

**KLASIFIKASI PENDERITA PENYAKIT *DIABETES MELLITUS*
MENGUNAKAN METODE *SUPPORT VECTOR MACHINE***

(Skripsi)

Oleh

FANNI LUFIANA

NPM 1617051022



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021**

**KLASIFIKASI PENDERITA PENYAKIT *DIABETES MELLITUS*
MENGUNAKAN METODE *SUPPORT VECTOR MACHINE***

Oleh

FANNI LUFIANA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA ILMU KOMPUTER

Pada

**Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung**



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021**

ABSTRACT

CLASSIFICATION OF DIABETES MELLITUS DISEASE USING THE SUPPORT VECTOR MACHINE METHOD

By

FANNI LUFIANA

Diabetes Mellitus (DM) is a non-communicable disease characterized by high blood glucose levels caused by disruptions in the body's metabolic system, which prevents it from producing the hormone insulin. Diabetes is due to a combination of heredity (genetics), environmental factors, and unhealthy lifestyles. Hemoglobin A1c is a blood test used to diagnose and manage diabetes patients. The purpose of this study is to provide a predictive model for diabetic classification, determine and compare the results of feature correlation analysis on the variables used, and evaluate the results of the support vector machine method's classification performance. There are many ways to diagnose diabetes, and the support vector machine is one of the machine learning algorithms used in this study's classification case (SVM). The dataset is cleaned and then transformed before being used in the SVM model. With testing techniques and dataset distribution using the 10-fold cross-validation method, the SVM model uses three kernels: linear, Gaussian, and polynomial. There are 34 variables in the evaluated dataset with 84900 records. The study's findings yielded a simple prediction system model for diabetic classification. It aims to make it easier for the user to find out the prediction results and get an accuracy value of 82.76% from the gaussian kernel.

Keywords: Diabetes Mellitus, HbA1c, Klasifikasi, K-Fold Cross Validation, Support Vector Machine

ABSTRAK

KLASIFIKASI PENDERITA PENYAKIT *DIABETES MELLITUS* MENGUNAKAN METODE *SUPPORT VECTOR MACHINE*

Oleh

FANNI LUFIANA

Diabetes Mellitus (DM) merupakan salah satu jenis penyakit tidak menular yang ditandai dengan meningkatnya kadar glukosa dalam darah akibat gangguan sistem metabolisme dalam tubuh sehingga tidak mampu memproduksi hormon insulin. Penyakit diabetes bisa terjadi karena faktor keturunan (genetika) serta faktor pengaruh lingkungan dan gaya hidup yang tidak sehat. Dalam mengukur kadar glukosa, hemoglobin A1c digunakan untuk mendiagnosa dan mengontrol kondisi penderita DM. Tujuan dari penelitian ini memberikan model prediksi untuk klasifikasi penderita diabetes, menentukan dan mengukur perbandingan dari hasil analisis korelasi fitur terhadap variabel yang digunakan, dan mengevaluasi hasil kinerja klasifikasi dari metode *support vector machine*. Terdapat banyak cara dalam mendiagnosa penyakit diabetes, salah satu algoritma *machine learning* yang digunakan dalam kasus klasifikasi pada penelitian ini yaitu *support vector machine* (SVM). *Dataset* dibersihkan lalu di transformasi terlebih dahulu sehingga siap untuk dimasukkan kedalam model SVM. Terdapat tiga jenis *kernel* yang digunakan dalam model SVM yaitu *linear*, *Gaussian*, dan *polynomial* dengan teknik pengujian dan pembagian *dataset* menggunakan metode *10-fold cross validation*. *Dataset* yang dievaluasi memiliki 34 variabel dengan 84900 *records*. Hasil penelitian memperoleh sebuah model sistem sederhana prediksi untuk klasifikasi penderita diabetes yang bertujuan untuk memudahkan pengguna dalam mengetahui hasil prediksi, serta mendapatkan nilai hasil akurasi sebesar 82,76% dari *kernel gaussian*.

Kata kunci: Diabetes Mellitus, HbA1c, Klasifikasi, K-Fold Cross Validation, Support Vector Machine

Judul Skripsi : **KLASIFIKASI PENDERITA PENYAKIT *DIABETES MELLITUS* MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE***

Nama Mahasiswa : **Fanni Lufiana**

Nomor Pokok Mahasiswa : **1617051022**

Jurusan : **Ilmu Komputer**

Fakultas : **Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing



Favorisen R. Lumbanraja, Ph.D.
NIP. 19830110 200812 1 002



Yunda Heningtyas, S.Kom., M.Kom.
NIP. 19890108 201903 2 014

2. Ketua Jurusan Ilmu Komputer



Didik Kurniawan, S.Si., M.T.
NIP 19800419 200501 1 004

MENGESAHKAN

1. Tim Penguji

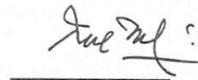
Ketua : Favorisen R. Lumbanraja, Ph.D.



Sekretaris : Yunda Heningtyas, S.Kom., M.Kom.



Penguji
Bukan Pembimbing : Dr. Ir. Kurnia Muludi, M.S.Sc.



2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



Dr. Eng. Satripto Dwi Yuwono, S.Si., M.T.
NIP 19740705 200003 1 001



Tanggal Lulus Ujian Skripsi : **01 Juli 2021**

PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya yang berjudul "**Klasifikasi Penderita Penyakit *Diabetes Mellitus* Menggunakan Metode *Support Vector Machine***" merupakan karya saya sendiri dan bukan karya orang lain. Semua tulisan yang tertuang di skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila dikemudian hari terbukti skripsi saya merupakan hasil penjiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang saya terima.

Bandar Lampung, 01 Juli 2021



Fanni Lufiana
NPM. 1617051022

RIWAYAT HIDUP



Penulis dilahirkan di Gisting pada tanggal 13 Januari 1999, sebagai anak pertama dari tiga bersaudara dengan ayah yang bernama Ibdiarman dan ibu bernama Gusmiati. Penulis menyelesaikan Pendidikan formal pertama kali di TK PKK Sukarame Talang Padang pada tahun 2004. Pendidikan dasar di SD Negeri 3 Talang Padang yang diselesaikan pada tahun 2010. Pendidikan menengah pertama di MTs Negeri 2 Tanggamus dan lulus pada tahun 2013, kemudian, penulis melanjutkan pendidikan menengah atas di MAN 1 Pringsewu dan lulus pada tahun 2016.

