PENANGANAN *IMBALANCE* DATA DENGAN *RANDOM OVERSAMPLING* (ROS) PADA KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM)

(Skripsi)

Oleh

DHIFA ZHAFIRAH

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG

2023

ABSTRACT

HANDLING OF IMBALANCE DATA WITH RANDOM OVERSAMPLING (ROS) IN CLASSIFICATION OF DIABETIC PATIENTS USING SUPPORT VECTOR MACHINE (SVM)

By

Dhifa Zhafirah

Diabetes mellitus is a health problem that often occurs in Indonesia, especially in Lampung Province. This disease is a condition in which the body does not produce enough or use the insulin hormone that carries glucose into the body's cells. The purpose of this research is to create a machine-learning model that can detect diabetes early using a Support Vector Machine (SVM). However, the dataset used in the research has data imbalance problems. Therefore, Random Oversampling (ROS) is used to overcome this problem. The results obtained from this study, ROS is able to handle imbalance data so that the accuracy value obtained reaches 96.43% (excellent classification) with the C-Classification model and Radial Basis Function (RBF) kernel, as well as sigma one and cost one parameter for the training data scheme 90% and 10% testing data. This accuracy value increases sharply compared to without ROS, which is only around 76%.

Keywords : Diabetes Melitus, Imbalance Data, Random Oversampling, Support Vector Machine

ABSTRAK

PENANGANAN *IMBALANCE* DATA DENGAN *RANDOM OVERSAMPLING* (ROS) PADA KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM)

Oleh

Dhifa Zhafirah

Diabetes melitus adalah salah satu masalah kesehatan yang sering terjadi di Indonesia khususnya Provinsi Lampung. Penyakit ini merupakan suatu kondisi dimana tubuh tidak cukup untuk menghasilkan atau menggunakan hormon insulin yang membawa glukosa ke dalam sel-sel tubuh. Tujuan penelitian ini adalah membuat model *machine learning* yang dapat mendeteksi dini penyakit diabetes menggunakan *Support Vector Machine* (SVM). Namun, pada dataset yang digunakan dalam penelitian memiliki masalah ketidakseimbangan data *(imbalance data).* Oleh karena itu, digunakan *Random Oversampling* (ROS) untuk mengatasi masalah tersebut. Hasil yang diperoleh dari penelitian ini, ROS mampu menangani *imbalance* data sehingga nilai akurasi yang didapatkan mencapai 96.43% *(excellent classification)* dengan model *type C-Classification* dan kernel *Radial Basis Function* (RBF), serta parameter *sigma* 1 dan *cost* 1 untuk skema data latih 90% dan data uji 10%. Nilai akurasi ini meningkat tajam dibandingkan tanpa ROS yang hanya sekitar 76%.

Kata Kunci : Diabetes Melitus, Ketidakseimbangan Data, Random Oversampling, Support Vector Machine

PENANGANAN *IMBALANCE* DATA DENGAN *RANDOM OVERSAMPLING* (ROS) PADA KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM)

Oleh
DHIFA ZHAFIRAH
1917031032

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar

SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	:	**PENANGANAN *IMBALANCE* DATA DENGAN *RANDOM OVERSAMPLING* (ROS) PADA KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM)**
Nama Mahasiswa	:	Dhifa Zhafirah
Nomor Pokok Mahasiswa	:	**1917031032**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

Bandar Lampung, 8 Februari 2023

MENYETUJUI

1. Komisi Pembimbing

Dr. Aang Nuryaman, S.Si., M.Si.	**Dr. Ahmad Faisol, S.Si, M.Sc**
NIP. 19740316200501100 1	NIP. 19800206200312100 3

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 19740316200501100 1

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Aang Nuryaman, S.Si, M.Si** .....

Sekretaris : **Dr. Ahmad Faisol, S.Si, M.Sc** .....

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

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, M. T.
NIP. 19740705200003100l

Tanggal Lulus Ujian Skripsi: 8 Februari 2023

PERNYATAAN SKRIPSI MAHASISWA

Saya yang bertanda tangan di bawah ini:

Nama	:	**Dhifa Zhafirah**
Nomor Pokok Mahasiswa	:	**1917031032**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Penanganan *Imbalance* Data dengan *Random Oversampling* (ROS) Pada Klasifikasi Penderita Diabetes Menggunakan *Support Vector Machine* (SVM)**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri dan semua tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah karya penulisan ilmiah Universitas Lampung.

Bandar Lampung, 8 Februari 2023

Penulis

Dhifa Zhafirah

NPM. 1917031032

RIWAYAT HIDUP

Penulis bernama Dhifa Zhafirah, dilahirkan di Kota Palembang, Provinsi Sumatera Selatan pada 23 Desember 2001. Penulis merupakan anak tunggal dari pasangan Bapak Andi Randi Jatmiko dan Ibu Merry Fransisca.

Penulis mengawali pendidikan di Taman Kanak-kanak (TK) Ar-Raudah 2006-2007. Kemudian menempuh pendidikan Sekolah Dasar (SD) di SDN 5 Talang pada tahun 2007-2013. Melanjutkan ke Sekolah Menengah Pertama (SMP) di SMPN 1 Bandar Lampung dan lulus pada tahun 2016. Kemudian penulis melanjutkan pendidikan Sekolah Menengah Atas (SMA) di SMA YP Unila Bandar Lampung dan lulus pada tahun 2019.

Pada tahun 2019, penulis terdaftar sebagai mahasiswa S1 Jurusan Matematika FMIPA Unila melalui jalur SNMPTN. Selama menjadi mahasiswa, penulis aktif dalam organisasi Pusat Informasi dan Konseling Remaja (PIK R) RAYA Unila dan Himpunan Mahasiswa Jurusan Matematika (HIMATIKA) FMIPA Unila. Selain itu, penulis mengikuti berbagai perlombaan salah satunya adalah terpilih menjadi Duta Generasi Berencana Provinsi Lampung Tahun 2021. Selama menempuh pendidikan di Jurusan Matematika, penulis berkesempatan untuk mendapatkan Beasiswa Unggulan Kemendikbudristek sejak tahun 2020. Pada tahun 2022, penulis melakukan Kerja Praktik (KP) di Badan Pemeriksa Keuangan Perwakilan Provinsi Lampung dan Kuliah Kerja Nyata (KKN) di Kelurahan Kota Karang Raya, Kecamatan Teluk Betung Utara, Bandar Lampung. Setelah melakukan kegiatan KP dan KKN, penulis mengikuti program Merdeka Belajar Kampus Merdeka (MBKM) dan berkesempatan untuk magang di PT. Nutrifood Indonesia.

