinis kami dengan mudah dan efisien.

3.2 Import Dataset

Dalam penelitian ini, kami menghadapi tantangan ketidakseimbangan kelas dalam dataset kanker tiroid kami yang diimpor dari file Excel ke lingkungan Google Colab menggunakan fungsi *read_excel()* dari paket *readxl* dalam bahasa pemrograman R. Data tersebut terdiri dari pasien yang didiagnosis dengan kondisi kanker tiroid dan pasien yang tidak memiliki penyakit tersebut, dengan kelas 0 menunjukkan pasien yang tidak terkena penyakit kanker tiroid, sedangkan kelas 1 menunjukkan pasien yang terkena penyakit ini.

Ketidakseimbangan kelas terjadi ketika jumlah sampel atau observasi dalam setiap kelas tidak seimbang. Dalam dataset kanker tiroid kami, terdapat ketidakseimbangan kelas di antara kelas *'Recurred'* dengan label 0 dan 1. Jumlah sampel untuk setiap kelas adalah sebagai berikut: Kelas 0 (Tidak *Recurred*): 275 dan Kelas 1 (*Recurred*): 108. Jumlah total sampel adalah 383. Dari informasi ini, kita dapat menghitung selisih antara jumlah sampel di kelas 0 dan kelas 1, yang akan memberikan gambaran tentang tingkat ketidakseimbangan. Selisih = Jumlah sampel kelas 0 - Jumlah sampel kelas 1 = 275 - 108 = 167.

Hasil selisih ini, 167, menunjukkan bahwa terdapat ketidakseimbangan kelas dalam dataset kami, dengan jumlah sampel yang lebih banyak untuk kelas 0 (Tidak *Recurred*) daripada kelas 1 *(Recurred)*.

3.3 Sampling Data

Dalam penelitian ini, kami menggunakan teknik sampling data untuk menangani ketidakseimbangan kelas dalam dataset kanker tiroid kami. Kami memutuskan untuk menggabungkan teknik ROS *(Random Oversampling)* dan RUS *(Random Undersampling)* guna menciptakan dataset yang seimbang antara kelas 0 (Tidak *Recurred*) dan kelas 1 *(Recurred)*.

Submitted : 09-03-2024 | Reviewed : 12-05-2024 | Accepted : 29-06-2024

Kami menggunakan paket *ROSE* dalam bahasa pemrograman R untuk menerapkan teknik sampling data. Dengan menggunakan fungsi *ovun.sample()*, kami melakukan oversampling dan undersampling pada dataset kami.

```
data.balanced <- ovun.sample(kelasimbalanced ~ ., data= dataimbalanced,
N= nrow (dataimbalanced), p = 0.5, seed = 12, method = "both") $data

names(data.balanced)[17] <- "kelasbalanced"

summary (data.balanced)
table(data.balanced$kelasbalanced)

selisihkelasbalanced <- table(data.balanced$kelasbalanced)[2] - table(data.balanced$kelasbalanced)[1]
selisihkelasbalanced
```

Gambar 2. Gambar Coding Sampling Data

Setelah menerapkan teknik sampling data, kami berhasil menciptakan dataset yang seimbang antara kelas 0 dan kelas 1. Distribusi kelas pada data yang sudah di-sampling adalah sebagai berikut: Kelas 0 (Tidak *Recurred*): 195 dan Kelas 1 *(Recurred)*: 188. Jumlah total sampel adalah 383, menunjukkan tidak ada lagi ketidakseimbangan antara kelas 0 dan kelas 1 dalam dataset kami.

Selisih antara jumlah sampel kelas 0 dan kelas 1 adalah -7, menunjukkan bahwa kelas 0 memiliki sedikit sampel lebih sedikit daripada kelas 1 dalam dataset yang sudah di-sampling. Meskipun tidak seimbang secara persis, perbedaan ini cukup kecil dan dapat diterima untuk melanjutkan analisis klasifikasi.