Tahun 2016 penulis melanjutkan pendidikan dan terdaftar sebagai mahasiswa jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN. Semasa kuliah, penulis terdaftar sebagai anggota bidang internal periode 2017 dalam organisasi Himpunan Mahasiswa Ilmu Komputer (HIMAKOM). Selama menjadi mahasiswa penulis melakukan beberapa kegiatan antara lain.

1. Melaksanakan Karya Wisata Ilmiah (KWI) di Desa Margosari Pagelaran Utara Pringsewu, pada bulan Januari tahun 2017.
2. Melaksanakan kerja praktik di Badan Pusat Statistik Kabupaten Pringsewu pada akhir bulan Desember 2018 hingga awal bulan Februari 2019.
3. Melaksanakan kegiatan Kuliah Kerja Nyata (KKN) di Desa Sukabumi, Kecamatan Batu Brak, Lampung Barat pada bulan Juni tahun 2019.

PERSEMBAHAN

Segala puji dan syukur saya ucapkan kepada Allah Subhanahu Wa Ta'ala atas segala rahmat dan ridho-Nya yang selalu memberikan keyakinan, kekuatan, kesabaran serta kelancaran kepadaku sehingga skripsi ini dapat diselesaikan.

Aku persembahkan karya kecilku ini untuk:

Mama dan Papa yang tak henti-hentinya mendoakan kesuksesan putra-putrinya, memberikan kasih sayang, pengorbanan, semangat, nasihat, perhatian, motivasi, dan dukungan moril maupun materiil dengan penuh kesabaran dan ketulusan hati, semua itu tidak ternilai berharganya dari apapun.

Abang-Abang, Uni-Uni, dan Adek-Adek yang aku sayangi. Terima kasih atas doa, kasih sayang, semangat, dan dukungan yang diberikan kepadaku.

Sahabat-Sahabatku, terimakasih telah menyemangatiku dan mendukungku dalam kebersamaan yang diberikan kepadaku tidak akan terlupakan

Keluarga Ilmu Komputer 2016,

*Serta almamater tercinta,
UNIVERSITAS LAMPUNG.*

MOTTO

“Boleh jadi kamu membenci sesuatu, padahal ia amat baik bagimu, dan boleh jadi (pula) kamu menyukai sesuatu, padahal ia amat buruk bagimu. Allah mengetahui, sedang kamu tidak mengetahui”

(Q.S. Al- Baqarah: 216)

“Be yourself, everyone else is already taken”

(Oscar Wild)

“Life isn't about being perfect, it's about accomplishing your dreams”

(JJK)

SANWACANA

Alhamdulillah rabbil'alam, puji syukur kehadiran Allah *Subhanahu wa ta'ala*, yang telah melimpahkan segala rahmat, hidayah dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Klasifikasi Penderita Penyakit *Diabetes Mellitus* Menggunakan Metode *Support Vector Machine*" ini dengan baik. Shalawat serta salam semoga senantiasa tercurahkan kepada Nabi Muhammad *Shalallahu 'alaihi wa sallam*, keluarganya, sahabatnya, dan mudah-mudahan kita adalah umat yang diberikan syafa'atnya di hari akhir kelak, aamiin.

Dalam penyusunan skripsi ini penulis menyadari banyak mendapat bimbingan, dukungan, dan motivasi dari berbagai pihak yang dengan tulus telah membantu penulis untuk menyelesaikan skripsi ini. Penulis sangat bersyukur dan mengucapkan terima kasih sedalam-dalamnya kepada:

1. Kedua orang tua, Papa dan Mama, adik-adikku Ferri Gusman dan Firmansyah, serta saudara-saudariku yang selalu mendoakan, memberikan kasih sayang, semangat, motivasi dan dukungan moril maupun materiil.
2. Bapak Favorisen R. Lumbanraja, Ph.D., selaku dosen pembimbing utama atas kesediaannya dan kesabarannya dalam membimbing, memberi arahan, nasihat dan ilmunya sehingga skripsi ini dapat diselesaikan dengan baik.
3. Ibu Yunda Heningtyas, S.Kom., M.Kom., selaku dosen pembimbing k atas kesediaannya dalam memberikan bimbingan, nasihat, juga kritik, saran guna penyempurnaan dalam penulisan skripsi ini.
4. Bapak Dr. Ir. Kurnia Muludi, M.S.Sc., selaku dosen pembahas skripsi yang telah memberikan saran dan masukan yang bermanfaat untuk perbaikan dalam penyusunan skripsi ini.
5. Bapak Dr. Eng. Suropto Dwi Yuwono, M.T. selaku Dekan FMIPA Universitas Lampung.

6. Bapak Didik Kurniawan, S.Si., MT., selaku Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung.
7. Ibu Astria Hijriani, S.Kom., M.Kom., selaku Sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
8. Bapak dan Ibu Dosen Jurusan Ilmu Komputer Universitas Lampung yang telah memberikan ilmu dan pengetahuan hidup selama penulis menjadi mahasiswa.
9. Ibu Ade Nora Maela, Pak Zainudin, dan Mas Nofal yang telah memudahkan segala urusan administrasi penulis di Jurusan Ilmu Komputer.
10. Teman-teman seperbimbingan Jojo, Rachel, Danu, Emes, Friska, Silfia, dan Indah yang telah menemani dan membantu untuk sama-sama belajar dalam menyusun sampai menyelesaikan skripsi ini.
11. Sahabat-sahabat terbaikku Yani, Suti, Nada, Imel, Mput, Sofa, Anita, Monica, Zakia, Shabrina, Dara, Anita Dwi, Mela, Yeni, Sisda, Maya, Irmaya, Vita, Ade ria, Sarifah, Destia, Rizka, Atika, dan Intan yang telah menjadi teman *sharing*, canda tawa, kebersamaan, dukungan dan semangat yang telah diberikan selama ini, semoga silaturahmi ini tetap selalu terjaga.
12. Teman-teman satu angkatan Jurusan Ilmu Komputer 2016 atas kebersamaan, kenangan, dan kerjasama selama menjalankan masa studi di Univ Lampung.

Penulis menyadari bahwa dalam penulisan skripsi ini masih banyak terdapat kekurangan karena masih terbatasnya pengetahuan, pengalaman, dan kemampuan penulis. Oleh karena itu, penulis sangat mengharapkan saran dan kritik yang membangun untuk penelitian ini sebagai bahan pertimbangan untuk karya tulis yang akan datang. Semoga skripsi ini dapat bermanfaat bagi semua pihak.