KATA INSPIRASI

"Sesungguhnya sesudah kesulitan itu ada kemudahan. Maka apabila kamu telah selesai (dari suatu urusan), kerjakanlah dengan sungguh-sungguh (urusan yang lain)."
(QS Al-Insyirah: 6)

"Intelligence plus character -that is the goal of true education."
(Martin Luther King Jr)

"Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum, sebelum mereka mengubah keadaan diri mereka sendiri."
(QS Ar Rad: 11)

PERSEMBAHAN

Dengan mengucapkan rasa syukur atas segala puji dan kehadirat Allah SWT yang telah melimpahkan nikmat serta hidayah-Nya sehingga skripsi ini dapat diselesaikan. Tak lupa sholawat serta salam selalu tercurahkan kepada junjungan besar Nabi Muhammad SAW yang telah memberikan tuntunan untuk selalu berada di jalan yang benar. Dengan penuh ketulusan, penulis mempersembahkan karya ini untuk :

Bunda dan Ayah

Bunda yang senantiasa memberikan doa, dukungan yang tiada henti dalam setiap keadaan dan keputusan, selalu menerima segala kekurangan, serta memberikan perhatian penuh yang tak ada habisnya.

Dosen Pembimbing dan Pembahas

Dosen pembimbing dan pembahas yang sangat berjasa, selalu membimbing, memberikan arahan, dan juga ilmu yang sangat bermanfaat saat proses pembuatan skripsi ini

Seluruh Keluargaku

Sahabat-sahabatku

Almamater Tercinta, Universitas Lampung.

SANWACANA

Puji syukur kehadirat Allah SWT, atas segala rahmat dan karunia-Nya. Sholawat serta salam selalu tercurahkan kepada junjungan besar Nabi Muhammad SAW, sehingga penulis dapat menyelesaikan skripsi dengan judul **"Penanganan *Imbalance* Data dengan *Random Oversampling* (ROS) Pada Klasifikasi Penderita Diabetes Menggunakan *Support Vector Machine* (SVM)"**. Penulis menyadari bahwa skripsi ini tidak akan terselesaikan dengan baik tanpa adanya arahan, bimbingan, serta kritik dan saran dari berbagai pihak.

Oleh karena itu, dalam kesempatan ini penulis ingin mengucapkan terima kasih kepada :

1. Bapak Dr. Aang Nuryaman, S.Si, M.Si., selaku Dosen Pembimbing I dan Ketua Jurusan Matematika yang selalu memberikan arahan, bimbingan, bantuan, motivasi, dan saran yang mendukung sehingga penulis dapat menyelesaikan skripsi ini.
2. Bapak Dr. Ahmad Faisol, S.Si, M.Sc., selaku Dosen Pembimbing II atas bantuan dan bimbingan kepada penulis selama proses penyusunan skripsi ini berlangsung.
3. Bapak Drs. Nusyirwan, M.Si., selaku Dosen Pembahas yang telah memberikan kritik dan saran yang membangun selama proses penyusunan skripsi.
4. Bapak Prof. Drs. Mustofa, M.A., Ph.D., selaku dosen pembimbing akademik.
5. Bapak Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Seluruh dosen, staff, dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Seluruh keluargaku, Bunda, Oma dan Opa yang senantiasa memberikan motivasi selama kuliah.
8. Sahabat seperjuangan, Kameela Maysoon yang telah menyemangati dan menemani penulis selama proses perkuliahan.
9. Koma *Squad* (Kak Ridho, Fadhlan, dan Mia) yang senantiasa memberikan semangat selama penulis menyelesaikan skripsi.
10. Tim KPS (Reynaldo, Novriansyah, Dimas Ferdiansyah, dan Zeda Erdian) atas dukungan selama proses menyelesaikan skripsi.
11. Tim Pejuang Muda Tanggamus (Kak Faishal, Cherissa, Beby, dan Riska) yang menjadi penyemangat penulis dalam menyelesaikan skripsi.
12. Kak Irvan Yama Pradipta yang senantiasa memberikan dorongan dan motivasi bagi penulis untuk menyelesaikan skripsi.
13. Kak Yazir, Ahmad Yusril Yusro, Rizqatasya A. Z, Azzahra Zulfa, dan Silvi Fitriani, Winda Apriliyanti, dan Sinta Andiana yang senantiasa membantu dan memberikan semangat kepada penulis.
14. Tim Gabut (Amanda, Ikhsan, Dewa, Mega, Siti, dan Elka) yang menjadi penyemangat dalam menjalankan proses perkuliahan.
15. Keluarga Besar UKM U PIK R RAYA dan Bidang Kaderisasi Kepemimpinan HIMATIKA FMIPA Universitas Lampung.sebagai wadah bagi penulis untuk mengembangkan minat dan bakat.
16. Semua pihak yang telah membantu penulis dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna dan masih terdapat banyak kekurangan baik dalam penyajian maupun penulisan. Oleh sebab itu, saran dan kritikan yang membangun senantiasa penulis harapkan demi menyempurnakan skripsi ini.

Bandar Lampung, Februari 2023

Penulis

Dhifa Zhafirah

NPM. 1917031032

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Pada tahun 2018, sebanyak 422 juta jiwa secara global menderita penyakit diabetes melitus, terutama di negara - negara berpenghasilan rendah atau menengah seperti negara di Afrika, India, Bangladesh, Filipina, dan Indonesia serta terdapat 1,6 juta kasus kematian setiap tahunnya (World Health Organization, 2018). Di Indonesia, prevalensi diabetes melitus pada tahun 2018 mencapai angka sebesar 2%, sedangkan prevalensi diabetes melitus di Provinsi Lampung pada tahun 2018 sebesar 1,4% naik dibandingkan tahun 2013 sebesar 0,7% (Riskesdas, 2018).

Selain itu, terdapat 3 kabupaten/kota di Provinsi Lampung yang memiliki presentasi terbesar masalah penyakit diabetes melitus yaitu Metro sebesar 3,3%, Bandar Lampung sebesar 2,3%, dan Pringsewu sebesar 1,8% (Riskesdas Provinsi Lampung, 2018). Sebagian besar pasien diabetes tidak menyadari risiko dari pra-diabetes yang mengarah pada penyakit yang sebenarnya. Oleh karena itu perlunya melakukan prediksi terkait risiko penyakit diabetes terhadap penderitanya. Permasalahan tersebut dapat diatasi dengan proses data *mining*.

Data *mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005). Data *mining* sering juga disebut dengan *Knowledge Discovery in Database* (KDD), yaitu merupakan proses mengekstraksi pola dan informasi yang berguna dari sebuah kumpulan data dan dapat menjadi salah satu alternatif penyelesaian untuk mengatasi permasalahan pengelompokkan suatu data. Adapun beberapa algoritma yang terdapat pada data *mining*, di antaranya

klasterisasi dan klasifikasi. Klasifikasi merupakan sebuah proses untuk menemukan sebuah model yang menjelaskan dan membedakan konsep atau kelas data dengan tujuan memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (Tan dkk., 2006). Klasifikasi dapat diterapkan pada berbagai aspek dan metode ini telah banyak dikembangkan dari waktu ke waktu. Namun, terkadang terdapat permasalahan yang ditemui dalam metode klasifikasi yaitu masalah ketidakseimbangan data (*imbalance* data) (Fitriani dkk., 2021).

