IMPLEMENTASI *SUPPORT VECTOR MACHINE* TERHADAP KLASIFIKASI DIAGNOSIS PENDERITA KANKER PAYUDARA

Skripsi

Oleh

ELIZABETH CEASARINA SITOMPUL
NPM. 2217031016

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2026

ABSTRACT

IMPLEMENTATION OF SUPPORT VECTOR MACHINE FOR CLASSIFICATION OF DIAGNOSIS OF BREAST CANCER PATIENTS

By

Elizabeth Ceasarina Sitompul

Support Vector Machine (SVM) is one of the machine learning methods used for classification problems by dividing data into two different classes through the formation of an optimal separating function called a hyperplane. The purpose of this study is to apply the SVM method and determine the best kernel function based on the highest accuracy value in the classification process. However, the dataset used in this study has a problem of data imbalance, so a special handling method is required. Therefore, the Random Oversampling (ROS) method is used to balance the data distribution before the classification process is carried out. In the classification process, several kernel functions were applied, namely linear, polynomial, sigmoid, and Radial Basis Function (RBF) kernels. The results of the study showed that the RBF kernel function provided the best performance compared to other kernels. In the data division scheme of 90% as training data and 10% as test data, with parameters C =10, 50, and 100 and $\gamma = 0.001$, the SVM model produced the highest accuracy value of 97.22%. These results indicate that the use of the RBF kernel is able to handle nonlinear data patterns more optimally. Using this kernel configuration, the SVM model parameters obtained, namely the $\mathbf{w}$ and b, are as follows:
$\mathbf{w}_{\text{radius\_mean}} = 377.7713,\ \mathbf{w}_{\text{texture\_mean}} = 340.6203,\ \mathbf{w}_{\text{perimeter\_mean}} = 2851.3717,$
$\mathbf{w}_{\text{area\_mean}} = 37057.1680,\ \ldots,\ \mathbf{w}_{\text{fractal\_dimension\_worst}} = 2.8828,\ b = 0.5153.$

Keywords: Support Vector Machine, Radial Basis Function Kernel, Data Imbalance, Random Oversampling, Breast Cancer.

ABSTRAK

IMPLEMENTASI *SUPPORT VECTOR MACHINE* TERHADAP KLASIFIKASI DIAGNOSIS PENDERITA KANKER PAYUDARA

Oleh

Elizabeth Ceasarina Sitompul

Support Vector Machine (SVM) merupakan salah satu metode *machine learning* yang digunakan untuk permasalahan klasifikasi dengan membagi data ke dalam dua kelas yang berbeda melalui pembentukan fungsi pemisah optimal yang disebut *hyperplane*. Tujuan penelitian ini adalah menerapkan metode SVM serta menentukan fungsi *kernel* terbaik berdasarkan nilai akurasi tertinggi pada proses klasifikasi. Namun, dataset yang digunakan dalam penelitian kali ini memiliki permasalahan ketidakseimbangan data (*imbalance data*), sehingga diperlukan metode penanganan khusus. Dengan demikian, digunakan metode *Random Oversampling* (ROS) untuk menyeimbangkan distribusi data sebelum proses klasifikasi dilakukan. Dalam Proses klasifikasi selanjutnya diterapkan menggunakan beberapa fungsi *kernel*, yaitu *kernel* linear, polinomial, sigmoid, dan *Radial Basis Function* (RBF). Hasil penelitian menunjukkan bahwa fungsi *kernel* RBF memberikan kinerja terbaik dibandingkan *kernel* lainnya. Pada skema pembagian data 90% sebagai data latih dan 10% sebagai data uji, dengan parameter C = 10, 50, dan 100 serta $\gamma = 0.001$, model SVM menghasilkan nilai akurasi tertinggi sebesar 97,22%. Hasil ini menunjukkan bahwa penggunaan *kernel* RBF mampu menangani pola data nonlinear secara lebih optimal. Dengan menggunakan konfigurasi *kernel* tersebut, diperoleh parameter $\mathbf{w}$ dan b, yakni sebagai berikut:
$\mathbf{w}_{\text{radius\_mean}} = 377,7713,\ \mathbf{w}_{\text{texture\_mean}} = 340,6203,\ \mathbf{w}_{\text{perimeter\_mean}} = 2851,3717,$
$\mathbf{w}_{\text{area\_mean}} = 37057,168,\ \ldots,\ \mathbf{w}_{\text{fractal\_dimension\_worst}} = 2,8828,\ b = 0,5153.$

Kata Kunci: *Support Vector Machine*, *Kernel Radial Basis Function*, Ketidakseimbangan Data, *Random Oversampling*, Kanker Payudara.

IMPLEMENTASI *SUPPORT VECTOR MACHINE* TERHADAP KLASIFIKASI DIAGNOSIS PENDERITA KANKER PAYUDARA

ELIZABETH CEASARINA SITOMPUL

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2026

Judul Skripsi	:	**IMPLEMENTASI *SUPPORT VECTOR MACHINE* TERHADAP KLASIFIKASI DIAGNOSIS PENDERITA KANKER PAYUDARA**
Nama Mahasiswa	:	Elizabeth Ceasarina Sitompul
Nomor Pokok Mahasiswa	:	**2217031016**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Subian Saidi, S.Si., M.Si	**Dr. Bernadhita H. S. U, S.Si., M.Sc**
NIP 198008212008121001	NIP 199206302023212034

2. Ketua Jurusan Matematika

Dr.Aang Nuryaman, S.Si.,M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Subian Saidi, S.Si., M.Si** .....

Sekretaris : **Dr. Bernadhita H. S. U, S.Si., M.Sc** .....

Penguji
Bukan Pembimbing : **Drs. Nusyirwan, M.Si** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **7 April 2026**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Elizabeth Ceasarina Sitompul**
Nomor Pokok Mahasiswa	:	**2217031016**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Implementasi *Support Vector Machine* Terhadap Klasifikasi Diagnosis Penderita Kanker Payudara**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 7 April 2026

Elizabeth Ceasarina Sitompul

RIWAYAT HIDUP

Penulis bernama Elizabeth Ceasarina Sitompul yang lahir di Bandar Lampung pada tanggal 16 Oktober 2003. Penulis merupakan anak pertama dari dua bersaudara dari pasangan Bapak Jun Bastian Sitompul dan Ibu Rintauli Purba.

Penulis mengawali pendidikan di Taman Kanak-Kanak (TK) Bratasena Adiwarna tahnun 2008-2010. Kemudian menempuh pendidikan Sekolah Dasar (SD) di SDS Tunas Bangsa tahun 2010-2016. Melanjutkan ke Sekolah Menengah Pertama (SMP) di SMPN 2 Bandar Lampung dan lulus pada tahun 2019. Lalu penulis melanjutkan pendidikan Sekolah Menengah Atas (SMA) di SMAN 13 Bandar Lampung dan lulus pada tahun 2022.

Pada tahun 2022, penulis berhasil lulus Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) dan diterima di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung. Selama menjadi mahasiswa, penusil aktif dalam organisasi Himpunan Mahasiswa Jurusan Matematika (HIMATIKA) sebagai anggota Kaderisasi & Kepemimpinan, Dies Natalis Jurusan Matematika (DINAMIKA) sebagai bendahara.

Pada tahun 2025, penulis melakukan Kerja Praktik (KP) di Dinas Perindustrian dan Perdagangan Provinsi Lampung (Disperindag) dan Kuliah Kerja Nyata (KKN) di Desa Padang Ratu Kecamatan Padang Ratu. Berdasarkan latar belakang pendidikan dan pengalaman yang dimiliki, penulis berharap penelitian ini dapat memberikan kontribusi yang bermanfaat, khususnya dalam bidang matematika, terutama pada ranah statistika.

KATA INSPIRASI

“Ombakku besar, perahuku kecil. Tapi, Yesusku Terlebih Besar”

– Penulis –

”Janganlah hendaknya kamu kuatir tentang apa pun juga, tetapi nyatakanlah dalam segala hal keinginanmu kepada Allah dalam doa dan permohonan dengan ucapan syukur”

– Filipi 4:6 –

”Karena itu Aku berkata kepadamu: apa saja yang kamu minta dan doakan, percayalah bahwa kamu telah menerimanya, maka hal itu akan diberikan kepadamu”

– Markus 11:24 –

Segala perkara dapat kutanggung di dalam Dia yang memberi kekuatan kepadaku”

– Filipi 4:13 –

”Janganlah takut, sebab Aku menyertai engkau, janganlah bimbang, sebab Aku ini Allahmu; Aku akan meneguhkan, bahkan akan menolong engkau; Aku akan memegang engkau dengan tangan kanan-Ku yang membawa kemenangan”

– Yesaya 41:10 –

”Tetap berseru dan andalkan Tuhan dalam hidupmu”

– Mama –

PERSEMBAHAN

Segala puji dan syukur kepada Tuhan Yesus Kristus atas segala kasih, anugerah, dan pernyertaan-Nya yang dinyatakan melalui orang-orang yang membimbing dan mendukung penulis dalam berbagai bentuk, sehingga skripsi ini dapat disusun dan diselesaikan dengan baik. Dengan penuh ketulusan dan rasa syukur, karya ini penulis persembahkan untuk:

Kedua Orang Tuaku Tercinta

Puji syukur kepada Tuhan Yesus Kristus atas kasih dan penyertaan-Nya melalui kehadiran Papa dan Mama dalam hidup penulis. Dengan segala kerendahan hati, penulis mengucapkan terima kasih yang sebesar-besarnya atas setiap pengorbanan, kasih sayang, doa yang tidak pernah putus, serta dukungan moril maupun materi yang senantiasa menguatkan penulis hingga dapat menyelesaikan skripsi ini. Kalian adalah sumber cinta pertama dan tempat penulis selalu kembali, yang telah mengajarkan makna kehidupan, pengharapan, dan iman di dalam Tuhan. Setiap doa dan perjuangan yang diberikan menjadi kekuatan bagi penulis dalam melewati setiap proses dan rintangan. Penulis menyadari bahwa karya sederhana ini tidak akan pernah cukup untuk membalas semua yang telah diberikan, namun kiranya ini menjadi tanda kecil dari rasa syukur dan kasih penulis. Kiranya Tuhan Yesus senantiasa melimpahkan kesehatan, sukacita, damai sejahtera, dan umur panjang kepada Papa dan Mama.