Bandar Lampung, 01 Juli 2021

Fanni Lufiana
NPM. 1617051022

DAFTAR ISI

Halaman

DAFTAR ISI	xiii
DAFTAR GAMBAR	xvi
DAFTAR TABEL	xvii
DAFTAR KODE PROGRAM	xviii
I. PENDAHULUAN	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	5
1.3. Batasan Masalah.....	5
1.4. Tujuan Penelitian.....	6
1.5. Manfaat Penelitian.....	6
II. TINJAUAN PUSTAKA	7
2.1. Penelitian Terdahulu.....	7
2.2. <i>Diabetes Mellitus</i> (DM).....	8
2.2.1 Pengertian <i>Diabetes Mellitus</i> (DM)	8
2.2.2 Klasifikasi <i>Diabetes Mellitus</i> (DM).....	9
2.3. <i>Machine Learning</i>	10
2.4. <i>Preprocessing</i>	11
2.5. Analisis Korelasi Fitur.....	12

2.6.	<i>Cross Validation</i>	13
2.6.1	<i>K-Fold Cross Validation</i>	13
2.6.2	<i>Leave-One-Out-Cross-Validation (LOOCV)</i>	14
2.6.3	<i>Hold-Out Validation</i>	15
2.7.	<i>Support Vector Machine (SVM)</i>	16
2.8.	<i>Confusion Matrix</i>	18
III. METODE PENELITIAN		20
3.1.	Tempat dan Waktu Penelitian	20
3.1.1	Tempat Penelitian.....	20
3.1.2	Waktu Penelitian	20
3.2.	Data dan Alat.....	22
3.2.1	Data	22
3.2.2	Alat.....	25
3.3.	Metode.....	26
IV. HASIL DAN PEMBAHASAN.....		29
4.1	<i>Preprocessing</i>	29
4.2	Analisis Korelasi Fitur.....	32
4.3	Percobaan Hasil Kinerja Dari Klasifikasi SVM.....	33
4.3.1	<i>Import Data</i> dan Penganalisan Data.....	33
4.3.2	Pengelompokan dan Pemanggilan Data Berdasarkan Kelas.....	34
4.3.3	Pembagian <i>Dataset</i>	35
4.3.4	Membuat Model Klasifikasi.....	35
4.3.5	Klasifikasi Dengan SVM <i>Linear, Gaussian, dan Polynomial</i>	36
4.4	Model Hasil Prediksi	38
4.5	Pembahasan	41

V. SIMPULAN DAN SARAN	42
5.1 Simpulan.....	42
5.2 Saran.....	43
DAFTAR PUSTAKA	44

DAFTAR GAMBAR

Gambar	Halaman
1. Prevelensi DM Berdasarkan Diagnosis Dokter Pada Penduduk Semua Umur Menurut Provinsi Lampung 2018.....	2
2. Prevalensi diabetes melitus berdasarkan diagnosis dokter pada penduduk semua umur menurut Kabupaten/Kota di Provinsi Lampung	3
3. Contoh Ilustrasi <i>k-fold cross validation</i> untuk $k=5$	14
4. Contoh Ilustrasi <i>Leave-One-Out-Cross-Validation (LOOCV)</i>	15
5. Contoh Ilustrasi <i>Hold-Out Validation</i>	15
6. Hyperplane terbaik yang memisahkan class-1 dan class +1	16
7. Tahapan Metode Penelitian.....	26
8. Tampilan Awal Sistem.....	38
9. Tampilan Hasil Analisis Model Prediksi Dengan <i>Shiny</i>	39
10. Grafik Hasil Prediksi Menggunakan <i>Kernel Linear, Gaussian, dan Polynomial</i>	41

DAFTAR TABEL

Tabel	Halaman
1. Penelitian Terdahulu	7
2. Interpretasi Terhadap Koefisien Korelasi (Astuti, 2017).....	13
3. Tabel <i>Confusion Matrix</i> (Kohavi & Provost, 1998)	18
4. Gantt chart pelaksanaan penelitian beserta metode yang digunakan	21
5. Dataset Penelitian.....	23
6. Deskripsi Atribut Penelitian.....	24
7. contoh data yang kosong / tidak dapat dibaca.....	29
8. Perubahan Jumlah Data Norm (DM kontrol baik), >7 (DM kontrol sedang), >8 (DM kontrol buruk) dan None Setelah Pembersihan Data.....	30
9. Data Hasil <i>Transformation Data</i>	31
10. Jumlah Data Hasil Analisis Korelasi.....	33
11. Hasil Percobaan Model SVM Klasifikasi Menggunakan 4 Kelas Data Dengan <i>Kernel Linear, Gaussian, dan Polynomial Dengan 10-Fold Cross Validation.</i>	37

DAFTAR KODE PROGRAM

Kode Program	Halaman
1. Kode Program Proses Data Transformasi	31
2. Kode Menampilkan Hasil Analisis Korelasi Fitur	32
3. Menggunakan library dari packages-packages yang telah diinstal.....	34
4. Kode Import Data dan Penganalisisan Data.....	34
5. Kode Pengelompokan dan Pemanggilan Data Berdasarkan Kelas.....	34
6. Kode Pembagian Data set.....	35
7. Kode Membuat Model Klasifikasi	35
8. Kode Melakukan Prediksi dengan Model Klasifikasi.....	36
9. Kode Mencari Kinerja Dari Hasil Prediksi.....	36

I. PENDAHULUAN

1.1. Latar Belakang

Diabetes Mellitus (DM) merupakan salah satu jenis penyakit tidak menular yang saat ini masih mengancam keberlangsungan hidup manusia. DM merupakan penyakit kronis yang ditandai dengan meningkatnya kadar glukosa dalam darah akibat gangguan sistem metabolisme dalam tubuh sehingga organ pankreas tidak mampu memproduksi hormon insulin sesuai kebutuhan tubuh (Gradiana & Irhamah, 2014). Gejala umum dari DM antara lain, yaitu *polyuria* (keinginan buang air kecil yang terlalu sering), *polydipsia* (rasa haus yang tidak berhenti), dan *polyphagia* (tingginya ritme rasa lapar) (Aljumah, Ahamad, & Siddiqui, 2012).