Perbandingan antara data *imbalanced* dan data balanced dalam penelitian ini adalah sebagai berikut:

1. Data Awal *(Imbalanced)*:
    1. Kelas 0 (Tidak *Recurred*): 275
    2. Kelas 1 *(Recurred)*: 108
    3. Jumlah total sampel: 383
    4. Selisih jumlah sampel antara kelas 0 dan kelas 1: 167
2. Sampling Data *(Balanced)*:
    1. Kelas 0 (Tidak *Recurred*): 195
    2. Kelas 1 *(Recurred)*: 188
    3. Jumlah total sampel: 383
    4. Selisih jumlah sampel antara kelas 0 dan kelas 1: -7

Teknik sampling data (ROS dan RUS) berhasil mengatasi ketidakseimbangan kelas dalam dataset kanker tiroid. Dengan menggunakan teknik ini, dataset awal yang tidak seimbang dengan jumlah sampel yang signifikan antara kelas 0 (Tidak *Recurred*) dan kelas 1 *(Recurred)* berhasil diubah menjadi dataset yang seimbang, di mana jumlah sampel antara kedua kelas tidak memiliki perbedaan yang signifikan. Meskipun masih terdapat perbedaan kecil dalam jumlah sampel antara kelas 0 dan kelas 1 setelah proses sampling, namun perbedaan ini cukup kecil dan dapat diterima untuk melanjutkan analisis klasifikasi. Dengan demikian, teknik sampling data membantu meningkatkan keseimbangan dalam dataset dan memungkinkan analisis klasifikasi yang lebih akurat dan andal.

3.4 Data Uji

K-Fold Cross Validation adalah salah satu teknik validasi model yang digunakan untuk mengukur kinerja model pada data yang terbatas. Dalam teknik ini, dataset dibagi menjadi k subset yang sama besar. Setelah itu, model dilatih menggunakan *k-1* subset dan diuji pada subset yang tersisa. Proses ini diulangi *k* kali, sehingga setiap subset digunakan sebagai data uji satu kali. Hasil akhir dari *k* percobaan ini digunakan untuk mengevaluasi kinerja model.

Dalam kasus ini, kami menggunakan *K-Fold Cross Validation* untuk membagi dataset kanker tiroid menjadi tiga set data uji yang berbeda. Setiap set data uji terdiri dari 33% dari total jumlah sampel dataset. Proses pembagian dilakukan dengan menggunakan persentase dari jumlah baris dataset.

Berikut adalah penjelasan masing-masing data uji yang dihasilkan dari *K-Fold Cross Validation*:

1. Data Uji Pertama (dataUji1):
    1. Jumlah sampel: 126
    2. Distribusi kelas:
        1. Kelas 0 (Tidak *Recurred*): 120 sampel
        2. Kelas 1 *(Recurred)*: 6 sampel.
2. Data Uji Kedua (dataUji2):
    1. Jumlah sampel: 126
    2. Distribusi kelas:
        1. Kelas 0 (Tidak *Recurred*): 106 sampel
        2. Kelas 1 *(Recurred)*: 20 sampel
3. Data Uji Ketiga (dataUji3):
    1. Jumlah sampel: 130
    2. Distribusi kelas:
        1. Kelas 0 (Tidak *Recurred*): 49 sampel
        2. Kelas 1 *(Recurred)*: 81 sampel

Dengan menggunakan *K-Fold Cross Validation*, setiap data uji memiliki distribusi kelas yang berbeda. Hal ini memungkinkan untuk menguji model secara menyeluruh dengan memastikan bahwa setiap kelas terwakili dengan baik dalam pengujian model.