Dosen Pembimbing dan Pembahas

Terima kasih kepada Bapak dan Ibu dosen pembimbing serta penguji atas bimbingan, arahan, dan sarannya sehingga skripsi ini dapat terselesaikan dengan baik.

Keluarga Besar dan Sahabat-sahabatku

Terima kasih untuk seluruh keluarga besar atas segala doa dan dukungan, serta sahabat-sahabat tercinta yang selalu menguatkan dengan kata dan kehadiran.

Almamater Tercinta, Universitas Lampung

SANWACANA

Puji dan syukur penulis panjatkan kepada Tuhan Yesus Kristus atas segala berkat, kasih, dan penyertaan-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul **"Implementasi *Support Vector Machine* Terhadap Klasifikasi Diagnosis Penderita Kanker Payudara"**.

Selama penyusunan skripsi ini, penulis memperoleh bantuan, dukungan, dan arahan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Tuhan Yesus Kristus yang selalu ada dalam setiap proses kehidupan penulis, baik dalam suka maupun duka, yang menjadi tempat penulis bersandar, mencurahkan isi hati, dan memperoleh kekuatan. Kasih dan penyertaan-Nya menjadi sumber pengharapan serta penguat bagi penulis hingga mampu menyelesaikan setiap proses dengan baik.
2. Bapak Dr. Subian Saidi, S.Si., M.Si. selaku Pembimbing I yang senantiasa memberikan arahan, bantuan, motivasi dan saran kepada penulis dalam penyusunan skripsi ini.
3. Ibu Bernadita Herindri Samoedra Utami, S.Si., M.Sc. selaku Pembimbing II yang telah memberikan arahan, bimbingan dan dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
4. Bapak Drs. Nusyirwan, M.Si., selaku Dosen Pembahas yang telah memberikan kritik dan saran yang membangun selama proses penyusunan skripsi.
5. Bapak Prof. Drs. Mustofa, M.A., Ph.D., selaku dosen pembimbing akademik.
6. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika.
7. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
8. Seluruh dosen, staff dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

9. Kedua orang tua penulis tersayang, Bapak Jun Bastian Sitompul dan Ibu Rintauli Purba, serta adik penulis, Graciella Sitompul, yang selalu memberikan kasih sayang, doa yang tiada henti, dukungan, cinta, semangat, serta sumber kekuatan dan pengharapan bagi penulis dalam melewati setiap langkah.
10. Sahabat seperjuangan, Herti, Dafiani, Novi, dan Veny, yang tidak hanya menjadi teman belajar dan teman kuliah, tetapi juga menjadi bagian dari perjalanan yang penuh cerita. Terima kasih atas tawa, canda, dan kebersamaan yang selalu mewarnai setiap proses, bahkan di tengah kesibukan. Segala dukungan, semangat, dan kehadiran kalian menjadi sumber kebahagiaan serta motivasi bagi penulis hingga mampu menyelesaikan skripsi ini.
11. Teman-teman Gereja, Mba Vita, Damai, Reza, Kezia, Felis, Aryana, Enjel, dan lainnya, yang telah menjadi bagian penting dalam pertumbuhan iman penulis. Terima kasih atas doa, dukungan, kebersamaan, serta motivasi rohani yang diberikan, sehingga penulis tetap dikuatkan dalam Tuhan dalam setiap proses.
12. Teman-teman SMA, Winda, Tiara, Rani, dan Rana, terima kasih atas kebersamaan, dukungan, dan persahabatan yang tetap terjaga hingga saat ini.
13. Teman-teman Bidang Kaderisasi Kepemimpinan HIMATIKA dan DINAMIKA FMIPA yang telah menjadi wadah bagi penulis dalam mengembangkan minat, bakat, dan pengalaman.
14. Rekan-Rekan KKN Unila Periode I Tahun 2025 Desa Padang Ratu.
15. Teruntuk diriku, terima kasih telah berjuang, bertahan, dan tidak menyerah dalam setiap proses, serta tetap kuat melewati rasa lelah, ragu, ketakutan, dan berbagai tantangan hingga mampu menyelesaikan skripsi ini. Semua usaha, doa, dan air mata menjadi bukti bahwa penulis mampu melangkah sejauh ini.

Penulis menyadari skripsi ini masih jauh dari sempurna. Oleh karena itu, saran dan kritik sangat diharapkan. Semoga skripsi ini bermanfaat bagi semua pihak.

Bandar Lampung, 7 April 2026
Penulis

Elizabeth Ceasarina Sitompul
NPM. 2217031016

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Kemajuan pesat di bidang teknologi dan ilmu pengetahuan menyebabkan penerapan *machine learning* semakin meluas di berbagai sektor. *Machine learning* memberikan kemudahan bagi pengguna dalam menyelesaikan berbagai permasalahan yang berkaitan dengan analisis data. Baik data berukuran kecil hingga berskala besar (*big data*), semuanya dapat diolah menggunakan metode *machine learning* untuk melakukan analisis seperti klasifikasi, regresi, maupun prediksi.

Klasifikasi merupakan salah satu teknik dalam *machine learning* yang bertujuan untuk mengelompokkan data ke dalam beberapa kategori atau kelas berdasarkan pola dari data sebelumnya (Han, *et al*., 2012). Dalam proses klasifikasi, model dibangun dari data pelatihan (*training*) untuk mempelajari hubungan antara variabel input dan target, sehingga model tersebut dapat memprediksi label kelas dari data baru yang belum pernah dilihat sebelumnya (James, *et al*., 2023). Namun, dalam penerapan metode klasifikasi sering muncul kendala berupa ketidakseimbangan data (*imbalanced* data) (Fitriani, *et al*., 2021).

Imbalanced data adalah kondisi yang terjadi ketika jumlah data pada suatu kelas jauh lebih banyak dibandingkan dengan kelas lainnya. Kelas dengan jumlah data lebih besar disebut *majority class*, sedangkan kelas dengan jumlah data lebih sedikit disebut *minority class* (Barro, *et al*., 2013). Ketidakseimbangan pada data dapat menurunkan kinerja model klasifikasi karena model cenderung lebih fokus mengenali kelas mayoritas dan mengabaikan kelas minoritas (Siringoringo & Jaya, 2018). Oleh karena itu, diperlukan metode klasifikasi yang mampu mengatasi permasalahan tersebut sekaligus memberikan tingkat akurasi tinggi, salah satunya adalah metode *Support Vector Machine* (SVM).

Salah satu algoritma *machine learning* yang paling banyak digunakan dalam permasalahan klasifikasi adalah *Support Vector Machine* (SVM), karena kemampuannya dalam memisahkan data ke dalam kategori tertentu secara optimal dan akurat. *Support Vector Machine* (SVM) adalah salah satu algoritma unggulan dalam bidang *machine learning* karena memiliki performa yang sangat baik dalam menangani permasalahan klasifikasi maupun prediksi. Inti dari metode ini adalah membangun sebuah model yang mampu menemukan batas pemisah optimal (*optimal hyperplane*) antar kelas pada data pelatihan, sehingga data dapat terklasifikasi ke dalam dua atau lebih kelompok yang berbeda. Dengan demikian, SVM dapat digunakan untuk memprediksi kategori suatu data baru secara akurat (Huang, *et al*., 2014).

Dalam penerapannya, SVM menggunakan fungsi *kernel* (*kernel function*) untuk mengatasi permasalahan data yang tidak dapat dipisahkan secara linear. Fungsi *kernel* berperan dalam mentransformasikan data dari ruang input berdimensi rendah ke ruang fitur berdimensi lebih tinggi sehingga data dapat dipisahkan secara linear di ruang tersebut (Hamel, 2009). Proses ini dikenal sebagai *kernel trick*, yang memungkinkan SVM melakukan pemetaan nonlinear tanpa harus menghitung transformasi secara eksplisit (Cortes & Vapnik, 2015). Beberapa jenis *kernel* yang umum digunakan dalam SVM yaitu *kernel* linear, polinomial, *radial basis function* (RBF), dan sigmoid (Huang, *et al*., 2014).

Implementasi metode SVM juga banyak diterapkan dalam bidang medis, salah satunya pada proses klasifikasi diagnosis kanker payudara. Kanker payudara merupakan salah satu jenis kanker yang paling banyak diderita baik di Indonesia maupun di dunia. Penyakit ini menjadi penyebab utama kematian pada wanita setelah kanker paru-paru (American Cancer Society, 2024). Menurut *World Health Organization* (WHO, 2022), kanker payudara menempati urutan pertama sebagai jenis kanker dengan kasus terbanyak di dunia. WHO juga memperkirakan lebih dari 2,3 juta kasus baru kanker payudara terjadi di seluruh dunia, dengan sekitar 670.000 kematian setiap tahunnya. Kanker payudara terjadi akibat pertumbuhan sel abnormal pada jaringan payudara yang dapat bersifat jinak maupun ganas. Identifikasi jenis kanker payudara sejak dini sangat penting untuk menentukan langkah pencegahan dan pengobatan yang tepat, sehingga risiko komplikasi serius dan kematian dapat diminimalkan (Wisudawati, 2021).

Berdasarkan penelitian sebelumnya, yakni Ardiansyah, *et al*., (2023) dalam "Klasifikasi Penyakit Diabetes Menggunakan Metode SVM dan KNN" menunjukkan bahwa SVM mencapai akurasi sempurna 100%, sedangkan KNN hanya 96%. Sementara itu, penelitian Alghifari *et al*., (2025) dalam penelitian berjudul "Comparison of SVM and Naïve Bayes Algorithms in Sentiment Analysis of User Reviews on Bukalapak" menunjukkan bahwa SVM memperoleh akurasi lebih tinggi (84,48%) dibandingkan Naïve Bayes (83,96%). Adapun penelitian yang dilakukan oleh Mutmainah (2021) membandingkan metode *Random Oversampling* (ROS) dan *Random Undersampling* (RUS) pada klasifikasi penyakit stroke. Ketidakseimbangan data antara kelas stroke (*class* 1) dan tidak stroke (*class* 0) ditangani dengan menyamakan distribusi kelas. Hasil penelitian menunjukkan bahwa ROS memberikan performa lebih baik dengan akurasi mencapai 95%, sedangkan RUS hanya memperoleh akurasi sekitar 76%. Secara keseluruhan, hasil-hasil tersebut menguatkan bahwa metode SVM memiliki kemampuan generalisasi yang lebih baik serta tingkat akurasi yang tinggi dalam berbagai kasus klasifikasi, baik dalam konteks medis maupun di luar bidang kesehatan.