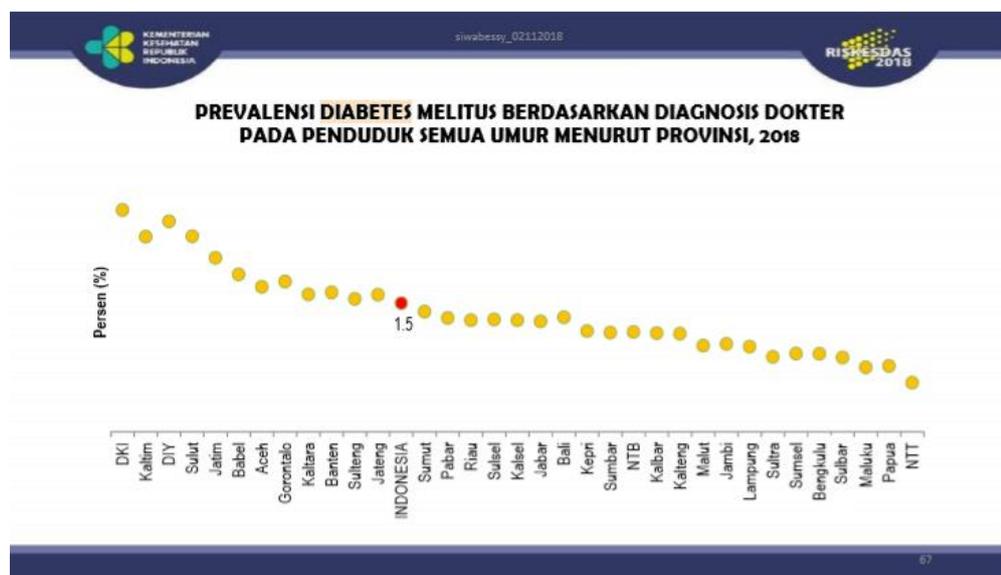
Seorang penderita DM cenderung tidak sadar bahwa dirinya memiliki penyakit DM, sehingga seringkali terjadi keterlambatan dalam penanganannya. DM biasa disebut dengan *the silent killer* karena penyakit ini dapat mengenai semua organ tubuh dan menimbulkan berbagai macam keluhan. Penyakit yang akan ditimbulkan antara lain gangguan penglihatan mata, katarak, penyakit jantung, sakit ginjal, impotensi seksual, luka sulit sembuh dan membusuk/gangren, infeksi paru-paru, gangguan pembuluh darah, stroke dan sebagainya. Adapun faktor penyebab terjadinya penyakit DM seperti faktor keturunan (genetika), faktor pengaruh lingkungan dan gaya hidup yang tidak sehat (American Diabetes Association, 2010).

HbA1C merupakan glukosa stabil yang terikat pada sel darah merah (hemoglobin) melalui proses non-enzimatis. Pemeriksaan HbA1C berfungsi

sebagai indikator dalam memantau kontrol gula darah jangka panjang, diagnosis, dan mengontrol kondisi penderita DM. pemeriksaan HbA1C dilakukan secara berkala setiap 3-6 bulan sekali. Pemeriksaan HbA1C dapat digolongkan menjadi tiga variabel, yakni DM kontrol baik (kadar HbA1C < 6.5%), DM kontrol sedang (kadar HbA1C 6.5% - 8%), DM kontrol buruk atau komplikasi (kadar HbA1C \geq 8%) (Que, Yasa, & Lestari, 2013).

Menurut data dari *International Diabetes Federation* (IDF) pada tahun 2019, Indonesia menempati urutan ke-7 dalam kasus jumlah penderita penyakit DM yaitu sekitar 10.7 juta jiwa, jumlah ini tentunya akan terus bertambah dan diprediksikan pada tahun 2030 jumlah penderita DM di Indonesia akan mencapai angka 13.7 juta jiwa (International Diabetes Federation, 2019).

Menurut hasil laporan nasional riset kesehatan dasar, prevalensi DM berdasarkan diagnosis dokter pada penduduk semua umur menurut provinsi, yaitu 1.5% dan Provinsi Lampung memiliki prevalansi DM, yaitu 0.99% (Tim Riskesdas, 2019). Berikut Prevelensi DM Berdasarkan Diagnosis Dokter Pada Penduduk Semua Umur Menurut Provinsi Lampung 2018 bisa dilihat di Gambar 1.



Gambar 1. Prevelensi DM Berdasarkan Diagnosis Dokter Pada Penduduk Semua Umur Menurut Provinsi Lampung 2018 (Tim Riskesdas, 2019)

Menurut hasil laporan Provinsi Lampung Riset kesehatan dasar pada Tabel 1, Kabupaten Metro memiliki prevalansi DM paling tinggi, yaitu 2.26% berdasarkan diagnosis dokter pada penduduk semua umur menurut kabupaten/kota di Provinsi Lampung. Prevalansi paling rendah ada di Kabupaten Pesisir Barat, yaitu 0.54% (Tim Riskesdas, 2019). Untuk Prevalensi DM berdasarkan diagnosis dokter pada penduduk semua umur menurut Kabupaten/Kota di Provinsi Lampung dapat dilihat pada Gambar 2.

Kabupaten/Kota	Diabetes Melitus berdasarkan Diagnosis Dokter*		
	%	95% CI	
Lampung Barat	0,68	0,37	1,23
Tanggamus	0,59	0,36	0,95
Lampung Selatan	0,82	0,48	1,38
Lampung Timur	1,14	0,76	1,69
Lampung Tengah	0,89	0,59	1,34
Lampung Utara	0,93	0,59	1,48
Way Kanan	1,05	0,69	1,60
Tulang Bawang	0,62	0,33	1,16
Pesawaran	0,70	0,40	1,22
Pringsewu	1,13	0,72	1,76
Mesuji	0,96	0,58	1,57
Tulangbawang Barat	0,66	0,36	1,22
Pesisir Barat	0,54	0,22	1,33
Kota Bandar Lampung	1,63	1,15	2,29
Kota Metro	2,26	1,56	3,27
Prov. Lampung	0,99	0,86	1,13

* Semua Umur

Gambar 2. Prevalensi diabetes melitus berdasarkan diagnosis dokter pada penduduk semua umur menurut Kabupaten/Kota di Provinsi Lampung (Tim Riskesdas, 2019)

Berbagai kasus yang berkaitan dengan pengelompokan objek dapat diselesaikan lebih mudah dengan menerapkan teknik klasifikasi. Sebagai contoh, pada bidang medis teknik klasifikasi dapat digunakan untuk mengklasifikasikan tingkat penyakit yang diderita oleh pasien sehingga memudahkan dokter dalam memberikan solusi pengobatan yang tepat. Untuk memecahkan masalah klasifikasi, berbagai macam metode dalam *machine*

learning dapat diterapkan. Ketepatan dalam pengklasifikasian objek sangat penting, teknik klasifikasi yang baik adalah metode yang menghasilkan kesalahan yang kecil dan metode tersebut dapat melakukan klasifikasi dengan baik, diantara teknik klasifikasi yang telah ada, yaitu metode SVM (*Support Vector Machine*) (Yunita, 2016).