Imbalance data merupakan kondisi dimana suatu kelompok kelas memiliki jumlah data yang jauh berbeda dibandingkan dengan kelas lainnya. Kelas yang memiliki jumlah data lebih banyak disebut dengan *majority class* dan kelas yang mempunyai jumlah data lebih sedikit disebut dengan *minority class* (Barro dkk., 2013). Hal ini disebabkan karena pembelajaran pengklasifikasian akan cenderung memprediksi kelas data yang banyak dibandingkan dengan kelas data yang sedikit.

Berbagai penelitian sebelumnya yang menggunakan metode *Support Vector Machine* (SVM) pada kasus *imbalance* data telah banyak dilakukan, diantaranya penelitian oleh Amelia, dkk. (2018), memprediksi keberhasilan studi mahasiswa menggunakan pemodelan SVM dengan mempertimbangkan karakteristik dan latar belakang pendidikan mahasiswa. Penanganan data tidak seimbang dilakukan menggunakan *Sythetic Minority Oversampling Technique* (SMOTE) yang berhasil meningkatkan kinerja SVM dalam mengklasifikasikan mahasiswa yang tidak lulus dengan akurasi sebesar 89.08%. Penelitian lainnya dilakukan Mutmainah (2021), dengan membandingkan *Random Oversampling* (ROS) dan *Random Undersampling* (RUS) pada kemungkinan penyakit stroke. Penanganan *imbalance* data dilakukan antar *class* 1 (stroke) dan *class* 0 (tidak stroke) dengan distribusi antar data sama. Hasil yang didapatkan pada penggunaan teknik ROS mendapat performa yang lebih tinggi yaitu 95% daripada teknik RUS yang mendapat performa 76%. Selain itu, terdapat penelitian yang telah dilakukan Mucholladin, dkk. (2021), melakukan klasifikasi penyakit diabetes menggunakan metode SVM

dengan jumlah sampel sebanyak 768 pasien diabetes yang diambil dari *Pima Indians Dataset*. Hasil penelitian ini menunjukkan bahwa akurasi klasifikasi mencapai 99,4% untuk memprediksi diabetes melitus.

Di sisi lain, untuk kasus klasifikasi penderita penyakit diabetes di Bandar Lampung, sejauh penelitian yang dilakukan belum ada yang meneliti hal ini. Data yang digunakan dalam penelitian ini mengalami *imbalance* data. Oleh karena itu, penulis menerapkan ROS dalam menangani *imbalance data* dengan menggunakan metode SVM pada klasifikasi penderita penyakit diabetes di Bandar Lampung.

1.2 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini antara lain :

1. Melakukan *balancing* data dengan menggunakan ROS.
2. Melakukan klasifikasi terhadap penderita diabetes pada Komunitas Kesehatan di Bandar Lampung dengan metode SVM.
3. Mengetahui kinerja metode klasifikasi penderita diabetes dengan menggunakan metode SVM.

1.3 Manfaat Penelitian

Manfaat yang ingin diperoleh dari penelitian ini adalah sebagai berikut :

1. Menambah pengetahuan dan wawasan dalam mengatasi permasalahan *imbalance* data.
2. Mengetahui klasifikasi risiko penyakit diabetes pada komunitas kesehatan di Bandar Lampung menggunakan metode SVM.

II. TINJAUAN PUSTAKA

2.1 Diabetes Melitus

Diabetes melitus umumnya disebut sebagai "diabetes" yaitu penyakit kronis yang berhubungan dengan tingginya kadar glukosa (gula) dalam darah (Devi dkk., 2019). Penyakit ini adalah salah satu masalah kesehatan yang sering terjadi di seluruh dunia. Diabetes merupakan suatu kondisi dimana tubuh tidak cukup untuk menghasilkan atau menggunakan hormon insulin yang membawa glukosa ke dalam sel-sel tubuh dan memungkinkan glukosa untuk masuk dan menjadi bahan bakar mereka.

Obesitas merupakan salah satu faktor untuk meningkatkan gula darah yang merupakan sebuah indikator dari diabetes melitus. Kadar gula darah dipengaruhi pula oleh faktor herediter, aktivitas fisik, asupan diet, keluaran energi, metabolisme, dan hormonal. Peningkatan glukosa dan lemak akan mengakibatkan transportasi asam lemak yang kedalam *adipose* dan *lipogenesis* meningkat. Program olah raga yang baik dan terukur akan menstabilkan kadar gula darah (Purwandari, 2014).

2.2 *Body Mass Index* (BMI)

Body Mass Index (BMI) atau Indeks Massa Tubuh (IMT) adalah parameter yang digunakan untuk mengetahui status berat badan seseorang apakah tergolong normal maupun tidak (*underweight* atau *overweight*), data yang diperlukan untuk mencari BMI adalah data selisih antara berat badan dan tinggi badan. BMI juga dapat

digunakan untuk menggambarkan komposisi tubuh secara kasar, meskipun tidak disertai dengan nilai dari konstribusi berat dari lemak dan otot (Supariasa, 2012). Menurut Sugondo (2009), hasil dari penghitungan Indeks Massa Tubuh (IMT) dapat diklasifikasikan bedasarkan klasifikasi menurut klasifikasi Kriteria Asia Pasifik menjadi *underweight*, normal, *overweight*, dan obesitas dengan rentang angka sebagai berikut:

Tabel 1. Klasifikasi IMT menurut Kriteria Asia Pasifik

Klasifikasi	Indeks Massa Tubuh
Underweight (badan berat kurang)	<18,5
Normal	18,5-22,9
Overweight (berat badan lebih)	≥23
Berisiko	23-24,9
Obesitas I	25-29,9
Obesitas II	≥30

2.3 Data *Mining*

Data *mining* adalah sebuah metode dalam menemukan informasi berharga dari sejumlah data yang dilakukan dengan memanfaatkan ilmu lain seperti statistik, matematika, pengenalan pola (Larose dan Larose, 2014). Data *mining* merupakan sebuah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengektraksi dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari dataset yang besar. Data *minin*g sering juga disebut sebagai *Knowledge Discovery in Database* (KDD).

Berikut pengelompokan data *mining* (Buulolo, 2020) :

1. Deskripsi (*Description*);
2. Estimasi (*Estimation*);
3. Prediksi (*Prediction*)

4. Klasifikasi (*Classification*);
5. Klasterisasi (*Clustering*); dan
6. Asosiasi (*Association*).

2.3.1 *Knowledge Discovery in Database* (KDD)

Knowledge Discovery in Database (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007).