Berdasarkan penjabaran di atas, *Support Vector Machine* (SVM) merupakan metode klasifikasi yang sangat andal untuk menganalisis berbagai jenis data. Oleh karena itu, peneliti tertarik menggunakan metode SVM dalam menentukan diagnosis kanker payudara. Proses pengklasifikasian yang akurat ini diharapkan dapat memudahkan dan mendukung penanganan yang tepat bagi penderita kanker payudara di masa yang akan datang.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini meliputi:

1. Menangani ketidakseimbangan kelas pada data pasien kanker payudara dengan menerapkan metode *Random Oversampling* (ROS) sehingga proporsi antara kelas positif dan negatif menjadi lebih seimbang dan model klasifikasi dapat mengenali data minoritas secara lebih optimal.
2. Membangun serta mengevaluasi model klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) dengan beberapa variasi fungsi *kernel* untuk mengetahui *kernel* yang memberikan performa terbaik berdasarkan nilai akurasi dan metrik evaluasi lainnya pada data kanker payudara.

1.3 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini meliputi:

1. Dapat memberikan pemahaman mengenai cara mengatasi ketidakseimbangan data pada kasus kanker payudara menggunakan *Random Oversampling* (ROS) serta mengetahui kinerja metode *Support Vector Machine* (SVM) dalam melakukan klasifikasi setelah dilakukan proses penyeimbangan data.
2. Penelitian ini diharapkan dapat menjadi referensi atau landasan bagi penelitian selanjutnya yang berkaitan dengan penerapan metode klasifikasi dan penanganan *imbalanced data*.

BAB II

TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining merupakan suatu proses menggali pengetahuan atau informasi berharga dari kumpulan data yang sangat besar dan kompleks. Tujuan utamanya adalah menemukan pola, keterkaitan, maupun informasi tersembunyi yang tidak langsung terlihat dalam data, sehingga mampu memberikan pemahaman yang lebih mendalam dan bernilai. Dalam praktiknya, *data mining* memanfaatkan berbagai metode statistik, matematika, serta kecerdasan buatan untuk menganalisis data secara sistematis dan otomatis. Hasil analisis ini dapat dimanfaatkan dalam mendukung pengambilan keputusan, memprediksi tren pasar, meningkatkan efisiensi operasional, hingga merancang strategi bisnis (Rahayu, *et al*., 2024).

Berikut pengelompokan *data mining* (Handoko, 2018):

1. Deskripsi, yaitu proses untuk menggambarkan pola atau karakteristik umum dari data.
2. Estimasi, yaitu menentukan nilai suatu variabel berdasarkan data historis yang sudah diketahui nilainya.
3. Prediksi, yaitu memperkirakan nilai di masa depan dengan memanfaatkan pola atau tren pada data sebelumnya.
4. Klasifikasi, yaitu proses pembagian data ke dalam kategori tertentu berdasarkan atribut atau karakteristik yang dimilikinya.
5. Klastering, yaitu mengelompokkan objek ke dalam kelompok yang memiliki kemiripan tertentu tanpa menggunakan label kelas.
6. Asosiasi, yaitu menemukan hubungan atau pola keterkaitan antar item dalam suatu dataset.

Data mining terdiri atas beberapa tahapan dalam pemerosesannya. Tahapan-tahapan tersebut meliputi sebagai berikut (Han, *et al*., 2012):

1. *Data Integration*
   Data integration adalah proses menggabungkan data dari berbagai sumber database menjadi satu database terpadu. Proses ini umumnya melibatkan penggabungan atribut-atribut, seperti nama, jenis produk, atau nomor pelanggan.

2. *Data Selection*
   Data selection adalah tahap pemilihan data yang berasal dari database operasional, di mana data disaring sesuai dengan kebutuhan dan tujuan penelitian sebelum proses *data mining* dilakukan. Data yang telah dipilih kemudian disimpan pada media atau berkas tersendiri, terpisah dari database operasional awal, sehingga memudahkan pengelolaan dan pemanfaatan data pada tahap selanjutnya.

3. *Prepocessing Data*
   Prepocessing data adalah proses pengolahan data mentah agar menjadi lebih bersih, terstruktur, dan siap digunakan untuk analisis. Tahap ini berfokus pada peningkatan kualitas data dengan cara mengatasi berbagai permasalahan seperti nilai yang hilang dan duplikasi. Melalui tahap ini, data disesuaikan agar memiliki format yang seragam dan relevan, sehingga dapat menghasilkan model analisis yang lebih akurat.

4. *Data Transformation*
   Data yang telah melalui tahap seleksi selanjutnya dikonversi ke dalam bentuk yang sesuai agar dapat diproses menggunakan metode *data mining*. Proses ini bisa meliputi penggabungan atribut, perhitungan nilai baru, atau normalisasi agar data berada dalam skala yang seragam. Dalam tahap ini, salah satu proses yang dilakukan adalah *scaling* data guna memastikan bahwa nilai data numerik berada pada skala yang sama.

   Terdapat dua cara yang digunakan dalam *scaling* data, yaitu:

   a. *Min Max Normalization* merupakan metode penskalaan data yang dilakukan melalui transformasi linear dengan memetakan nilai data ke dalam rentang tertentu berdasarkan nilai minimum dan maksimum data.

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

   dengan:
   x = nilai yang diamati

$x_{\min}$ = nilai x minimum
$x_{\max}$ = nilai x maksimum.

b. *Z-Score Normalization (Standard Scaler)* adalah metode normalisasi data di mana setiap nilai atribut diubah berdasarkan rata-rata (*mean*) dan simpangan baku dari data tersebut.

$$x_{\text{standard}} = \frac{x - \bar{x}}{s} \tag{2}$$

dengan:
x = nilai yang diamati
$\bar{x}$ = nilai rata-rata (mean)
s = simpangan baku.

5. Proses *mining*
Data mining adalah tahap inti yang dilakukan untuk mengekstraksi pengetahuan penting dan tersembunyi dari data. Proses ini melibatkan pemilihan metode *data mining* yang sesuai, seperti *summarization, classification, clustering*, atau *regression*, penggunaan algoritma yang tepat untuk menghasilkan representasi output yang relevan.

6. *Interpretation* (Evaluasi)
Pada tahap ini, pola-pola menarik yang diperoleh melalui teknik *data mining* dievaluasi untuk menilai apakah hipotesis yang diajukan dapat dibuktikan. Hasil analisis dapat berupa pola-pola khas maupun model prediksi yang kemudian diperiksa kesesuaiannya.

2.2 *Machine Learning*

Machine Learning (ML) adalah cabang dari kecerdasan buatan (*Artificial Intelligence* atau AI) yang berfokus pada pengembangan algoritma agar komputer dapat mempelajari pola dari data tanpa harus diprogram secara langsung. Secara umum, ML bertujuan untuk membuat sistem yang dapat mengenali pola dalam data, melakukan prediksi, serta meningkatkan kinerjanya seiring dengan bertambahnya pengalaman. Konsep ini telah merevolusi berbagai bidang, termasuk kesehatan, keuangan, industri, serta teknologi digital (Budiasto & Tallulembang, 2025). Proses *learning* memiliki dua tahapan, yaitu latihan (*training*) yang merupakan tahap proses pembelajaran terhadap suatu data yang telah diketahui kategori dan pengujian

(*testing*) yang merupakan tahapan evaluasi terhadap kinerja model dari hasil pelatihan (Solihin, *et al*., 2022). Algoritma dalam *machine learning* dibedakan berdasarkan empat jenis, antara lain (Permana, *et al*., 2023):

1. *Supervised Learning*
   Supervised learning, yaitu metode pembelajaran terarah yang memanfaatkan data berlabel untuk membangun model yang mampu memprediksi atau mengklasifikasikan data baru.

2. *Unsupervised Learning*
   Unsupervised learning, yaitu metode pembelajaran tak terarah di mana data tanpa label dikelompokkan untuk mengidentifikasi pola atau struktur tersembunyi.

3. *Semi Supervised Learning*
   Semi supervised learning, yaitu gabungan *supervised* dan *unsupervised* yang memanfaatkan data berlabel dan tanpa label untuk meningkatkan akurasi model.

4. *Reinforcement Learning*
   Reinforcement learning, yaitu metode pembelajaran berbasis pengalaman di mana agen belajar menentukan tindakan optimal melalui interaksi dengan lingkungan untuk memaksimalkan kinerja.

2.3 Klasifikasi

Klasifikasi adalah proses pengolahan data untuk mengelompokkan objek ke dalam kategori tertentu. Proses ini membangun model dari *training* data yang kemudian digunakan untuk mengklasifikasikan *testing* data. Klasifikasi merupakan tantangan yang memerlukan penelitian lebih lanjut dengan tujuan menghasilkan prediksi kelas target yang setepat mungkin pada setiap kasus (Rahmat, *et al*., 2021). Proses klasifikasi terdiri dari empat komponen utama, yaitu (Iriadi, 2013):

1. Kelas
   Kelas, yaitu variabel dependen yang bersifat kategorikal dan berfungsi sebagai label pada objek hasil klasifikasi, sering disebut juga sebagai variabel target.

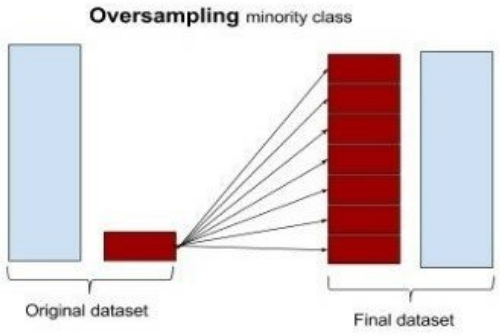
2. Prediktor
   Prediktor, yaitu variabel independen yang menggambarkan karakteristik atribut dari data yang akan diklasifikasikan.