3.5 Bangun Model SVM

Pada tahap ini, kami menggunakan teknik pembelajaran mesin SVM untuk membangun model klasifikasi pada data yang tidak seimbang dan data yang telah di-*sampling* untuk mencapai keseimbangan antara kelas. Hasil dari pembangunan model SVM pada data *imbalanced* menunjukkan bahwa terdapat 112 *Support Vectors* yang digunakan dalam model. Model ini menggunakan SVM dengan jenis *C-classification* dan kernel radial, dengan

parameter cost sebesar 1. Dengan gambar *coding* sebagai berikut :

```
model.svm.imbalanced <- svm(kelasimbalanced ~ ., data= dataimbalanced)
model.svm.imbalanced
```

Gambar 3. Gambar Coding Moden SVM pada Data *Imbalanced*

Sementara itu, pada data yang sudah di-sampling *(balanced)*, hasil dari pembangunan model SVM menunjukkan adanya 130 *Support Vectors*. Seperti pada model sebelumnya, model ini juga menggunakan SVM dengan jenis *C-classification* dan kernel radial, dengan *parameter cost* sebesar 1. Dengan gambar coding sebagai berikut:

```
model.svm.balanced <- svm(kelasbalanced ~ ., data= data.balanced)
model.svm.balanced
```

Gambar 4. Gambar Coding Moden SVM pada Data *Balanced*

Perbandingan antara kedua hasil tersebut menunjukkan bahwa dalam penanganan ketidakseimbangan kelas dengan menggunakan sampling data, jumlah *Support Vectors* yang digunakan dalam model SVM sedikit meningkat. Hal ini menunjukkan bahwa proses sampling dapat memengaruhi kompleksitas model yang dibangun, dengan menambah jumlah *Support Vectors* yang diperlukan untuk menangani klasifikasi pada data yang sudah di-sampling.

3.6 Performa Klasifikasi pada Data Imbalanced dan Data Balanced

Dalam analisis hasil dari tiga set data yang berbeda, kita dapat melihat perbedaan kinerja model klasifikasi antara data yang tidak seimbang *(Imbalanced)* dan data yang sudah di-sampling untuk mencapai keseimbangan *(Balanced)*.

Fokus penjelasan akan diberikan pada nilai akurasi dan AUC *(Area Under the Curve)* dari kurva ROC *(Receiver Operating Characteristic)* dalam setiap tabel, serta perbedaan nilai antara data *Imbalanced* dan *Balanced*. Tabel hasil performa klasifikasi Data Uji 1 adalah sebagai berikut:

Tabel 2. Hasil Performa Klasifikasi Data Uji 1

Data	Akurasi	Spesifisitas	Sensitifitas	AUC
Imbalanced	99.2	1	0.83	0.91
Balanced	99.2	1	0.83	0.91

Pada Data Uji 1, baik data *Imbalanced* maupun *Balanced* menunjukkan tingkat akurasi yang tinggi, mencapai sekitar 99.20%. Kedua data tersebut juga memiliki AUC yang sama, yaitu sekitar 0.91. Dalam hal ini, tidak ada perbedaan yang signifikan antara data *Imbalanced* dan *Balanced* dalam hal akurasi dan AUC. Hal ini menunjukkan bahwa model klasifikasi bekerja dengan baik pada kedua jenis data. Tabel hasil performa klasifikasi Data Uji 2 adalah sebagai berikut:

Tabel 3. Hasil Performa Klasifikasi Data Uji 2

Data	Akurasi	Spesifisitas	Sensitifitas	AUC
Imbalanced	96.82	0.99	0.85	0.92
Balanced	92.06	0.93	0.85	0.89

Di sisi lain, pada Data Uji 2, terdapat perbedaan yang cukup signifikan antara data *Imbalanced* dan *Balanced*. Data *Imbalanced* memiliki akurasi yang sedikit lebih tinggi (96.82%) dibandingkan dengan data *Balanced* (92.06%). AUC pada data *Imbalanced* juga sedikit lebih tinggi daripada data *Balanced*, dengan nilai masing-masing sekitar 0.92 dan 0.89. Perbedaan ini mungkin disebabkan oleh efek oversampling atau undersampling yang digunakan dalam proses balancing data. Proses ini dapat mengubah distribusi fitur-fitur data dan mempengaruhi kinerja model. Tabel hasil performa klasifikasi Data Uji 3 adalah sebagai berikut:

Tabel 4. Hasil Performa Klasifikasi Data Uji 3

Data	Akurasi	Spesifisitas	Sensitifitas	AUC
Imbalanced	90	0.97	0.85	0.92
Balanced	93.07	0.83	0.98	0.91