Proses KDD adalah urutan iteratif dan interaktif dari tahapan-tahapan utama berikut (Gullo, 2015) :

1. Data *Integration*, yang tujuan utamanya adalah untuk membuat kumpulan data target dari data asli yaitu, memilih *subset variable* atau *sample* data, dimana penemuan harus dilakukan;
2. Data *Selection*, merupakan proses seleksi data, data yang relevan digunakan terhadap analisis yang akan dilakukan.
3. Data *Preprocessing*, yang bertujuan untuk membersihkan data dengan melakukan berbagai operasi, seperti pemodelan dan pengulangan, mendefinisikan strategi yang tepat untuk menangani bidang data yang hilang, akuntansi untuk informasi urutan waktu;
   a. Data *cleansing*

      Umumnya data yang didapatkan memiliki data hilang, ataupun kesalahan pada *input* data. Data *cleansing* merupakan proses menghilangkan *noise* dan data yang tidak relevan.
   b. *Scaling data*

      Scaling data merupakan proses transformasi data dari bentuk asli ke dalam bentuk lain yang sesuai untuk data mining. *Scaling data* digunakan untuk menyesuaikan data yang diolah berdasarkan algoritma yang digunakan.

Terdapat dua cara yang biasanya digunakan untuk *scaling data* yaitu:

a. *Min-max normalization*, merupakan proses transformasi data yang bekerja dengan cara menempatkan data dalam range 0 sebagai nilai terkecil dan 1 sebagai nilai terbesar. Rumus perhitungan pada *min-max normalization* yaitu:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

dengan :

x' = nilai x baru

x_{min} = nilai x minimum

x_{max} = nilai x maksimum

b. *Z-score normalization* (*standard scaler*) merupakan metode transformasi data bedasarkan nilai rata-rata dan standar deviasi yang bertujuan untuk mencegah adanya data yang memiliki nilai terlalu besar dibandingkan dengan nilai yang lain. Rumus perhitungan pada *Z-score normalization* yaitu:

$$Z = \frac{x - \mu}{\sigma} \quad (2.2)$$

dengan :

x = nilai yang diamati

μ = rata-rata nilai (*mean*)

σ = standar deviasi

c. *Handling data categorical*

Dataset dapat terdiri dari data numerik maupun data kategorik, pada data kategorik diberikan label agar lebih mudah dipahami oleh komputer. Terdapat dua cara yang biasa digunakan untuk *handling data categorical* yaitu l*abel encoding* dan *one hot encoding*. *Label encoding* digunakan ketika data memiliki tingkatan yang berbeda (Winata dkk ,2020). Berikut merupakan contoh data yang sudah dilakukan *label encoding* :

Tabel 2. Contoh *label encoding*

Penilaian	*Label encoding*
Buruk	1
Cukup	2
Baik	3

One hot encoding digunakan ketika data yang dimiliki tidak terdapat tingkatan yang berbeda. Berikut merupakan contoh data yang sudah dilakukan *one hot encoding* :

Tabel 3. Contoh *one hot encoding*

Diabetes	Non diabetes
0	1
0	1
1	0

4. Data *Mining*, yang berkaitan dengan pengambilan pola yang menarik dengan memilih metode data mining tertentu atau tugas (misalnya, *summarization, classification, clustering, regression*, dan sebagainya), algoritma yang tepat untuk melakukan tugas di tangan, dan representatif yang tepat dari hasil *output*;
5. Data *Interpretation/Evaluation*, yang di eksploitasi oleh pengguna untuk menafsirkan dan mengekstrak pengetahuan dari pola yang ditambang, dengan menvisualisasikan pola. Interpretasi ini biasanya dilakukan dengan memvisualisasikan pola, model, atau data yang diberikan model tersebut dan, dalam kasus, secara iteratif melihat kembali langkah-langkah sebelumnya dari proses.

2.4 *Machine Learning*

Machine Learning merupakan sebuah studi tentang algoritma untuk mempelajari sesuatu dalam melakukan beberapa hal tertentu yang dilakukan oleh manusia secara

otomatis. Belajar dalam hal ini berkaitan dengan bagaimana menuntaskan berbagai tugas yang ada, atau membuat suatu prediksi kesimpulan baru yang akurat dari berbagai pola yang sudah dipelajari sebelumnya (Shwartz dan David, 2014). Algoritma dalam *machine learning* bekerja dengan cara membangun sebuah model dari masukan agar dapat menghasilkan suatu prediksi atau pengambilan keputusan berdasarkan data yang ada. Penelitian terkini mengungkapkan bahwa *machine learning* terbagi menjadi tiga kategori yaitu, *Supervised Learning, Unsupervised Learning,* dan *Reinforcement Learning* (Somvanshi dan Chavan, 2016).

1. *Supervised Learning*

   Supervised Learning adalah metode klasifikasi di mana kumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang tidak dikenal. Metode *supervised learning* didasarkan pada kumpulan sampel data yang memiliki label. Kumpulan sampel digunakan untuk meringkas karakteristik distribusi ukuran perilaku dalam setiap jenis aplikasi sehingga membentuk model perilaku dari data (Amei dkk., 2011).
2. *Unsupervised Learning*

   Unsupervised Learning sering disebut *cluster* dikarenakan tidak ada kebutuhan untuk pemberian label dalam kumpulan data dan hasilnya tidak mengidentifikasi contoh di kelas yang telah ditentukan (Thupae dkk., 2018).
3. *Reinforcement Learning*

   Reinforcement Learning biasanya berada antara *supervised learning* dan *unsupervised learning*, teknik ini bekerja dalam lingkungan yang dinamis di mana konsepnya harus menyelesaikan tujuan tanpa adanya pemberitahuan dari komputer secara eksplisit jika tujuan tersebut telah tercapai (Das dan Nene, 2017).

2.5 Klasifikasi

Klasifikasi merupakan proses menemukan model atau fungsi yang membedakan kelas atau konsep data (Utami dkk, 2020). Klasifikasi termasuk ke dalam metode

supervised learning, yaitu metode yang berfungsi untuk menemukan hubungan antara atribut masukan dan atribut target. Tujuan dari metode klasifikasi ini untuk meningkatkan kehandalan hasil yang diperoleh dari suatu data. Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning*, dimana algoritma klasifikasi dibuat untuk menganalisis data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data *testing* digunakan untuk memperkirakan akurasi dari *rule* klasifikasi (Han dan Kamber, 2006). Menurut Gorunescu (2011), proses klasifikasi memiliki empat komponen dasar, yaitu :

1. Kelas

   Variabel dependen yang berupa kategorikal yang merepresentasikan “label” yang terdapat pada suatu objek klasifikasi.
2. Prediktor

   Prediktor merupakan variabel independen yang direpresentasikan oleh karakteristik (atribut) data.
3. *Training dataset*

   Training dataset merupakan set data yang berisi nilai dari kedua komponen sebelumnya (kelas dan prediktor) yang digunakan untuk menentukan kelas yang cocok berdasarkan prediktor.
4. *Testing dataset*

   Testing dataset berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat serta untuk mengukur tingkat akurasi dari klasifikasi.