3. *Training dataset*
   Training dataset, yaitu sekumpulan data yang terdiri atas variabel kelas dan prediktor, yang digunakan untuk melatih model agar mampu mempelajari pola hubungan antara atribut dan kelas.

4. *Testing dataset*
   Testing dataset, yaitu data baru yang digunakan untuk mengevaluasi model yang dibangun, sehingga dapat menilai akurasi dan kinerja model dalam klasifikasi.

2.4 *Imbalanced Data*

Imbalanced data terjadi ketika proporsi jumlah data pada masing-masing kelas tidak seimbang, di mana satu kelas memiliki jumlah sampel jauh lebih banyak dibanding kelas lainnya. Kondisi ini dapat menjadi tantangan dalam proses klasifikasi karena model cenderung memprioritaskan kelas mayoritas, sehingga akurasi untuk kelas mayoritas tinggi, namun untuk kelas minoritas rendah (Akbar & Hayaty, 2020). Salah satu pendekatan yang umum digunakan untuk mengatasi masalah ini adalah *Random Oversampling* (ROS). ROS bekerja dengan menambah jumlah sampel pada kelas minoritas secara acak hingga jumlahnya sebanding dengan kelas mayoritas. Proses ini dilakukan dengan menghitung selisih jumlah sampel antar kelas, kemudian menyalin sampel minoritas secara acak sebanyak selisih tersebut ke dalam *training* dataset (Chawla, 2003). Tujuan ROS adalah menciptakan keseimbangan antar kelas sehingga model *machine learning* dapat melakukan proses klasifikasi dengan lebih adil dan menghasilkan prediksi yang lebih akurat pada kedua kelas (Yu, *et al*., 2017).



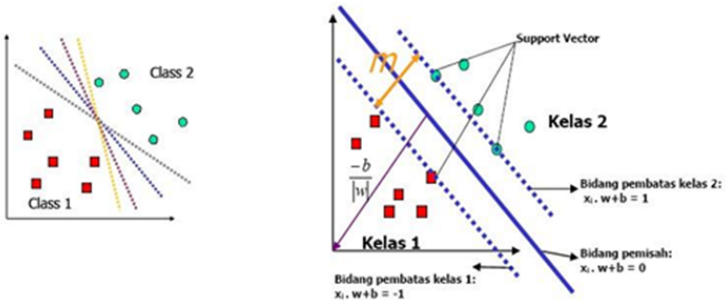
Gambar 1. Proses *Random Oversampling*

2.5 *Hyperparameter Tuning Grid Search*

Parameter merupakan variabel yang dipelajari langsung oleh model selama proses pelatihan, sedangkan *hyperparameter* adalah variabel pengatur yang menentukan cara kerja model dan turut memengaruhi kualitas prediksi, namun nilainya tidak berubah selama proses optimisasi dan tidak bergantung pada data pelatihan (Amalia *et al*., 2022). Untuk mendapatkan performa model yang optimal, diperlukan proses penentuan *hyperparameter* terbaik melalui *hyperparameter tuning*. Salah satu teknik *tuning* yang banyak digunakan adalah *Grid Search*, yaitu metode yang menelusuri berbagai kombinasi *hyperparameter* secara sistematis berdasarkan nilai yang telah ditetapkan sebelumnya. Metode ini membantu menemukan konfigurasi parameter yang memberikan performa terbaik sehingga model mampu menghasilkan prediksi yang lebih akurat pada data uji (Gunawan *et al*., 2020). Keunggulan *Grid Search* terletak pada proses evaluasinya yang terstruktur, yaitu setiap kombinasi diuji dengan prosedur yang sama sehingga hasilnya objektif dan dapat dibandingkan secara adil (Andini *et al*., 2022).

2.6 *Support Vector Machine*

Support Vector Machine (SVM) adalah metode *supervised learning* yang digunakan untuk klasifikasi dan regresi, berlandaskan prinsip *Structural Risk Minimization* (Cortes & Vapnik, 2015), di samping itu SVM juga merupakan salah satu metode dalam *machine learning* yang relatif baru namun memiliki kinerja tinggi dalam berbagai bidang, seperti bioinformatika, klasifikasi teks dan dokumen, serta pengenalan tulisan tangan. Metode ini memiliki ciri khas yaitu kemampuan menemukan fungsi pemisah optimal yang memisahkan dua set data dari dua kelas berbeda. SVM bekerja dengan membangun sebuah *hyperplane* atau bidang pemisah yang memaksimalkan jarak (*margin*) antara kedua kelas tersebut. Proses pencarian *hyperplane* terbaik inilah yang menjadi inti dari metode *Support Vector Machine* (Munawarah, *et al*., , 2016). *Margin* sendiri merupakan jarak antara *hyperplane* dengan data terdekat dari masing-masing kelas.



Gambar 2. Menemukan *hyperplane* terbaik pada *Support Vector Machine*

Gambar 2 menggambarkan konsep pemisahan kelas dalam metode *Support Vector Machine* (SVM). Bagian sebelah kiri menunjukkan beberapa alternatif bidang pemisah (*discrimination boundaries*) yang dapat memisahkan dua kelas data, yaitu kelas -1 yang diwakili oleh kotak merah dan kelas +1 yang diwakili oleh lingkaran hijau. Sementara itu, bagian sebelah kanan memperlihatkan optimal *hyperplane*, yaitu bidang pemisah terbaik yang memiliki margin terbesar antara kedua kelas. *Margin* (m) ini adalah jarak terjauh yang masih mempertahankan pemisahan kedua kelas. Data yang terletak tepat pada batas *margin* tersebut disebut *support vectors*, yang berperan penting dalam menentukan posisi dan orientasi *hyperplane*. Dengan demikian, tujuan dari proses klasifikasi SVM adalah menemukan *hyperplane* yang mampu memaksimalkan *margin* untuk memperoleh akurasi klasifikasi yang optimal (Adinegoro *et al*., 2015). Dalam klasifikasi, tujuan utamanya yaitu menemukan *hyperplane* yang dapat memisahkan kedua kelas.

Misalnya data yang ada direpresentasikan ke dalam bentuk vektor berikut:

$$\vec{\mathbf{d}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\},$$

dengan $x_i \in R$ dan $y_i \in \{-1, 1\}$.
Diasumsikan bahwa data dapat sepenuhnya dipisahkan oleh sebuah *hyperplane* menjadi dua kelas berbeda, yaitu kelas, -1 dan $+1$, yang dapat dirumuskan sebagai berikut (Zaki & Meira, 2014):

$$\mathbf{w^T} \cdot \vec{\boldsymbol{x}} + b = 0 \tag{3}$$

dengan:
w = vektor normal pada *hyperplane*

b = jarak dari *hyperplane* ke titik pusat.

Menurut Cortes dan Vapnik (1995), didapatkan persamaan sebagai berikut:

$$[(\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}}) + b] \geq +1 \text{ untuk } y_{\boldsymbol{i}} = +1 \tag{4}$$

$$[(\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}}) + b] \leq -1 \text{ untuk } y_{\boldsymbol{i}} = -1 \tag{5}$$

dengan:
$\boldsymbol{x_i}$ = himpunan data *training*, $i = 1, 2, \ldots, n$
y_i = label kelas dari x_i.

Persamaan (4) dan (5) dapat disederhanakan menjadi:

$$y_i(\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}} + b) \geq 1, \quad i = 1, 2, 3, \ldots, n. \tag{6}$$

Pemaksimalan jarak terdekat antara *hyperplane* dan pola data dilakukan untuk menentukan *margin* maksimum antar kelas. *Margin* tersebut didefinisikan dengan $d = d_1 + d_2$, di mana nilai *margin* akan mencapai nilai maksimum ketika $d_1 = d_2$. Dengan demikian, *margin* maksimum diperoleh melalui upaya memaksimalkan jarak antara *hyperplane* dengan titik data terdekat, yang dinyatakan sebagai $\frac{1}{\|\vec{\mathbf{w}}\|}$.

$$d = d_1 + d_2 = \frac{1}{\|\vec{\mathbf{w}}\|}(|\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\mathbf{1}} + b||\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\mathbf{2}} + b|) = \frac{2}{\|\vec{\mathbf{w}}\|}. \tag{7}$$

Berdasarkan persamaan tersebut, penentuan *margin* maksimum dapat dilakukan dengan meminimalkan nilai $\|\mathbf{w}\|^2$, secara matematis, persamaan tersebut dirumuskan sebagai berikut:

$$\min \frac{1}{2}\|\vec{\mathbf{w}}\|^2. \tag{8}$$

Proses optimasi dapat diselesaikan dengan menerapkan metode *Lagrange Multiplier* sebagaimana ditunjukkan sebagai berikut :

$$L = \frac{1}{2}\|\vec{\mathbf{w}}\|^2 - \sum_{i=1}^{l} a_i[y_i(\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}} + b) - 1]$$

$$L = \frac{1}{2}\|\vec{\mathbf{w}}\|^2 - \sum_{i=1}^{l} a_i y_i(\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}} + b) - \sum_{i=1}^{l} a_i. \tag{9}$$

Parameter a_i berperan sebagai *Lagrange multiplier* yang bernilai nol atau positif ($a_i \geq 0$). Proses Optimasi dilakukan dengan meminimalkan L terhadap **w** dan b sebagai berikut (Hamel, 2009):

$$\frac{\partial L}{\partial b} = 0$$

$$\sum_{i=1}^{l} a_i y_i = 0 \tag{10}$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0$$

$$\vec{\mathbf{w}} - \sum_{i=1}^{l} a_i y_i \vec{\boldsymbol{x}}_{\boldsymbol{i}} = 0$$

$$\vec{\mathbf{w}} = \sum_{i=1}^{l} a_i y_i \vec{\boldsymbol{x}}_{\boldsymbol{i}}. \tag{11}$$