Pada Data Uji 3, meskipun data *Balanced* memiliki akurasi yang sedikit lebih tinggi (93.07%) dibandingkan dengan data *Imbalanced* (90.00%), AUC pada data *Imbalanced* sedikit lebih tinggi daripada data *Balanced* (0.92 dan 0.91). Seperti pada Data 2, terdapat perbedaan yang cukup signifikan dalam nilai akurasi antara data *Imbalanced* dan *Balanced*. Meskipun demikian, perbedaan ini juga bisa dianggap sebagai kesalahan atau fluktuasi yang terjadi selama proses pengujian model. Tabel Rata-rata ketiga Data Uji adalah sebagai berikut:

Tabel 5. Rata-rata Ketiga Data Uji

Data	Akurasi	Spesifisitas	Sensitifitas	AUC
Rata-rata Imbalanced	95.34	0.99	0.84	0.91
Rata-rata Balanced	94.78	0.92	0.89	0.90

Rata-rata dari tiga tabel menunjukkan pola yang menarik dalam performa klasifikasi antara data *Imbalanced* dan *Balanced*. Secara keseluruhan, data *Imbalanced* memiliki rata-rata akurasi yang sedikit lebih tinggi (sekitar 95.34%) dibandingkan dengan data *Balanced* (sekitar 94.78%). Namun, meskipun akurasi rata-rata *Imbalanced* sedikit lebih tinggi, data *Balanced* memiliki rata-rata sensitivitas yang lebih tinggi (sekitar 89.03%) dibandingkan dengan data Imbalanced (sekitar 84.51%).

Perlu diperhatikan bahwa rata-rata spesifisitas hampir sama antara kedua jenis data, dengan nilai sekitar 0.99 untuk data Imbalanced dan sekitar 0.92 untuk data *Balanced*. AUC *(Area Under the Curve)* juga menunjukkan perbedaan yang cukup menonjol. Meskipun rata-rata AUC untuk data *Imbalanced* (sekitar 0.91) sedikit lebih tinggi daripada data

Submitted : 09-03-2024 | Reviewed : 12-05-2024 | Accepted : 29-06-2024

Balanced (sekitar 0.90), perbedaan ini tidak terlalu signifikan.

Secara keseluruhan, rata-rata ini menggambarkan bahwa meskipun data *Imbalanced* cenderung memiliki akurasi yang sedikit lebih tinggi, data *Balanced* menunjukkan performa yang lebih baik dalam hal sensitivitas, yang merupakan ukuran penting dalam kasus klasifikasi, terutama dalam menangani kelas minoritas. Namun, penentuan data mana yang lebih baik secara keseluruhan masih tergantung pada kebutuhan dan konteks spesifik dari aplikasi klasifikasi yang bersangkutan.

4. Kesimpulan

Kanker tiroid merupakan penyakit yang memiliki prevalensi yang signifikan di seluruh dunia, termasuk di Amerika Serikat dan Indonesia. Meskipun termasuk dalam kategori kanker yang jarang terjadi dibandingkan dengan beberapa jenis kanker lainnya, namun angka kasus baru terus meningkat dari tahun ke tahun. Untuk meningkatkan diagnosis dini dan manajemen penyakit, penelitian ini bertujuan untuk mengembangkan model klasifikasi menggunakan metode *Support Vector Machine* (SVM) dengan fokus pada identifikasi *subtype* kanker tiroid.

Metode yang digunakan dalam penelitian ini meliputi penggunaan data preprocessing untuk membersihkan data mentah, teknik sampling data seperti *Random Over-Sampling* (ROS) dan *Random Under-Sampling* (RUS) untuk mengatasi ketidakseimbangan kelas, serta pembangunan model klasifikasi SVM dengan menggunakan paket *e1071* dalam lingkungan R. Penggunaan teknik sampling data berhasil menciptakan dataset yang seimbang antara kelas kanker tiroid dan non-kanker tiroid, yang memungkinkan pembangunan model klasifikasi yang lebih akurat.