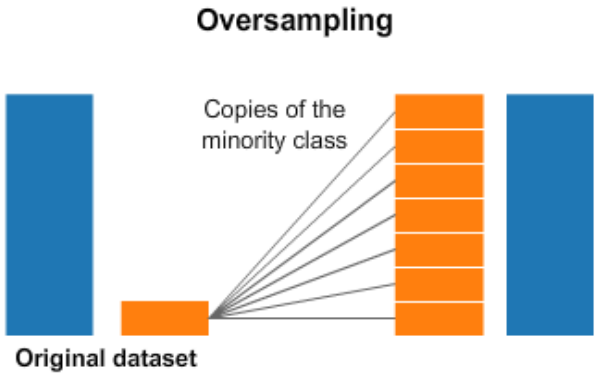
2.6 *Imbalance Data*

Imbalance data merupakan kondisi dimana suatu kelompok kelas memiliki jumlah data yang jauh berbeda dibandingkan dengan kelas lainnya. Kelas yang memiliki jumlah data lebih banyak disebut dengan *majority class* dan kelas yang mempunyai jumlah data lebih sedikit disebut dengan *minority class* (Barro dkk., 2013). Untuk mengatasi *imbalance* data, terdapat beberapa metode yang dapat digunakan. Salah

satunya dengan menggunakan *resampling.* Pendekatan *resampling* terdiri menjadi 3 yaitu, *Random Oversampling* (ROS), *Random Undersampling* (RUS), dan *Hibrida* yang menggabungkan kedua pendekatan *sampling* (Jian dkk., 2016).

2.7 *Random Oversampling*

Random Oversampling merupakan penambahan data dari kelas minoritas ke dalam data *training* secara acak. Proses penambahan ini diulang sampai jumlah data kelas minoritas sama dengan jumlah kelas mayoritas. *Random Oversampling* bertujuan untuk meningkatkan ukuran kelas minoritas dengan mensintesis sampel baru atau dataset *training* dengan menduplikasi secara acak sampel kelas minoritas (Yu dkk., 2017).

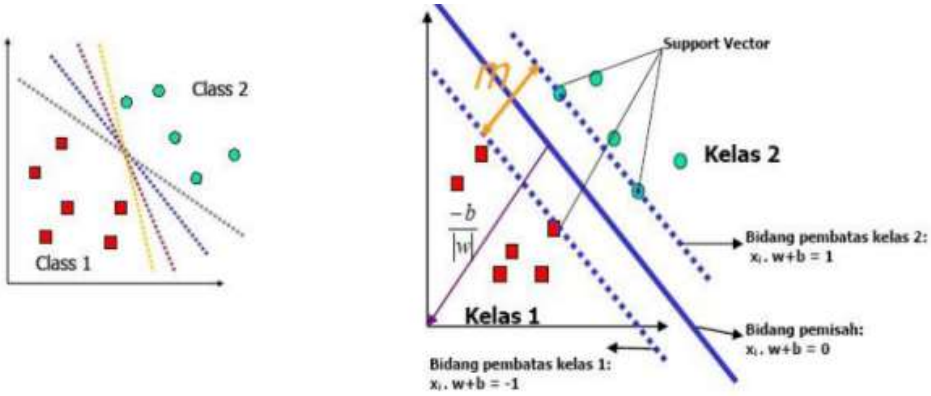


Gambar 1. Proses *Random Oversampling*

2.8 *Support Vector Machine*

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada *input space* (Foeady dkk., 2018). Batas pemisah antar kelas (*hyperplane*) dapat memaksimalkan jarak atau *margin* antara kelas data. *Hyperplane* terbaik antara kedua kelas dapat

ditemukan dengan mengukur *margin* dan kemudian mencari titik maksimalnya. Usaha dalam mencari *hyperplane* yang terbaik sebagai pemisah kelas-kelas adalah inti dari proses pada metode SVM (Assaffat, 2015).



Gambar 2. *Support Vector Machine* menemukan *hyperplane* terbaik

Pada Gambar 2 menunjukkan bahwa terdapat dua pola yang merupakan anggota dari dua buah kelas, yaitu +1 dan -1. Pola yang tergabung dalam kelas -1 disimbolkan dengan kotak berwarna merah, sedangkan kelas +1 disimbolkan dengan lingkaran berwarna hijau. Pada Gambar 2 terlihat berbagai alternatif bidang pemisah yang dapat memisahkan semua data set sesuai dengan kelasnya.

Gambar sebelah kiri merupakan alternatif bidang pemisah sesuai kelasnya, sedangkan gambar sebelah kanan merupakan bidang pemisah terbaik optimal (*hyperplane*) dengan jarak terbesar. Adapun data yang terletak pada bidang pembatas disebut dengan *support vector* (Adinegoro dkk., 2015). Tujuan utama dari klasifikasi adalah mencari *hyperplane* permisah antara kedua kelas.

Misalkan data yang tersedia direpresentasikan dalam bentuk vektor :

$$\vec{d} \coloneqq \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\},$$

dengan $x_i \in R \; dan \; y_i \in \{-1,1\}$

Diasumsikan data tersebut terpisah secara sempurna ke dalam dua kelas yaitu -1 dan 1 oleh *hyperplane,* yang didefinisikan (Zaki, 2014) :

$$w^T . \vec{x} + b = 0 \qquad (2.3)$$

dengan :

w = matriks vektor normal pada *hyperplane*

b = jarak dari *hyperplane* ke titik pusat

sehingga menurut Cortes dan Vapnik (1995), diperoleh persamaan :

$$[(w^T.\overrightarrow{x_i}) + b] \geq +1 \; untuk \; y_i = +1 \tag{2.4}$$

$$[(w^T.\overrightarrow{x_i}) + b] \leq -1 \; untuk \; y_i = -1 \tag{2.5}$$

dengan :

x_i = himpunan data *training, i* = 1,2,…,n

y_i = label kelas dari x_i

Persamaan (2.4) dan (2.5) dapat disederhanakan menjadi :

$$y_i(w^T.\overrightarrow{x_i} + b) \geq 1, i = 1,2,3..,N \tag{2.6}$$

Pemaksimalan jarak terdekat antara *hyperplane* dengan *pattern* dilakukan untuk menghitung *margin* maksimum antar kelas. *Margin* didefinisikan sebagai $d = d_1 + d_2$, sehingga *margin* akan memiliki nilai maksimum jika $d_1 = d_2$. *Margin* maksimum dapat didapatkan dengan memaksimalkan jarak antara *hyperplane* dengan titik terdekatnya yaitu $\frac{1}{\|\vec{w}\|}$.