Selain itu, proses optimasi juga bisa dilakukan dengan memaksimalkan L terhadap a_i melalui substitusi Persamaan (10) dan (11) ke Persamaan (9) sehingga diperoleh sebagai berikut:

$$L = \frac{1}{2}\|\vec{\mathbf{w}}\|^2 - \sum_{i=1}^{l} a_i y_i(\mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}} + b) - \sum_{i=1}^{l} a_i$$

$$L = \frac{1}{2}(\mathbf{w}^T\vec{\mathbf{w}}) - (\sum_{i=1}^{l} a_i y_i \mathbf{w^T} \cdot \vec{\boldsymbol{x}}_{\boldsymbol{i}} + \sum_{i=1}^{l} a_i y_i b - \sum_{i=1}^{l} a_i)$$

$$L = \frac{1}{2}(\sum_{i=1}^{l} a_i y_i \vec{\boldsymbol{x}}_{\boldsymbol{i}} \sum_{j=1}^{l} a_j y_j \vec{\boldsymbol{x}}_{\boldsymbol{j}}) - ((\sum_{i=1}^{l} a_i y_i \vec{\boldsymbol{x}}_{\boldsymbol{i}} \sum_{j=1}^{l} a_j y_j \vec{\boldsymbol{x}}_{\boldsymbol{j}}) + 0 - \sum_{i=1}^{l} a_i)$$

$$L = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \vec{\boldsymbol{x}}_{\boldsymbol{j}} \vec{\boldsymbol{x}}_{\boldsymbol{i}} - (\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \vec{\boldsymbol{x}}_{\boldsymbol{i}} \vec{\boldsymbol{x}}_{\boldsymbol{j}} - \sum_{i=1}^{l} a_i)$$

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \vec{\boldsymbol{x}}_{\boldsymbol{i}} \vec{\boldsymbol{x}}_{\boldsymbol{j}}. \tag{12}$$

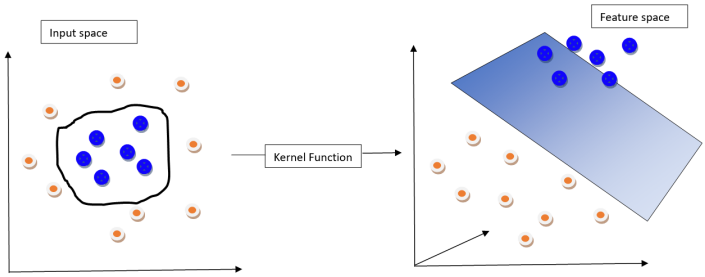
Dimana $a_i \geq 0, \sum_{i=1}^{l} a_i y_i = 0$.

Nilai a_i diperoleh melalui penyelesaian Persamaan (12) yang selanjutnya digunakan untuk menentukan variabel primal, dengan rumus sebagai berikut:

$$\vec{\mathbf{w}_1} = \sum_{i=1}^{l} a_i y_i K(\vec{\boldsymbol{x_i}}\vec{\boldsymbol{x_j}}), b = -\frac{1}{2}(\mathbf{w}^{\boldsymbol{T}}\boldsymbol{x}^{+} + \mathbf{w}^{\boldsymbol{T}}\boldsymbol{x}^{-}). \tag{13}$$

Setelah proses optimasi, nilai $a_i > 0$ yang diperoleh disebut sebagai *support vector*, sementara nilai $a_i = 0$ menunjukkan sisa lainnya. Hasil keputusan pada metode SVM sepenuhnya bergantung pada data yang menjadi *support vector* tersebut. Namun, pada kenyataannya, data yang benar-benar dapat dipisahkan secara linear jarang dijumpai di dunia nyata. Dengan demikian, untuk mengatasi pola data yang bersifat nonlinear, SVM memanfaatkan fungsi *kernel* untuk memetakan data asli ke ruang fitur berdimensi tinggi (*feature space*), sehingga pemisahan antar kelas dapat dilakukan secara linear di ruang tersebut. Dengan kata lain, permasalahan nonlinear dapat diselesaikan menggunakan pendekatan *kernel trick*, yaitu suatu teknik yang memungkinkan data diubah menjadi linear di ruang fitur tanpa perlu melakukan transformasi eksplisit (Hamel, 2009). Metode *kernel trick* dapat dinyatakan melalui persamaan berikut:

$$K(\vec{\boldsymbol{x_i}}, \vec{\boldsymbol{x_j}}) = \varnothing(\vec{\boldsymbol{x_i}}).\varnothing(\vec{\boldsymbol{x_j}}). \tag{14}$$



Gambar 3. Proses Pemetaan Data dari Ruang Asli ke *Feature Space*

Gambar 3 menunjukkan proses transformasi data menggunakan *kernel function*. Pada bagian kiri (*input space*), terlihat bahwa data dari dua kelas yang ditandai dengan warna biru dan *oranye* tidak dapat dipisahkan secara linear karena berada dalam ruang berdimensi rendah. Melalui penerapan *kernel function*, data tersebut kemudian dipetakan ke ruang berdimensi lebih tinggi (*feature space*), seperti yang terlihat pada bagian kanan gambar. Di ruang baru ini, data dari kedua kelas menjadi lebih mudah dipisahkan secara *linear* oleh sebuah *hyperplane*. Proses ini menggambarkan prinsip utama dari *kernel trick* dalam metode *Support Vector Machine* (SVM), yaitu

mentransformasikan data nonlinear menjadi linear di ruang fitur berdimensi lebih tinggi. Pemetaan data ke ruang berdimensi lebih tinggi dapat dinyatakan melalui notasi matematika sebagai berikut:

$$\phi; R^d \rightarrow R^q,\ d < q. \tag{15}$$

Umumnya, bentuk eksplisit dari transformasi ϕ umumnya tidak diketahui. Oleh karena itu, proses pemetaan tersebut digantikan dengan penggunaan fungsi *kernel* yang dinyatakan sebagai $K = (\boldsymbol{x_i}, \boldsymbol{x_j})$. Melalui fungsi *kernel* ini, hasil klasifikasi dapat ditentukan berdasarkan persamaan berikut:

$$\begin{aligned} f(\phi(\vec{\boldsymbol{x_i}})) &= sign(\mathbf{w^T} \cdot \phi(\vec{\boldsymbol{x_i}}) + b) \\ f(\phi(\vec{\boldsymbol{x_i}})) &= sign(\sum_{i=1}^{n} a_i y_i\, \phi(\vec{\boldsymbol{x_i}}) \cdot \phi(\vec{\boldsymbol{x_j}}) + b) \\ f(\phi(\vec{\boldsymbol{x_i}})) &= sign(\sum_{i=1}^{n} a_i y_i\, K(\vec{\boldsymbol{x_i}}, \vec{\boldsymbol{x_j}}) + b) \end{aligned} \tag{16}$$

dengan :

x_i = data *input* x baris ke-i

x_j = data *input* x kolom ke-j

y_i = kelas *output* baris ke-i

b = bias

a_i = *support vector*

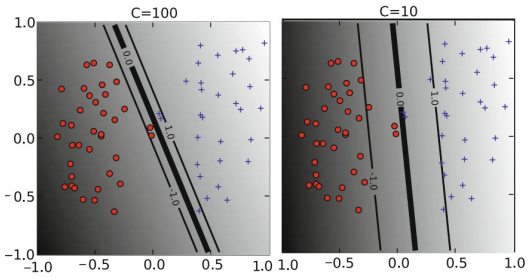
$sign$ = notasi (+ atau –), jika $f(\phi(x)) > 0$ maka data dimasukkan ke kelas $+1$, sedangkan jika $f(\phi(x)) < 0$ maka data dimasukkan ke kelas -1.

Beberapa jenis fungsi *kernel* yang umum diterapkan pada metode SVM antara lain sebagai berikut:

1. *Kernel* Linear
   Kernel linear merupakan fungsi *kernel* yang paling sederhana dan digunakan ketika data dapat dipisahkan secara linear. Pada *kernel* ini, proses pemetaan data tidak dilakukan ke ruang fitur berdimensi lebih tinggi, sehingga model bekerja langsung pada ruang aslinya. *Kernel* linear juga memanfaatkan parameter *cost* (C), yaitu parameter regulasi yang mengatur seberapa besar model memberikan penalti terhadap kesalahan klasifikasi (Praghakusma &

Charibaldi, 2021). Rumus *kernel* linear dituliskan sebagai berikut:

$$K(\vec{x_i}, \vec{x_j}) = \vec{x_i}.\vec{x_j}. \tag{17}$$



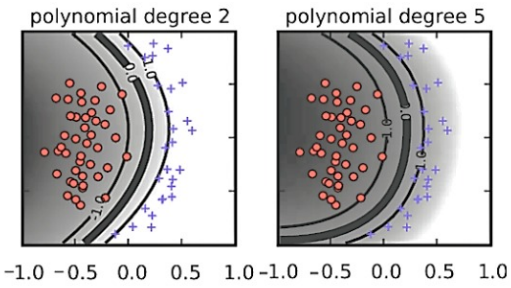
Gambar 4. *Kernel* Linear dengan Parameter C

Pada Gambar 4, dapat dilihat bahwa penggunaan nilai C yang rendah (contohnya 10) membuat SVM menghasilkan margin yang relatif lebar serta masih mentoleransi beberapa titik data yang berada dekat dengan garis pemisah, sehingga tingkat kesalahan marginnya tetap rendah. Namun, ketika nilai C dibuat lebih tinggi (seperti 100), model berfokus untuk mengurangi kesalahan secara lebih tegas, yang berdampak pada margin yang semakin sempit dan berkurangnya toleransi terhadap data yang berada di dekat batas keputusan (Ben-Hur & Weston, 2009).