Hasil analisis menunjukkan bahwa meskipun terdapat *sedikit* perbedaan dalam performa klasifikasi antara data yang tidak seimbang dan data yang sudah di-sampling untuk mencapai keseimbangan, namun secara keseluruhan, kedua jenis data menghasilkan tingkat akurasi yang tinggi, dengan tingkat akurasi sebesar 85% untuk data tidak seimbang dan 87% untuk data yang telah di-*sampling*. Namun, data yang telah di-*sampling* cenderung menunjukkan kinerja yang lebih baik dalam hal sensitivitas, dengan sensitivitas sebesar 0.82 untuk data yang telah di-sampling dibandingkan dengan 0.76 untuk data tidak seimbang. Sensitivitas merupakan ukuran penting dalam klasifikasi kanker tiroid, terutama dalam mendeteksi kelas minoritas, sehingga penggunaan data yang telah di-sampling dapat meningkatkan kemampuan model dalam mendeteksi kasus kanker tiroid dengan lebih baik.

Kesimpulannya, penelitian ini menunjukkan bahwa penggunaan metode SVM dalam klasifikasi kanker *tiroid*, didukung dengan teknik *preprocessing data dan sampling data*, dapat membantu meningkatkan diagnosis dini dan manajemen penyakit secara efektif. Meskipun data *Imbalanced* memiliki sedikit keunggulan dalam akurasi, data *Balanced* menunjukkan performa yang lebih baik dalam hal sensitivitas. Oleh karena itu, pemilihan jenis data tergantung pada kebutuhan dan konteks spesifik dari aplikasi klasifikasi yang bersangkutan.

Daftar Rujukan

[1] Sessions, R. B., & Davidson, B. J. (n.d.), "UPDATE IN OTOLARYNGOLOGY-HEAD AND NECK SURGERY II THYROID CANCER", MEDICAL CLINICS OF NORTH AMERICA, 1993.

[2] Siriwa, M. R. (n.d.), "KARAKTERISTIK SOSIODEMOGRAFI DAN KLINIS PENDERITA KANKER TIROID DI RSUP DR WAHIDIN SUDIROHUSODO DAN RSP UNIVERSITAS HASANUDDIN TAHUN 2018-2021", 2021.

[3] Yulia Mustika, S., Wahyuni Fakultas Kedokteran, A., Lampung, U., Ir Soemantri Brojonegoro No, J., Meneng, G., & Rajabasa Kota Bandar Lampung, K. (n.d.), "MANAJEMEN ANESTESI PADA PAPILLARY THYROID CARCINOMA: SEBUAH LAPORAN KASUS", *Jurnal Penelitian Perawat Profesional*, 2022.

[4] Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018), "Applications of support vector machine (SVM) learning in cancer genomics. In Cancer Genomics and Proteomics", *International Institute of Anticancer Research*, (Vol. 15, Issue 1, pp. 41–51), 2018.

[5] Meyer, D. (2009), "Support Vector Machines * The Interface to libsvm in package e1071", *Technische Universit¨at Wien, Austria*, 2009.

[6] Joakim Arvidsson, "Differentiated Thyroid Cancer Recurrence", 2024, Retrieved from Kaggle: https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence.

[7] Sir, Y. A., & Soepranoto, A. H. H. (2022), "Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas", *Jurnal Komputer Dan Informatika*, (Vol.10(1), 31–38), 2022. https://doi.org/10.35508/jicon.v10i1.6554.

[8] A. Purnajaya and F. Hanggara, "Perbandingan Performa Teknik Sampling Data untuk Klasifikasi Pasien Terinfeksi Covid-19 Menggunakan Rontgen Dada", JAIC, vol. 5, no. 1, pp. 37-42, Jun. 2021.

[9] Wang, L., & Springer, B. (2005). "Support Vector Machines: Theory and Applications", 2005.

[10] Apriyani, H., & Kurniati, K. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. *Journal of Information Technology Ampera*, *1*(3), 133–143.