$$d = d_1 + d_2 = \frac{1}{\|\vec{w}\|}(|w^T.\overrightarrow{x_1} + b||w^T.\overrightarrow{x_2} + b|) = \frac{2}{\|\vec{w}\|} \tag{2.7}$$

Berdasarkan persamaan di atas, maka untuk mencari *margin* maksimal sama dengan meminimumkan nilai $\|w\|^2$, secara matematis dinyatakan sebagai berikut :

$$min\frac{1}{2}\|\vec{w}\|^2 \tag{2.8}$$

Optimasi dapat dilakukan dengan menggunakan *Lagrange Multiplier* sebagai berikut :

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i[y_i(w^T.\overrightarrow{x_i} + b) - 1]$$

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i y_i(w^T.\overrightarrow{x_i} + b) - \sum_{i=1}^{l} a_i \tag{2.9}$$

a_i merupakan *lagrange multiplier* dengan nilai nol atau positif ($a_i \geq 0$). Optimasi dilakukan dengan meminimalkan L terhadap w dan b sebagai berikut (Hamel, 2009):

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

$$\sum_{i=1}^{l} a_i y_i = 0 \tag{2.10}$$

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

$$\vec{w} - \sum_{i=1}^{l} a_i y_i \overrightarrow{x_i} = 0$$

$$\vec{w} = \sum_{i=1}^{l} a_i y_i \overrightarrow{x_i} \tag{2.11}$$

Selain itu, optimasi dapat dilakukan dengan memaksimalkan *L* terhadap a_i dengan subtitusi Persamaan (2.10) dan (2.11) ke dalam Persamaan (2.9) sebagai berikut :

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i y_i (w^T \overrightarrow{x_i} + b) - \sum_{i=1}^{l} a_i$$

$$L = \frac{1}{2}(w^T . \vec{w}) - (\sum_{i=1}^{l} a_i y_i w^T \overrightarrow{x_i} + \sum_{i=1}^{l} a_i y_i b - \sum_{i=1}^{l} a_i)$$

$$L = \frac{1}{2}(\sum_{i=1}^{l} a_i y_i \vec{x}_i . \sum_{j=1}^{l} a_j y_j \overrightarrow{x_j}) - (\sum_{i=1}^{l} a_i y_i \overrightarrow{x_i} . \sum_{j=1}^{l} a_j y_j \overrightarrow{x_j}) + 0 - \sum_{i=1}^{l} a_i)$$

$$L = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \overrightarrow{x_i}\overrightarrow{x_j} - (\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \overrightarrow{x_i}\overrightarrow{x_j} - \sum_{i=1}^{l} a_i)$$

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \overrightarrow{x_i}\overrightarrow{x_j} \tag{2.12}$$

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

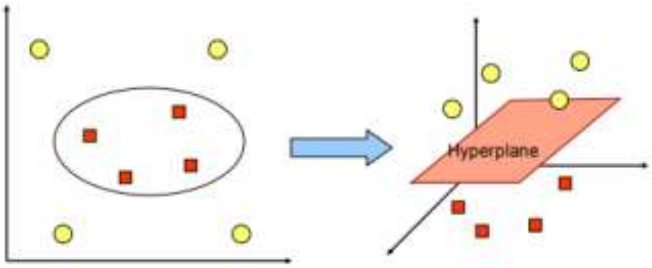
Nilai a_i akan diperoleh dengan penyelesaian Persamaan (2.12) yang digunakan untuk mencari *primal variable* dengan rumus :

$$\overrightarrow{w_i} = \sum_{i=1}^{l} a_i y_i \, K(\overrightarrow{x_i}, \overrightarrow{x_j}) \;, b = -\frac{1}{2}(w^T x^+ + w^T x^-) \tag{2.13}$$

Setelah proses telah dilakukan, maka diperoleh $a_i > 0$ yang disebut dengan *support vector* dan sisanya memiliki nilai $a_i = 0$. Fungsi keputusan yang dihasilkan hanya dipengaruh oleh nilai *support vector*. Pada umumnya terdapat permasalahan-permasalahan di dunia nyata sangat jarang ditemukan data yang terpisah secara linear, sehingga dalam mengatasi permasalahan nonlinear SVM dapat menggunakan fungsi *kernel*. Konsep kerja *kernel* adalah dengan mentransformasi

data ke dalam dimensi ruang fitur (*feature space*). Penyelesaian kasus non-linier dapat diatasi dengan SVM yang telah dikembangkan yaitu dengan menggunakan *kernel trick* yang dapat mengubah data menjadi linier (Hamel, 2009). Adapun *kernel trick* dirumuskan dengan :

$$K(\overrightarrow{x_i}, \overrightarrow{x_j},) = \phi(\overrightarrow{x_i},).\phi(\overrightarrow{x_j},) \tag{2.14}$$



Gambar 3. Transformasi Data dalam *Feature Space*

Pada Gambar 3 menunjukkan data berdimensi dua tidak dapat dipisahkan secara linear oleh *hyperplane*. Pada Gambar 3 mengilustrasikan pemetaan data ke dalam ruang dengan dimensi lebih tinggi (dimensi tiga) sehingga dua kelas dapat dipisahkan secara linear oleh *hyperplane*. Berikut notasi matematika dari *mapping* tersebut :

$$\phi: R^d \rightarrow R^q, \mathrm{d} < \mathrm{q} \tag{2.15}$$

Umumnya, transformasi ϕ tidak diketahui sehingga diganti dengan fungsi *kernel* $K(x_i, x_j)$. Hasil klasifikasi dapat diperoleh dari persamaan :

$$f(\phi(\overrightarrow{x_i})) = sign(w^T.\phi(\overrightarrow{x_i}) + b)$$

$$f(\phi(\overrightarrow{x_i})) = sign(\textstyle\sum_{i=1}^{n} a_i y_i\, \phi(\overrightarrow{x_i}).\phi(\overrightarrow{x_j}) + b)$$

$$f(\phi(\overrightarrow{x_i})) = sign(\textstyle\sum_{i=1}^{n} a_i y_i\, K(\overrightarrow{x_i}, \overrightarrow{x_j}) + b) \tag{2.16}$$

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 fungsi *kernel* yang umumnya digunakan dalam SVM sebagai berikut (Han dkk., 2012) :

a. Kernel Linear

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \overrightarrow{x_i}\,\overrightarrow{x_j} \tag{2.17}$$

b. Kernel Polynomial

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i}.\overrightarrow{x_j} + 1)^d \tag{2.18}$$

c. Kernel Sigmoid

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \tanh(\gamma \overrightarrow{x_i}.\overrightarrow{x_j} - \delta) \tag{2.19}$$

d. Kernel *Radial Basis Function* (RBF)

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \mathrm{e}^{-\left(\frac{\|\overrightarrow{x_i} - \overrightarrow{x_j}\|^2}{2\sigma^2}\right)}, \sigma > 0 \tag{2.20}$$

Berdasarkan penelitian sebelumnya, di antara jenis-jenis kernel yang ada dinyatakan bahwa kernel RBF mampu meningkatkan nilai akurasi yang lebih baik dan akurat.