2. *Kernel* Polinomial

   Kernel polinomial digunakan ketika data tidak dapat dipisahkan menggunakan garis lurus pada ruang awal. Melalui penggunaan derajat tertentu, *kernel* ini memproyeksikan data ke ruang dengan dimensi yang lebih tinggi sehingga pola hubungan yang bersifat nonlinier dapat dimodelkan dengan lebih baik. Pada *kernel* polinomial, terdapat parameter *cost* (C) dan *degree*, dimana parameter *degree* yang menentukan derajat polinomial yang digunakan. Nilai *degree* ini pada dasarnya menunjukkan tingkat kompleksitas kurva pemisah yang dibentuk untuk menghasilkan *hyperplane* yang mampu memisahkan data secara lebih efektif (Praghakusma & Charibaldi, 2021). Rumus *kernel* polinomial dituliskan sebagai berikut:

$$K(\vec{x_i}, \vec{x_j}) = (\vec{x_i}.\vec{x_j} + 1)^d. \tag{18}$$

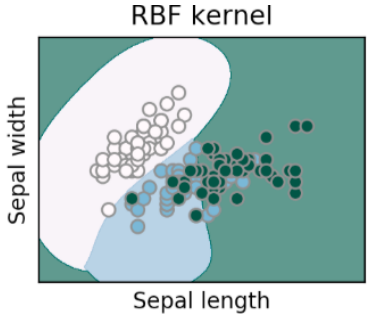


Gambar 5. *Kernel* Polinomial dengan Parameter C dan *Degree*

Pada Gambar 5, Parameter derajat (*degree*) pada *kernel* polinomial berperan dalam mengatur tingkat fleksibilitas model. Semakin besar nilai *degree* yang digunakan, semakin kompleks dan lentur bentuk *decision boundary* yang dapat dibentuk oleh model dalam proses klasifikasi (Ben-Hur & Weston, 2009).

3. *Kernel Radial Basis Function* (RBF)
   Kernel RBF digunakan ketika data tidak dapat dipisahkan secara linear. *Kernel* ini menggunakan parameter *gamma* (γ) dan *cost* (C), di mana γ merupakan parameter yang mengatur seberapa jauh pengaruh satu sampel data terhadap bentuk keputusan model, sehingga nilai *gamma* menentukan tingkat sensitivitas SVM terhadap pola yang bersifat nonlinear.(Praghakusma & Charibaldi, 2021) . Rumus *kernel* RBF dituliskan sebagai berikut:

$$K(\vec{\boldsymbol{x_i}}, \vec{\boldsymbol{x_j}}) = e^{-\left(\frac{\|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j\|^2}{2\sigma^2}\right)}, \quad \sigma > 0. \tag{19}$$



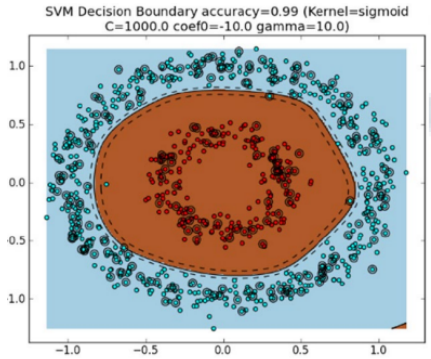
Gambar 6. *Kernel* RBF dengan Parameter C dan *Gamma*

Pada Gambar 6, menunjukkan *kernel* RBF menghasilkan batas keputusan yang melengkung dan fleksibel sehingga mampu mengikuti pola sebaran

data dengan baik. Pada gambar terlihat bahwa *kernel* ini dapat menyesuaikan bentuk batas pemisah secara optimal, sehingga pemisahan antar kelas menjadi lebih jelas. Hal ini menunjukkan bahwa *kernel* RBF efektif dalam menangani data berpola nonlinier karena mampu merepresentasikan hubungan kompleks antar variabel (Pedregosa, *et al*., 2011).

4. *Kernel* sigmoid
   Kernel sigmoid merupakan fungsi *kernel* yang memiliki bentuk serupa dengan fungsi aktivasi pada jaringan saraf tiruan. *Kernel* ini menggunakan parameter *gamma* (γ) dan *cost* (C), di mana γ merupakan parameter yang mengatur seberapa jauh pengaruh satu sampel data terhadap bentuk keputusan model, sehingga nilai *gamma* menentukan tingkat sensitivitas SVM terhadap pola yang bersifat nonlinear.(Praghakusma & Charibaldi, 2021). Rumus *kernel* sigmoid dituliskan sebagai berikut:

$$K(\vec{\boldsymbol{x_i}}, \vec{\boldsymbol{x_j}}) = tanh(y\vec{\boldsymbol{x_i}}.\vec{\boldsymbol{x_j}} - \delta). \tag{20}$$



Gambar 7. *Kernel* Sigmoid dengan Parameter C dan *Gamma*

Pada Gambar 7, menunjukkan bahwa penggunaan nilai *gamma* yang terlalu tinggi pada *kernel* sigmoid berdampak pada menurunnya tingkat akurasi klasifikasi. Namun, efek ini tetap dipengaruhi oleh jumlah fitur yang digunakan. Jika jumlah fitur banyak, *gamma* yang optimal cenderung lebih kecil, sedangkan pada jumlah fitur yang lebih sedikit, *gamma* yang digunakan biasanya lebih besar (Al-Mjibli *et al*., 2020).

2.7 Evaluasi Model

Tahapan evaluasi model dilakukan untuk menilai kemampuan metode *Support Vector Machine* (SVM) dalam mengklasifikasikan data secara akurat. Penilaian ini dilakukan dengan mencocokkan hasil prediksi yang dihasilkan model dengan label kelas sebenarnya pada data pengujian. Proses evaluasi menggunakan *confusion matrix* yang menunjukkan jumlah prediksi yang tepat maupun keliru untuk setiap kelas. Berdasarkan matriks tersebut, selanjutnya dihitung sejumlah indikator kinerja, antara lain *accuracy*, *precision*, *recall*, dan *F1-score*, yang digunakan untuk menggambarkan tingkat keberhasilan model secara menyeluruh (Atmanegara & Handayani, 2024).

Tabel 1. *Confusion matrix*

Kelas Asli	Kelas Prediksi	
	Prediksi Positif	Prediksi Negatif
Aktual Positif	*True Positive (TP)*	*False Positif (FP)*
Aktual Negatif	*False Negatif (FN)*	*True Negatif (TN)*

a. *True Positive* (TP) , adalah data yang diprediksi positif dan data sebenarnya adalah positif.

b. *True Negative* (TN), adalah data yang diprediksi negatif dan data sebenarnya adalah negatif.

c. *False Positive* (FP) , adalah data yang diprediksi positif dan data sebenarnya adalah negatif.

d. *False Negative* (FN), adalah data yang diprediksi negatif dan data sebenarnya adalah positif.

Berdasarkan tabel *confusion matrix* di atas, dapat dilakukan perhitungan untuk menilai kinerja model menggunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall* dan *F1-score*. Berikut adalah perhitungan dari masing-masing model (Saputro & Sari, 2020):

a. *Accuracy*, yaitu ukuran yang digunakan untuk menilai sejauh mana model mampu memberikan hasil klasifikasi yang benar pada seluruh data yang diuji. Nilai ini diperoleh dari perbandingan antara jumlah prediksi yang sesuai dengan kondisi sebenarnya dan total keseluruhan data. Semakin besar nilai *accuracy*, semakin baik kemampuan model dalam melakukan klasifikasi.

$$Accuracy = \frac{TP + TN}{Total}. \tag{21}$$

b. *Precision*, metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan model dalam memprediksi kelas positif, dengan melihat perbandingan antara jumlah prediksi positif yang benar dan seluruh data yang diprediksi sebagai positif.

$$Precision = \frac{TP}{FP + TP}. \tag{22}$$

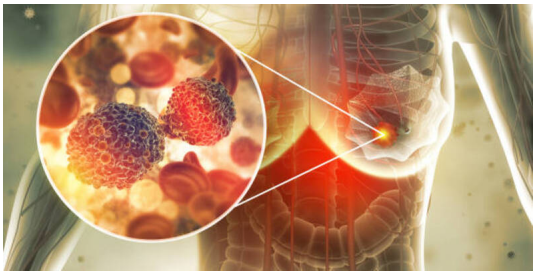
c. *Recall*, metrik evaluasi yang digunakan untuk menilai kemampuan model dalam mengenali data dengan kelas aktual positif, dengan melihat perbandingan antara jumlah prediksi positif yang benar dan seluruh data yang benar-benar termasuk ke dalam kelas positif.

$$Recall = \frac{TP}{FN + TP}. \tag{23}$$

d. *F1-score*, digunakan untuk menilai keseimbangan kinerja model berdasarkan nilai *precision* dan *recall* secara bersamaan. Metrik ini dihitung menggunakan rata-rata harmonik, sehingga nilai yang lebih rendah antara *precision* dan *recall* akan memberikan pengaruh yang lebih besar. Oleh karena itu, *F1-score* hanya akan bernilai tinggi apabila kedua metrik tersebut juga menunjukkan hasil yang baik.

$$F1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall}. \tag{24}$$

2.8 Kanker Payudara (*Breast Cancer*)



Gambar 8. Ilustrasi Kanker Payudara

Kanker payudara (*breast cancer*) merupakan kondisi terjadinya pertumbuhan sel pada jaringan payudara yang bersifat abnormal dan tidak mengikuti mekanisme pertumbuhan sel normal. Kanker payudara juga penyebab utama kematian pada wanita (Wisudawati, 2021). Menurut Yayasan Kanker Indonesia (YKI), kanker ditandai dengan pembelahan sel yang berlangsung cepat, berkelanjutan, dan tidak terkendali sehingga dapat merusak jaringan di sekitarnya. Pada dasarnya, pertumbuhan sel abnormal dapat berbentuk tumor, yang diklasifikasikan menjadi dua jenis, yaitu jinak (*benign*) dan ganas (*malignant*). Tumor jinak tidak menyebar ke jaringan lain, sedangkan tumor ganas memiliki kemampuan untuk menyebar dan menyerang jaringan tubuh lainnya. Tumor ganas inilah yang dikenal sebagai kanker (Putra, 2015).

Secara umum, faktor risiko kanker payudara dapat dikelompokkan ke dalam faktor yang bersifat pasti dan faktor yang masih bersifat potensial. Faktor risiko yang telah diketahui dengan jelas antara lain usia, adanya riwayat kanker payudara dalam keluarga, kondisi reproduksi seperti usia saat menstruasi pertama dan kehamilan, serta riwayat penyakit payudara jinak yang pernah dialami sebelumnya. Sementara itu, faktor lain yang mungkin meningkatkan peluang terjadinya kanker payudara adalah penggunaan estrogen dari luar tubuh, pemakaian kontrasepsi oral dalam jangka panjang, kelebihan berat badan atau obesitas, pola makan dengan kadar lemak tinggi, kebiasaan mengonsumsi minuman beralkohol, dan kebiasaan merokok. Meskipun tidak semua faktor tersebut langsung menimbulkan kanker, kombinasi dari beberapa faktor dapat memperbesar risiko seorang wanita untuk mengalaminya (Hero, 2021).