SVM merupakan salah satu algoritma *machine learning* yang populer untuk melakukan klasifikasi dan regresi. SVM menjadi metode yang kuat untuk pola klasifikasi dan regresi, memiliki tingkat keberhasilan yang tinggi saat diterapkan di berbagai bidang (Qi, Tian, & Shi, 2013). Keunggulan dari metode SVM yaitu, dapat mengatasi masalah klasifikasi dan regresi dengan linier ataupun non-linier *kernel* yang dapat menjadi satu kemampuan algoritma pembelajaran untuk klasifikasi serta regresi, memiliki akurasi tinggi dan tingkat kesalahan yang *relative* kecil dan dapat digunakan untuk melakukan klasifikasi (Maimon & Rokach, 2010).

Penelitian mengenai penyakit DM menggunakan *machine learning* pernah dilakukan oleh Yusa, Utami, & Luthfi (2016), yaitu tentang analisis komperatif evaluasi algoritma klasifikasi pada readmisi pasien diabetes. Teknik klasifikasi pada penelitian ini dibutuhkan guna mencari solusi dalam penentuan algoritma yang memiliki performa terbaik. Model algoritma klasifikasi berupa *Decision Tree*, *k-Nearest Neighbor (k-NN)* dan *Naïve Bayes* dengan berbagai parameter *setting* yang akan dievaluasi performanya berdasarkan nilai performa *accuracy* dengan metode *10-Fold Cross Validation*.

Penelitian yang dilakukan oleh Puspitasari, Ratnawati, & Widodo (2018) tentang klasifikasi penyakit gigi dan mulut menggunakan metode *support vector machine*. Sistem klasifikasi yang digunakan pada penelitian ini yakni menggunakan metode SVM, karena metode SVM dapat mengatasi masalah klasifikasi dan regresi dengan linier maupun non-linier. Proses yang

dilakukan bersifat non-linear sehingga kernel yang digunakan, yakni kernel RBF.

Penelitian yang dilakukan oleh Kumari & Chitra (2013), yaitu tentang *Classification Of Diabetes Disease Using Support Vector Machine*. Penelitian ini dibuatkan untuk mendapatkan hasil pengukuran performa seperti *accuracy*, *sensitivity*, dan *specificity* dari metode SVM untuk klasifikasi penderita penyakit DM.

1.2. Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah mengklasifikasi penderita penyakit DM menggunakan metode SVM serta mengevaluasi hasil kinerja klasifikasi metode SVM.

1.3. Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan untuk meneliti penyakit DM, yaitu data dari *Diabetes 130-US Hospitals For Years 1999-2008 dataset* yang bersumber dari *UCI Machine Learning Repository* yang dapat di akses melalui <https://archive.ics.uci.edu/ml/datasets/diabetes+130us+hospitals+for+years+1999-2008> dengan data awal yang terdiri dari 101766 *dataset*
2. Metode yang digunakan dalam klasifikasi penderita penyakit DM yaitu SVM dengan *kernel* yang digunakan yaitu *Linear*, *Polynomial*, dan *Gaussian*.
3. Hasil evaluasi dari kinerja metode SVM menggunakan 3 parameter yaitu *accuracy*, *sensitivity* dan *specificity*.

1.4. Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini, yaitu:

1. Memberikan model prediksi untuk klasifikasi penderita DM menggunakan data *Diabetes 130-US Hospitals For Years 1999-2008 dataset*
2. Menentukan dan mengukur perbandingan dari hasil analisis korelasi fitur terhadap variabel
3. Mengevaluasi hasil kinerja klasifikasi metode SVM pada penderita penyakit DM berupa *accuracy*, *sensitivity*, dan *specificity*.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini antara lain, yaitu:

1. Memberikan model yang bisa digunakan untuk pemerintah ataupun masyarakat dalam melakukan penanganan kejadian penyakit DM
2. Mengetahui hasil kinerja klasifikasi dari metode SVM pada penderita penyakit DM.

II. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Untuk melakukan suatu penelitian diperlukan adanya hasil-hasil penelitian sebelumnya yang berkaitan dengan penelitian ini guna untuk membandingkan hasil akurasi pada penelitian tersebut. Adapun Tabel 1 menunjukkan beberapa penelitian sebelumnya yang akan membantu dan membandingkan hasil akurasi pada penelitian yang akan dibuat.

Tabel 1. Penelitian Terdahulu

Judul	Tahun	Data	Metode	Hasil / Evaluasi
Andika Prasatya, Riki Ruli A Siregar & Rakhmat Arianto	2020	<i>Diabetes 130-US Hospitals For Years 1999-2008 Dataset.</i> Adapun data awal berjumlah 101767 data. Setelah dilakukan pembersihan data, data berjumlah menjadi 17019 data.	<i>K-Means</i> dan C4.5	<i>Accuracy</i> didapat sebesar 72%
V. Anuja Kumari & R. Chitra	2013	<i>The Pima Indian Diabetes Dataset</i> dengan data yang terdiri dari 460 data.	<i>Support Vector Machine (SVM)</i>	<i>Accuracy</i> didapat sebesar 78%, <i>Sensitivity</i> 80%, dan <i>Specificity</i> 76.5%

Berdasarkan Tabel 1 terdapat dua penelitian terdahulu yang dijadikan sebagai acuan untuk melakukan penelitian ini. Penelitian sebelumnya yang pertama dilakukan oleh Prasatya, Siregar, & Arianto (2020) meneliti tentang "Penerapan Metode *K-Means* dan C4.5 Untuk Prediksi Diabetes". Penelitian ini melakukan prediksi HbA1C dengan kombinasi metode *K-Means* dengan C4.5. Data yang digunakan dalam penelitian ini yaitu *Diabetes 130-US Hospitals For Years 1999-2008 Dataset*. Adapun data awal berjumlah 101767 data, namun setelah dilakukan pembersihan data, data berjumlah menjadi 17019 data. Hasil prediksi yang didapatkan akan dilakukan proses validasi untuk mengetahui tingkat keakurasian dengan menggunakan *K-Fold Cross Validation*. Nilai akurasi yang didapat sebesar 72%.