2.9 Evaluasi Model

Pengujian atau evaluasi model memiliki tujuan agar dapat mengetahui seberapa baik kinerja dari model pada tahapan pembelajaran menggunakan kernel SVM. Ada banyak indikator penilaian dalam bidang klasifikasi salah satunya adalah *confusion matrix*. *Confusion matrix* menggambarkan performa model melalui tabel (Saputro dan Sari, 2019). Setiap baris dari matriks tersebut mempresentasikan klasifikasi aktual dari data dan setiap kolom dari matriks tersebut mempresentasikan klasifikasi prediksi dari data atau sebaliknya. Berikut adalah tabel bentuk *confusion matrix* secara umum untuk jumlah label/kelas sebanyak 2 (*binary classification*) :

Tabel 4. *Confusion Matrix* untuk Jumlah k=2

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

1. *True Postive* (TP) adalah data diprediksi positif dan data sebenarnya adalah positif.
2. *True Negative* (TN) adalah data diprediksi negatif dan data sebenarnya adalah negatif.
3. *False Positive* (FP) adalah data diprediksi positif dan data sebenarnya adalah negatif.
4. *False Negative* (FN) adalah data diprediksi negatif dan data sebenarnya adalah positif.

Berdasarkan hasil dari *confusion matrix*, dapat pula dilakukan perhitungan untuk mengukur performa model, yaitu *accuracy, precision*, *recall* (*sensitivity*/*true positive rate*), dan *f1- score* (Saputro dan Sari, 2019). Masing-masing perhitungannya didefinisikan sebagai berikut:

a. *Accuracy*, efektivitas keseluruhan dari hasil klasifikasi yang telah dilakukan.

$$Accuracy = \frac{TP + TN}{Total} \tag{2.21}$$

b. *Precision*, yaitu melihat seberapa sering model memprediksi positif dan secara aktual prediksi itu benar dengan perumusan sebagai berikut:

$$Precision = \frac{TP}{FP + TP} \tag{2.22}$$

c. *Recall*, yaitu seberapa sering model memprediksi positif pada data yang memiliki klasifikasi aktual yang positif dengan perumusan sebagai berikut:

$$Recall = \frac{TP}{FN + TP} \tag{2.23}$$

d. *F1-score*, yaitu merupakan hubungan antara data berlabel positif dari hasil klasifikasi yang menunjukan keseimbangan antara *precision* dan *recall*

$$F_1 = 2 \text{ x } \frac{Precision \times Recall}{Precision + Recall} \tag{2.24}$$

Selain dengan menggunakan tabel *confusion matrix* dan perhitungan *accuracy, precision, recall,* serta *f1-score*. Evaluasi model SVM dapat dilakukan dengan melihat kurva *Receiver Operating Characteristic* (ROC). Kurva ROC menunjukkan hubungan antara *observed class* dan *predicted class*. Akurasi klasifikasi ROC dilakukan dengan cara menghitung luas daerah di bawah kurva ROC. Kriteria keakuratan tes diagnostik menggunakan AUC disajikan pada Tabel 5 (Gorunescu, 2011) :

Tabel 5. Kriteria Nilai AUC

Nilai AUC	Interpretasi
0,90-1,00	*excellent classification*
0,80-0,90	*good classification*
0,70-0,80	*fair classification*
0,60-0,70	*poor classification*
0,50-0,60	*failure*

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester ganjil tahun ajaran 2022/2023 bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data primer, yaitu data penderita penyakit diabetes yang diperoleh dari hasil pengecekan kesehatan *Body Mass Index* (BMI) pada Komunitas Kesehatan di Bandar Lampung tepatnya pada bulan Oktober tahun 2022. Jumlah data yang digunakan pada penelitian ini yaitu sebanyak 154 data dan terdapat 10 variabel yaitu, *name*, *age*, *weight*, *height*, *body fat*, *visceral fat*, *resting metabolism*, *body mass index*, *body age,* dan *outcome*.

3.3 Metode Penelitian

Langkah-langkah yang dilakukan pada penelitian ini sebagai berikut :

1. Melakukan visualisasi data untuk menggambarkan dan mendeskripsikan data yang digunakan. Pada tahap ini ditampilkan *bar chart* untuk menunjukkan perbandingan jumlah antara penderita diabetes dan non diabetes. Selain itu dilakukannya analisis eksplorasi data dengan statistik deskriptif sederhana.
2. Melakukan *preprocessing* data, yaitu :

a. *Cleansing Data,* memastikan data tidak memiliki data hilang (*missing value*) dan data duplikasi.
b. *Scaling Data*, melakukan transformasi data menggunakan *standard scaler*.
c. *Handling Data Categorical*¸memberikan label pada data yang berbentuk kategorik dengan menggunakan *one hot encoding*.

3. *Handling Imbalance Data*

   Mengatasi *imbalance* data dengan menggunakan *Random Oversampling* (ROS). *Random Oversampling* bertujuan untuk meningkatkan ukuran kelas minoritas dengan mensintesis sampel baru atau dataset *training* dengan menduplikasi secara acak sampel kelas minoritas.

4. Melakukan *splitting data* menjadi 2 (*training* dan *testing*) dengan 4 skema yaitu, 60% data *training* dan 40% data *testing*, skema 70% data *training* dan 30% data *testing*, skema 80% data *training* dan 20% data *testing*, skema 90% data *training* dan 10% data *testing* yang diambil secara acak dari dataset penelitian.
5. Membangun model *Support Vector Machine*, dengan menggunakan bantuan *hyperparameter tuning* yang berfungsi untuk mendapatkan parameter terbaik dalam melakukan prediksi.
6. Melakukan evaluasi terhadap model

   Pada tahap ini model yang sudah dibangun selanjutnya diuji untuk mengetahui seberapa baik performa model yang dihasilkan dengan *confusion matrix* dan kurva *Area Under Curve* (AUC)-*Receiver Operating Characteristic* (ROC).