BAB III

METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester ganjil tahun ajaran 2025/2026 bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu pengetahuan Alam Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan pada penelitian ini merupakan data sekunder mengenai penderita kanker payudara yang diambil dari *website* `https://www.kaggle.com/datasets/wasiqaliyasir/breast-cancer-dataset`. Dengan jumlah data sebanyak 569 sampel dan 32 variabel yang terdiri dari satu variabel Y berupa diagnosis dan 30 variabel X yang bersifat numerik, yaitu hasil pengukuran karakteristik sel yang mencakup *radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry*, dan *fractal dimension* masing-masing dalam tiga bentuk yaitu *mean*, *standard error* (SE), dan *worst* (nilai terburuk), serta satu variabel ID sebagai penanda data.

Variabel yang digunakan dalam penelitian ini meliputi Diagnosis, yaitu klasifikasi kanker payudara yang terdiri dari *Malignant* (M) yang menunjukkan tumor bersifat ganas, dan *Benign* (B) yang menunjukkan tumor bersifat jinak. Selain itu, setiap karakteristik inti sel dalam dataset ini terdiri dari tiga jenis pengukuran, yaitu *mean* yang menunjukkan nilai rata-rata dari karakteristik tersebut, *standard error* (se) yang menggambarkan tingkat variasi atau kesalahan standar dari pengukuran, serta *worst* yang menunjukkan nilai terbesar atau kondisi paling ekstrem dari karakteristik tersebut. Adapun variabel karakteristik inti sel yang digunakan meliputi

Radius yang menunjukkan jari-jari inti sel, *Texture* yang menggambarkan tekstur permukaan inti sel, *Perimeter* yang menyatakan panjang keliling inti sel, *Area* yang menunjukkan luas area inti sel, *Smoothness* yang menunjukkan tingkat kehalusan tepi sel, *Compactness* yang menggambarkan tingkat kepadatan bentuk inti sel, *Concavity* yang menunjukkan tingkat kecekungan kontur inti sel, *Concave Points* yang menyatakan jumlah titik cekung pada kontur inti sel, *Symmetry* yang menunjukkan tingkat kesimetrian inti sel, dan *Fractal Dimension* yang menggambarkan kompleksitas kontur inti sel.

3.3 Metode Penelitian

Penelitian ini dilakukan melalui beberapa tahapan, yaitu sebagai berikut:

1. Analisis Deskriptif
   Tahap awal penelitian dilakukan dengan analisis deskriptif untuk memperoleh gambaran umum mengenai dataset kanker payudara yang digunakan. Pada tahap ini, distribusi data berdasarkan kelas jinak (*benign*) dan ganas (*malignant*) divisualisasikan menggunakan diagram lingkaran. Selain itu, dilakukan analisis eksploratif dengan menghitung ukuran statistik deskriptif sederhana, seperti nilai rata-rata, median, simpangan baku, nilai minimum, dan maksimum pada setiap variabel penelitian.

2. *Preprocessing Data*

   a. *Cleaning* Data $\rightarrow$ memastikan dataset terbebas dari nilai hilang (*missing value*) maupun data duplikat.

   b. *Scaling* Data $\rightarrow$ melakukan penyesuaian skala pada data numerik menggunakan metode *standard scaler* agar setiap fitur berada pada skala yang sama.

   c. *Handling Data Categorical* $\rightarrow$ mengubah variabel kategorik pada atribut diagnosis jinak dan ganas ke dalam bentuk numerik menggunakan teknik *label encoding*.

3. *Handling Imbalance Data*
   Untuk mengatasi permasalahan ketidakseimbangan data (*imbalance*) dengan menerapkan metode *Random Oversampling* (ROS). Metode ini dilakukan dengan

menambah jumlah data pada kelas minoritas melalui penggandaan data secara acak, sehingga proporsi antar kelas menjadi lebih seimbang.

4. *Splitting Data*
   Dataset selanjutnya dipisahkan ke dalam dua kelompok, yaitu data latih (*training*) dan data uji (*testing*), dengan menerapkan tiga skema pembagian sebagai berikut:

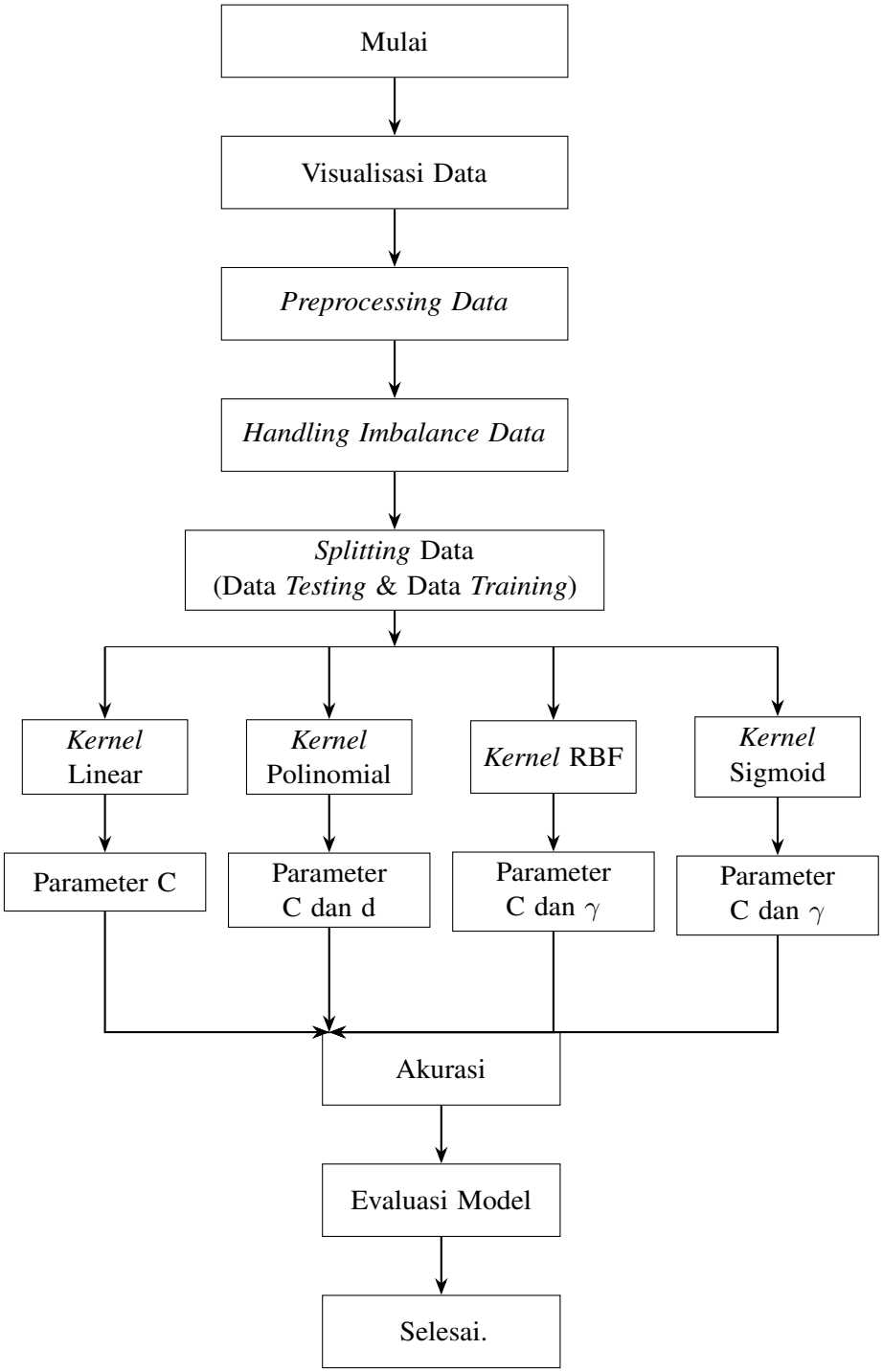
   a. 70% data latih dan 30% data uji

   b. 80% data latih dan 20% data uji

   c. 90% data latih dan 10% data uji.

5. Membangun Model *Support Vector Machine* (SVM)
   Pada tahap ini, model klasifikasi dibangun dengan menerapkan metode *Support Vector Machine* (SVM) serta proses *hyperparameter tuning* untuk memperoleh konfigurasi parameter terbaik dan optimal. Jenis *kernel* yang digunakan meliputi *kernel* Linear, Polinomial, Sigmoid, dan *Radial Basis Function* (RBF).

6. Evaluasi Model
   Model yang telah dihasilkan kemudian diuji untuk menilai kinerja klasifikasi yang dihasilkan. Proses evaluasi dilakukan dengan menggunakan *confusion matrix* serta beberapa metrik pendukungnya, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*.



Gambar 9. Diagram Alir Penelitian

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulam

Berdasarkan hasil penerapan metode *Support Vector Machine* (SVM) pada data penderita kanker payudara, dapat disimpulkan sebagai berikut:

1. Penerapan teknik *Random Oversampling* (ROS) mampu menangani ketidakseimbangan kelas pada data pasien kanker payudara. Ketidakseimbangan tersebut terjadi karena jumlah data pada kelas kanker jinak (*Benign*) lebih banyak dibandingkan dengan kelas kanker ganas (*Malignant*). Melalui metode ROS, jumlah data pada kelas minoritas (kanker ganas) ditingkatkan dengan cara melakukan duplikasi sampel secara acak hingga jumlahnya menjadi sebanding dengan kelas mayoritas (kanker jinak). Dengan distribusi data yang lebih seimbang, model klasifikasi dapat mempelajari pola dari masing-masing kelas secara lebih optimal sehingga kemampuan model dalam mengenali data minoritas menjadi lebih baik.