Penelitian yang kedua dilakukan oleh Kumari & Chitra (2013) meneliti tentang "*Classification Of Diabetes Disease Using Support Vector Machine*". Penelitian ini sistem klasifikasi yang digunakan yakni menggunakan metode SVM dengan kernel yang digunakan yakni kernel RBF. Data yang digunakan pada penelitian ini yaitu data *The Pima Indian diabetes dataset* dengan data *samples* yang digunakan yaitu 460 data. Hasil klasifikasi yang diperoleh dengan menggunakan metode SVM mempunyai rata – rata nilai *accuracy* sebesar 78%, *sensitivity* 80%, dan *specificity* 76.5%.

2.2. Diabetes Mellitus (DM)

2.2.1 Pengertian Diabetes Mellitus (DM)

Diabetes Mellitus (DM) merupakan suatu penyakit tidak menular kronis yang ditandai dengan ketidakmampuan tubuh untuk melakukan metabolisme karbohidrat, lemak, dan protein sehingga terjadinya peningkatan kadar gula darah (*hiperglikemia*) yang disebabkan oleh *resistensi* (menurunnya kadar hormon) insulin (Sulistria, 2013).

DM merupakan salah satu penyakit menahun yang ditandai dengan kadar gula darah melebihi normal yaitu kadar gula darah sewaktu $> 200 \text{ mg/dl}$, dan kadar gula darah puasa $> 126 \text{ mg/dl}$. DM tidak bisa disembuhkan, tetapi kadar gula dalam darah dapat dikendalikan melalui pilar pengendalian DM, yaitu dengan latihan jasmani, terapi gizi medis, intervensi farmakologis, dan edukasi, sehingga untuk dapat mencegah terjadinya komplikasi kronis maka diperlukannya pengendalian DM yang baik (PERKENI, 2011).

Pengobatan DM yang paling utama, yaitu mengubah gaya hidup terutama mengatur pola makan yang sehat dan seimbang, minum air hangat serta banyak berolahraga. Penelitian (Afrianti, Bahri, Yuniarti, & Elmatris, 2012) menyimpulkan bahwa dengan minum air putih menyebabkan terjadinya pemecahan gula, untuk membantu mengeluarkan zat-zat kimia seperti glukosa dan zat-zat melalui ginjal serta proses pembersihan organ tubuh, diperlukan jumlah cairan yang banyak dalam satu kali pemberian di pagi hari.

2.2.2 Klasifikasi *Diabetes Mellitus* (DM)

- a. *Diabetes Mellitus* tipe 1 atau IDDM (*Insulin Dependent Diabetes Mellitus*)
DM tipe 1 adalah penyakit gangguan metabolik yang ditandai oleh naiknya kadar gula darah akibat *destruksi* (kerusakan) sel beta pancreas karena suatu sebab tertentu yang menyebabkan produksi insulin tidak ada sama sekali, sehingga penderita sangat memerlukan tambahan insulin dari luar dengan cara disuntikan. DM tipe 1 biasa terjadi pada usia muda dibawah 30 tahun (Direktur Pengendalian PTM, 2008).

- b. *Diabetes Mellitus* tipe 2 atau INDDM (*Insulin Non-dependent Diabetes Mellitus*)
DM tipe 2 adalah penyakit gangguan metabolik yang ditandai oleh kenaikan kadar gula darah akibat kadar insulin yang ada pada tubuh

manusia tidak dapat bekerja dengan baik dan fungsi insulin untuk metabolisme tidak ada atau kurang. DM tipe 2 sering di derita oleh seorang yang berusia diatas 40 tahun. Biasanya disebabkan faktor keturunan, obesitas, kurang aktivitas, penyakit lain dan usia (Direktur Pengendalian PTM, 2008).

c. *Diabetes Melitus Gestasional* (DMG)

DMG adalah penyakit gangguan metabolik yang ditandai oleh kenaikan kadar gula darah yang terjadi pada wanita hamil. Biasanya terjadi pada usia 24 minggu masa kehamilan dan setelah melahirkan kadar gula darah kembali normal (Direktur Pengendalian PTM, 2008).

2.3. Machine Learning

Machine Learning (ML) atau pembelajaran mesin adalah salah satu disiplin ilmu dari *Computer Science* yang mempelajari bagaimana membuat komputer atau mesin agar mempunyai suatu kecerdasan. Menurut Ahmad (2017). ML yaitu pendekatan dalam AI (*Artificial Intelegent*) yang digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan suatu masalah atau melakukan otomatisasi. ML biasanya digunakan dalam perencanaan, diagnosis, prediksi, pengenalan, kontrol robot, serta tugas-tugas lainnya yang terkait dengan kecerdasan buatan. Ciri khas dari ML yaitu adanya proses pelatihan, pembelajaran atau *training*. Machine Learning mempunyai 2 tipe teknik yaitu *supervised learning* dan *unsupervised learning*.

1. *Supervised Learning*

Supervised Learning adalah salah satu tipe algoritma *machine learning* yang menggunakan *dataset* yang dikenal (*training dataset*) untuk membuat prediksi atau klasifikasi. Tujuan pembelajaran adalah membangun model yang dapat menghasilkan output yang benar untuk suatu data *input*. Contoh algoritma dalam *supervised learning* yaitu, *Decision tree*, KNN (*K-Nearest Neighbors*), *Logistic Regression*, SVM,

dan *Naive Bayes* (Nurhayati, Busman, & Iswara, 2019). Terdapat dua aplikasi utama dalam ML pada *supervised learning*, yaitu klasifikasi dan regresi.

Klasifikasi adalah metode dalam ML yang digunakan oleh mesin untuk memilah atau mengklasifikasikan obyek berdasarkan ciri tertentu sebagaimana manusia mencoba membedakan benda satu dengan yang lain, klasifikasi bertujuan untuk memisahkan data menjadi kelas-kelas tertentu. Sedangkan regresi digunakan oleh mesin untuk menerka keluaran dari suatu data masukan berdasarkan data yang sudah dipelajari dalam *data training* (Ahmad, 2017).