V. KESIMPULAN

Setelah melakukan proses *machine learning* dengan menggunakan metode *Support Vector Machine* (SVM) pada klasifikasi penderita diabetes, dapat diambil kesimpulan sebagai berikut :

1. Aspek yang digunakan pada penelitian ini diantaranya, ROS untuk menangani *imbalance data,* pembagian skema data 90% data latih dan 10% data *testing* menjadi skema terbaik untuk pembagian data, model dengan parameter terbaik yaitu *sigma* sebesar 1 dan *cost* sebesar 1.
2. Data penderita diabetes pada komunitas kesehatan di Bandar Lampung telah dilakukan *balancing data* menggunakan ROS sehingga jumlah data menjadi seimbang (*balance*) yaitu 138 data untuk kategori "Diabetes" dan 138 data untuk kategori "Non diabetes".
3. Hasil uji SVM menggunakan data yang telah dilakukan *balancing* dengan ROS didapat model terbaik yang layak digunakan dalam melakukan klasifikasi penderita diabetes pada Komunitas Kesehatan di Bandar Lampung. Model dengan kernel RBF yaitu menggunakan parameter *sigma* sebesar 1 dan *cost* sebesar 1 pada skema data latih 90% dan uji 10% didapat nilai akurasi sebesar 96.43%, *precision* sebesar 93.33%, *recall* sebesar 100%, *f1-score* sebesar 96.55%, serta memiliki nilai AUC terbesar yaitu 0.964. Maka, dapat disimpulkan metode SVM memberikan kinerja klasifikasi yang sangat baik (*excellent classification*) pada data penderita diabetes Komunitas Kesehatan di Bandar Lampung Tahun 2022.

DAFTAR PUSTAKA

Abdillah, Abdul A. dan Suwarno. 2016. Diagnosis of Diabetes Using Support Vector Machines with Radial Basis Function Kernels. *International Journal of Technology*. **5**: 849-858.

Adinegoro, A., Atmaja, R. D., dan Purnamasari, R. 2015. Deteksi tumor otak dengan ektrasi ciri dan feature selection mengunakan linear. *Proceeding of Engineering*. **2**(2): 2532.

Amei, W., Huailin, D., Qingfeng, W., dan Ling, L. 2011. A survey of application-level protocol identification based on machine learning. *2011 International Conference on Information Management, Innovation Management and Industrial Engineering.* **3**: 201–204.

Amelia, O. D., Soleh, A. M., dan Rahardiantoro, S. 2018. Pemodelan Support Vector Machine Data Tidak Seimbang Keberhasilan Studi Mahasiswa Magister IPB. *Institute of Electrical and Electronics Engineers*. **2**(1): 33-40.

Assaffat. 2015. Analisis Akurasi Support Vector Machine Dengan Fungsi Kernel Gaussian RBF Untuk Prakiraan Beban Listrik Harian Sektor Industri. *Jurnal Momentum*. **11**(2).

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).

Buulolo, E. 2020. *Data Mining Untuk Perguruan Tinggi*. Deepublish Publisher.

Cortes., dan Vapnik. 1995. Support Vector Network, Machine Learning. *Kluwer Academic Publisher.* 273-297.

Das, S., dan Nene, M. J. *2017.* A survey on types of machine learning techniques in intrusion prevention systems. *International Conference on Wireless Communications*.

Devi, R. D., Bai, A., dan Nagarajan, N. 2019. A Novel Hybrid Approach for Diagnosing Diabetes Mellitus using Farthest First and Support Vector Machine Algorithms. *Obesity Medicine*.

Fitriani, Reza Dwi., Yasin, Hasbi., dan Tarno. 2021. Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes. *Jurnal Gaussian*. **10**(1) :11-20.

Foeady, A. Z., Novitasari, D. C., dan Asyhar, A. H. 2018. Automated Diagnosis System of Diabetic Retinopathy Using GLCM Method and SVM Classifier. *Proceeding of EECSI*. 154-160.

Gorunescu, F. 2011. *Data Mining : Concept, Model and Techniques*. Springer. Berlin.

Gullo, Francesco. 2015. From patterns in data to knowledge discovery: what data mining can do. *Physics Procedia*. **62**: 18–22.

Hamel, L. 2009. Knowledge Discovery with Support Vector Machines. In *Knowledge Discovery with Support Vector Machines.*

Han, J.,dan Kamber, M. 2006. *Data Mining Concept and Tehniques*. Morgan Kauffman. San Fransisco.

Han, J., Kamber, M., dan Pei, J. 2012. Third Edition : Data Mining Concepts and Techniques. *Journal of Chemical Information and Modeling*. **53**(9): 1689-1699.

Jian, C., Gao, J., dan Ao, Y. 2016. A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Journal of Neurocomputing*. **193**: 115–122.

Larose, D. T. 2005. *Discovering Knowledge in Data: An Introduction to Data mining*. John Willey and Sons, Inc. New York.

Larose, D. T., dan Larose, C. D. 2014. *Discovering Knowledge in Data*. New York.

Mucholladin, A. W., Bachtiar, F. A., dan Furqon, M. T. 2021. Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. **5(**2): 622-633.

Mutmainah, S. 2021. Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke. *Jurnal SNATi*. **1**: 10–16.

Purwandari, H. 2014. Hubungan Obesitas Dengan Kadar Gula Darah Pada Karyawan Di RS Tingkat IV Madiun. *Jurnal Kesehatan STIKES*. **1**:65–72.

Riskesdas. 2018. *Profil Riset Kesehatan Dasar Tahun 2018*. Jakarta. Riskesdas RI

Riskesdas Provinsi Lampung. 2018. *Profil Riskesdas Provinsi Lampung Tahun 2018*. Lampung. Provinsi Lampung.

Santosa, B. 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu, Yogyakarta.

Saputro, I. W. dan Sari, B. W. 2019. Uji Performa Algoritma Naive Bayes untuk Prediksi Masa Studi Mahasiswa. *Citec Journal*. **6(**1): 1-11.

Shwartz, S dan David, S. B. 2014. *Understanding Machine Learning From Theory to Algorithm*. Cambridge University Press. New York.

Somvanshi, M dan Chavan, P. 2016. A review of machine learning techniques using decision tree and support vector machine. 2016 *International Conference on Computing Communication Control and Automation.*

Sugondo. 2009. *Buku Ajar Penyakit Dalam*. EGC. Jakarta.

Supariasa, N. 2012. *Penilaian Status Gizi.* EGC. Jakarta.

Tan, P., Steinbach, M., dan Kumar, V. 2006. *Introduction to Data Mining*. KMedia, Yogyakarta.

Thupae, R., Isong, B., Gasela, N., dan Abu, M. 2018. Machine Learning Techniques for Traffic Identification and Classification in SDWSN. *Institute of Electrical and Electronics Engineers*. 4645–4650.

Utami, Y. T., Shofiana, D. A., dan Heningtyas, Y. 2020. Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi. *Jurnal Komputasi*. **8**(2): 69-76.

World Health Organization. 2018. Diabetes : Key facts. https://www.who.int/news-room/fact-sheets/detail/diabetes/. Diakses pada tanggal 30 Oktober 2022.

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

Zaki. 2014. *Data Mining and Analysis : Fundamental Concepts and Algorithms*. Cambridge University Press 32. New York.