2. Hasil dari pengujian menunjukkan bahwa *kernel* RBF memiliki kinerja paling optimal dengan akurasi tertinggi sebesar 97,22% pada parameter C = 10, 50, dan 100 serta $\gamma = 0{,}001$. Evaluasi menggunakan *Confusion Matrix* pada pembagian data latih 90% dan data uji 10% menghasilkan nilai akurasi, *recall*, *precision*, dan *F1-score* masing-masing sebesar 97,22% yang menunjukkan bahwa *kernel* RBF memiliki performa terbaik dibandingkan *kernel* lainnya. Selain itu, hasil analisis deskriptif menunjukkan bahwa variabel *area* dan *perimeter* memiliki nilai rata-rata serta sebaran data yang tinggi, sehingga menjadi faktor utama dalam membedakan sel kanker payudara jinak dan ganas, serta mendukung keunggulan *kernel* RBF dalam menghasilkan prediksi yang akurat, konsisten, dan stabil. Dengan menggunakan konfigurasi *kernel* tersebut, diperoleh parameter $\mathbf{w}$ dan b, yakni sebagai berikut:

$\mathbf{w}_{\text{radius\_mean}} = 377,7713,\ \mathbf{w}_{\text{texture\_mean}} = 340,6203,\ \mathbf{w}_{\text{perimeter\_mean}} = 2851,3717,$
$\mathbf{w}_{\text{area\_mean}} = 37057,1680,\ \ldots,\ \mathbf{w}_{\text{fractal\_dimension\_worst}} = 2,8828,\ b = 0,5153.$

5.2 Saran

Penelitian selanjutnya disarankan untuk menggunakan dataset dengan kualitas dan validitas yang lebih terjamin, seperti data klinis yang bersumber langsung dari rumah sakit atau institusi kesehatan, agar hasil klasifikasi yang diperoleh lebih akurat dan relevan secara medis. Selain itu, juga disarankan untuk menerapkan metode klasifikasi lain atau mengombinasikan beberapa metode (*hybrid method*) guna membandingkan dan meningkatkan kinerja model dalam melakukan klasifikasi kanker payudara.

DAFTAR PUSTAKA

Adinegoro, A., Atmaja, R. D., Purnamasari, R., Prodi, S., Telekomunikasi, T., & Elektro, F. T. 2015. Deteksi Tumor Otak dengan Ektrasi Ciri & Feature Selection mengunakan Linear Discriminant Analysis (LDA) dan Support Vector Machine (SVM). *e-Proceeding Engineering* . **2**(2): 2532-2539.

Akbar, K., & Hayaty, M. 2020. Data Balancing untuk Mengatasi Imbalance Dataset pada Prediksi Produksi Padi. *Information Technology Journal of UMUS*. **2**(2): 1-14.

Alghifari, M. Y., Sanjaya, M. R., Indah, D. R., & Ruskan, E. L. 2025. Comparison of SVM and Naive Bayes Algorithms in Sentiment Analysis of User Reviews on Bukalapak. *INOVTEK Polbeng-Seri Informatika*. **10**(3): 1623-1633.

Al-Mejibli, I. S., Alwan, J. K., & Abd, D. H. 2020. The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering*. **10**(5): 5497–5506.

Amalia, H. 2018. Perbandingan Metode Data Mining Svm Dan Nn Untuk Klasifikasi Penyakit Ginjal Kronis. *Jurnal PILAR Nusa Mandiri*. **14**(1): 1-6.

Amalia, Radhi, M., Sitompul, D.R.H., Sinurat, S.H., & Indra, E. 2022. Prediksi Harga Mobil Menggunakan Algoritma Regresi dengan Hyperparameter Tuning. *Jurnal Sistem Informasi dan Ilmu Komputer Prima*. **4**(2): 28–32.

American Cancer Society. 2024. Breast Cancer Facts & Figures 2024-2025.https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf

Andini, E., Faisal, M.R., Herteno, R., Nugroho, R.A., Abadi, F., & Muliadi. 2022. Peningkatan Kinerja Prediksi Cacat Software dengan Hyperparameter Tuning pada Algoritma Klasifikasi Deep Forest. Jurnal Mnemonic. 5(2): 119–27.

Ardiansyah, A., Telaumbanua, E. C., Gultom, A. S., & Limbong, A. A. 2024. Klasifikasi Penyakit Diabetes Menggunakan Metode SVM Dan KNN. *Jurnal Penelitian Rumpun Ilmu Teknik*. **3**(1): 77-83.

Atmanegara, R.C.,& Handayani, W. 2024. Customer Churn Analysis Using Machine Learning to Improve Customer Retention on Vissie Net. *International Journal of Scientific Research and Management*. **12**(09): 7379–7387.

Barro, R. A., Sulvianti, I. D., dan Afendi, F. M. 2013. Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu. *Journal of Statistics*. **1**(1): 23-39.

Ben-Hur, A., & Weston, J. 2009. *A user's guide to support vector machines*. Totowa, NJ: Humana Press.

Budiasto, J., & Tallulembang, T. M. 2025. *Machine Learning untuk Pemula: Konsep dan Implementasi*. Penerbit Buku Indonesia. Jakarta.

Chawla, N. V. 2003. C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. *In Proceedings of the ICML*. **3**(66): 1-9.

Cortes, C., & Vapnik, V. 1995. Support-vector networks. *Journal Machine learning*. **20**(3): 273-297.

Fitriani, R. D., Yasin, H., & Tarno, T. 2021. Penanganan klasifikasi kelas data tidak seimbang dengan random oversampling pada naive bayes (Studi kasus: Status peserta KB IUD di Kabupaten Kendal). *Jurnal Gaussian*. **10**(1): 11-20.

Gunawan, M. I., Sugiarto, D., & Mardianto, I. 2020. Peningkatan Kinerja Akurasi prediksi penyakit diabetes mellitus menggunakan metode grid Seacrh Pada

algoritma logistic regression. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*. **6**(3): 280-284.

Hamel, L. 2009. *Knowledge Discovery with Support Vector Machines*. Boken-New Jersey. Canada.

Han, J., Kamber. M., & Pei, J. 2012. *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann Publishers. Amsterdam.

Handoko, K. 2018. Pengelompokkan Data Mining Pada Jumlah Penumpang Di Bandara Hang Nadim. *Computer Based Information System Journal*. ***6***(2): 60-68.

Hero, S. K. 2021. Faktor Risiko Kanker Payudara. *Jurnal Medika Hutama*. **3**(1): 1533-1538.

Hikmayanti, H., Nurmasruriyah, A. F., Fauzi, A., Nurjanah, N., & Rani, A. N. 2024. Performance Comparison of Support Vector Machine Algorithm and Logistic Regression Algorithm. *International Journal of Artificial Intelligence Research*. **7**(1): 2579-7298.

Huang, M. L., Hung, Y. H., Lee, W. M., Li, R. K., & Jiang, B. R. 2014. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. *The Scientific World Journal*. 14(1): 1-10.

Iriadi, N. 2013. Komparasi Algoritma Klasifikasi Data Mining Dalam Penentuan Resiko Kredit Pada Koperasi Serba Usaha. *Paradigma*. **15**(2): 192-204.

James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. *An introduction to statistical learning*. New York: springer.

Munawarah, R., Soesanto, O., & Faisal, M. R. 2016. Penerapan Metode Support Vector Machine Pada Diagnosa Hepatitis. *Klik-Kumpulan Jurnal Ilmu Komputer*. **3**(1): 103-113.

Mutmainah, S. 2021. Penanganan Imbalance Data pada Klasifikasi Kemungkinan Penyakit Stroke. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*. **1**(1): 10-16.

Natsir, F. M. 2024. Analisis Deteksi Dini Penyakit Jantung dengen Pendetekatan Support Vector Machine pada Data Pasien (skripsi). Teknik Informatika FT Universitas Muhammadiyah Makkasar, Makasar.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. **12**: 2825–2830.

Permana, A. A., Wahyuddin, S., Santoso, L. W., Wibowo, G. W. N., Wardhani, A. K., Rahmadeni, Wahidin, A. J., Yuliastuti, G. E., Elisawati, Wijayanti, R. R., & Abdurraasyid. 2023. *Machine Learning*. Padang: PT Global Eksekutif Teknologi.

Praghakusma, A. Z., & Charibaldi, N. 2021. Komparasi fungsi kernel metode Support Vector Machine untuk analisis sentimen Instagram dan Twitter (Studi Kasus: Komisi Pemberantasan Korupsi). *Jurnal Sarjana Teknik Informatika*. **9**(2): 33–42.

Putra, Sitiatava Rizema. 2015. *Buku Lengkap Kanker Payudara*. Yogyakarta: Laksana.

Rahayu, P. W., Sudipa, I. P., Suryani., Surachman, A., Ridwan, A., Darmawiguna, I. G. M., & Maysanjaya. I. M, D. 2024. *Buku Ajaran Data Mining*. PT. Sonpedia Publishing Indonesia. Jambi.

Rahmat, H., Iwan Tri Riyadi, Y., Azizul Azhar, R., & Ansari Saleh, A. 2021. Generalized Normalized Euclidean Distance Based Fuzzy Soft Set Similarity for Data Classification. *Computer Systems Science & Engineering*. **38**(1): 119-130.

Saputro, I.W., & Sari, B.W. 2020. Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*. **6**(1): 1-11.

Siringoringo, R., & Jaya, I.K. 2018. Ensemble Learning dengan metode Smote Bagging pada Klasifikasi Data Tidak Seimbang. *Information System Development*. **3**(2).

Solihin, A., Mulyana, D. I., & Yel, M. B. 2022. Klasifikasi jenis alat musik tradisional Papua menggunakan metode transfer learning dan data augmentasi. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*. **5**(2): 36-44.

Song, T., Wang, Y., Du, W., Cao, S., Tian, Y., & Liang, Y. 2017. The method for breast cancer grade prediction and pathway analysis based on improved multiple kernel learning. *Journal of Bioinformatics and Computational Biology*. 15(01): 2957-9511.

Wisudawati, L. M. 2021. Klasifikasi Tumor Jinak dan Tumor Ganas pada Citra Mammogram Menggunakan Gray Level Co-Occurrence Matrix (GLCM) dan Support Vector Machine (SVM). *Jurnal Ilmiah Informatika Komputer*. **26**(2): 176-186.

World Health Organization. 2022. Breast cancer. `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`

Yu, D., Hu, J., Tang, Z., & Shen, H. 2017. Neurocomputing Improving protein ATP binding residues prediction by boosting SVMs with random under sampling. *Journal of Neurocomputing*. **104** :180–190.

Zaki, M.J., & Meira Jr, W. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.