2. *Unsupervised learning*

Unsupervised learning adalah salah satu tipe algoritma *machine learning* yang digunakan untuk menarik kesimpulan dari *dataset* yang terdiri dari *input data labeled response*. Metode *unsupervised learning* yang paling umum adalah analisa cluster, yang digunakan pada analisa data untuk mencari pola-pola tersembunyi atau pengelompokan dalam data. Tujuan pembelajaran adalah membangun model yang dapat menemukan variabel tersembunyi pada data pelatihan. Selanjutnya, variabel tersembunyi tersebut dapat digunakan untuk kebutuhan pemodelan (*latent variable models*). Contoh *unsupervised learning* yaitu *K-means*, *Fuzzy clustering*, dan *Hierarchical clustering* (Nurhayati, Busman, & Iswara, 2019).

2.4. *Preprocessing*

Preprocessing merupakan tahapan sebelum proses pengklasifikasian yang diperlukan untuk membersihkan, menghilangkan, mengubah sumber data, baik itu berupa karakter non-alfabet maupun kata-kata yang tidak diperlukan. Hal ini bertujuan agar data yang digunakan lebih optimal ketika digunakan pada proses pengklasifikasiannya. Ada banyak faktor yang menimbulkan *problem* performansi klasifikasi, pertama dan terutama adalah bentuk dan

kualitas data, bila data mengandung noise, *redundansi* dan mempunyai data yang tidak relevan, kondisi tersebut membuat proses pengekstrasian fitur selama fase pelatihan lebih sulit (Muttaqin & Bachtiar, 2019). Sehingga, *preprocessing* data merupakan langkah yang penting dalam proses *machine learning*.

2.5. Analisis Korelasi Fitur

Analisis korelasi didefinisikan sebagai metode statistika yang digunakan untuk mengukur keeratan hubungan antara dua variabel tanpa mempersoalkan apakah variabel tersebut saling bergantung. Terdapat dua teknik korelasi yang sangat populer sampai sekarang, yaitu Korelasi Pearson Product Moment dan Korelasi Rank Spearman. Dari analisis korelasi yang dilakukan didapatkan suatu nilai yang disebut sebagai koefisien korelasi (Astuti, 2017).

Koefisien korelasi bisa bernilai positif atau negatif dan nilai koefisien korelasi berkisar antara -1 sampai dengan +1. Apabila nilai koefisien korelasi bernilai negatif (-1) maka korelasi kedua variabel memiliki hubungan yang berlawanan, artinya jika nilai variabel X tinggi, maka nilai variabel Y akan menjadi rendah (dan sebaliknya). dan apabila korelasi bernilai (+1) maka korelasi kedua variabel memiliki hubungan yang searah. Artinya jika nilai variabel X tinggi, maka nilai variabel Y akan tinggi pula.. Jika koefisien korelasi bernilai 0 artinya korelasi kedua variabel tidak terdapat hubungan (Astuti, 2017). Berikut adalah tabel interpretasi terhadap koefisien korelasi yang dapat digunakan sebagai acuan untuk mengetahui tinggi atau lemahnya hubungan korelasi antar variabel yang dapat dilihat pada Tabel 2.

Tabel 2. Interpretasi Terhadap Koefisien Korelasi (Astuti, 2017)

Besar Koefisien Korelasi (Positif atau Negatif)	Interpretasi Koefisien Korelasi
0,00	Tidak Ada Korelasi (Hubungan)
0,01 - 0,20	Korelasi Sangat Lemah
0,21 - 0,40	Korelasi Lemah
0,41 - 0,70	Korelasi Sedang
0,71 - 0,99	Korelasi Tinggi
1,00	Korelasi Sempurna

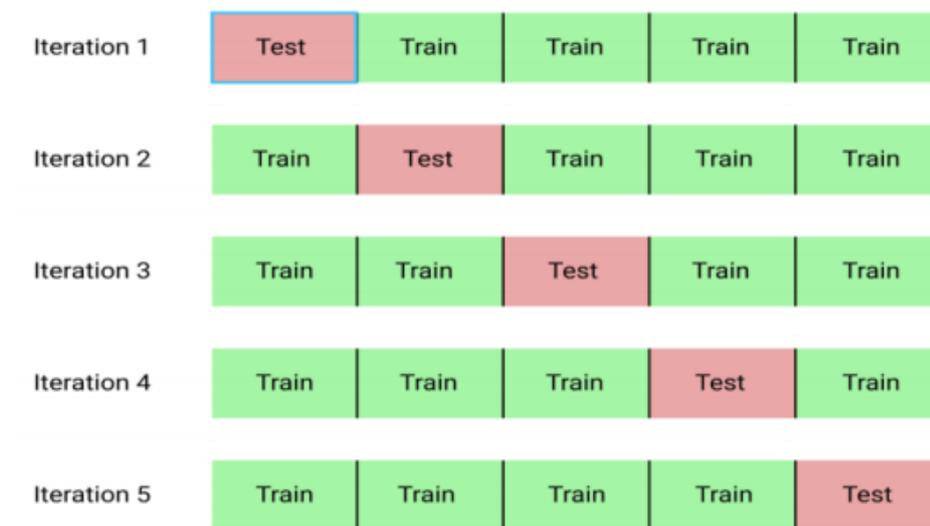
2.6. Cross Validation

Menurut Santoso (2007) *cross-validation* adalah salah satu cara menentukan parameter terbaik dari satu model dengan cara menguji besarnya error pada *data testing*. Dalam *cross-validation*, data dibagi sebanyak k bagian dengan ukuran sampel yang sama. Kemudian $k-1$ bagian digunakan sebagai data untuk *training* sistem dan sisanya untuk *testing*. Pengujian ini dilakukan sebanyak k kali dengan mengganti-ganti partisi yang berfungsi sebagai *data training* dan *data testing*. *Cross Validation* meliputi:

2.6.1 K-Fold Cross Validation

K-Fold Cross Validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut *input* yang acak (Pandie, 2012). *K-fold cross validation* diawali dengan membagi data sejumlah n *fold* yang diinginkan. Untuk penggunaan jumlah *fold* terbaik untuk uji validitas, dianjurkan menggunakan *10-fold cross validation* dalam model. Penggunaan *10 fold* ini dianjurkan karena merupakan jumlah *fold* terbaik untuk uji validitas (Kohavi & Provost, 1998).

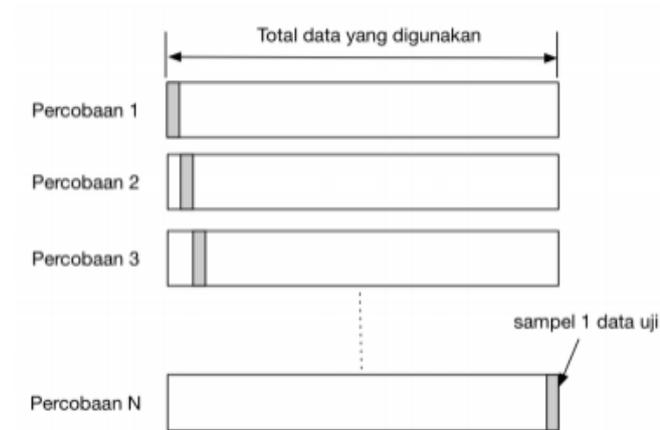
Menurut Fu (1994) *K-Fold Cross Validation* mengulang k-kali untuk membagi sebuah himpunan contoh secara acak menjadi k *subset* yang saling bebas, setiap ulangan disisakan satu *subset* untuk pengujian dan *subset* lainnya untuk pelatihan. Menurut Trevor, Robert, & Friedman (2008) dengan $K=5$ atau 10 dapat digunakan untuk memperkirakan tingkat kesalahan yang terjadi, sebab data *training* pada setiap *fold* cukup berbeda dengan data *training* yang asli. Adapun contoh *k-fold cross validation* bisa dilihat di Gambar 3.



Gambar 3. Contoh Ilustrasi *k-fold cross validation* untuk $k=5$
(Azis, Purnawansyah, Fattah, & Putri, 2020)

2.6.2 *Leave-One-Out-Cross-Validation (LOOCV)*

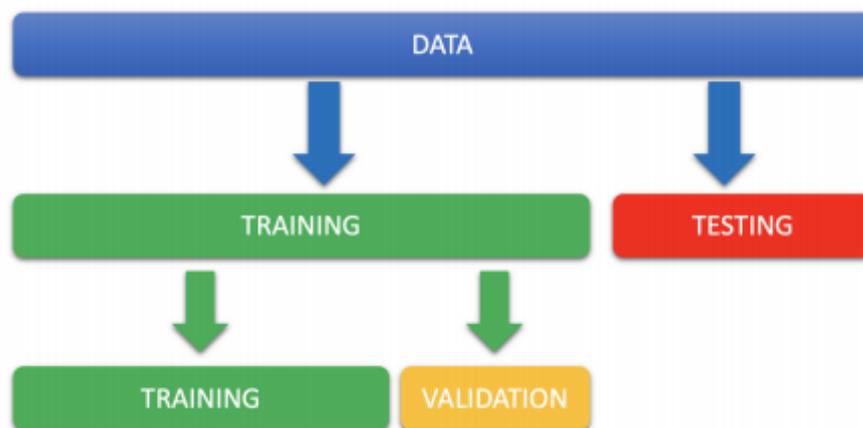
Leave-One-Out-Cross-Validation (LOOCV) Merupakan turunan dari metode *K-Fold Cross Validation*, di mana K yang dipilih adalah sebesar jumlah data. Misalkan dataset yang memiliki N sample, maka percobaan dilakukan sebanyak N kali. Setiap percobaan menggunakan sebanyak $N - 1$ sample untuk data training dan sisanya untuk testing. LOOCV membutuhkan waktu yang lebih lama yang berpotensi menyulitkan proses komputasi (Mamase & Buliali, 2016). Berikut contoh *Leave-One-Out-Cross-Validation (LOOCV)* pada Gambar 4.



Gambar 4. Contoh Ilustrasi *Leave-One-Out-Cross-Validation (LOOCV)* (Refaeilzadeh, Tang, & Liu, 2008)

2.6.3 *Hold-Out Validation*

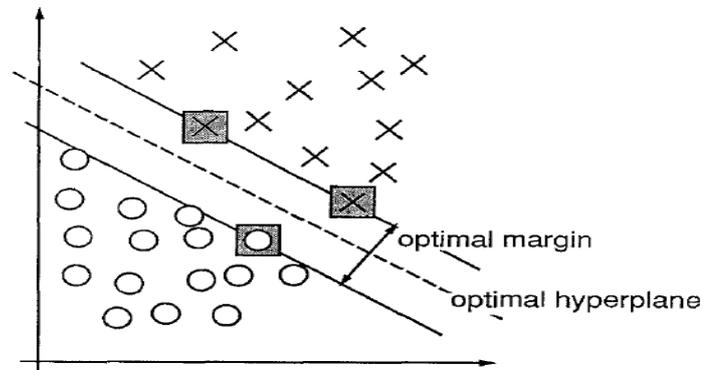
Metode *holdout* adalah metode yang akan menyediakan sejumlah data untuk digunakan sebagai data *testing*, dan sisanya sebagai data *training*. Kelemahannya adalah prosedur ini tidak menggunakan semua data yang tersedia dan hasilnya sangat tergantung pada pilihan untuk pelatihan (Witten, Frank, & Hall, 2011). Berikut Gambar 5 contoh ilustrasi cara kerja *Hold-Out Validation*.



Gambar 5. Contoh Ilustrasi *Hold-Out Validation* (Refaeilzadeh, Tang, & Liu, 2008).

2.7. Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan algoritma dengan metode *supervised learning* yang menganalisa data dan mengenali pola, digunakan untuk analisis klasifikasi dan regresi. Tujuan SVM adalah menghasilkan model (berdasarkan data training) yang memprediksi nilai target dari data *testing*. Menurut Cortes & Vapnik (1995) SVM merupakan sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi *linier* dalam sebuah ruang fitur berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan *learning bias* yang berasal dari teori pembelajaran statistik. SVM dijelaskan sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada ruang *input*. *Hyperplane* membagi ruang vektor ke dalam dua bagian, yang masing-masing berkorespondensi pada kelas yang berbeda. Berikut *Hyperplane* terbaik yang memisahkan *class-1* dan *class +1* pada Gambar 6.



Gambar 6. Hyperplane terbaik yang memisahkan class-1 dan class +1 (Cortes & Vapnik, 1995)

Gambar 6 memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class* yaitu bulat dan silang. *Problem* klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. Garis pemisah (*discrimination boundaries*) menunjukkan *hyperplane* pemisah terbaik antara kedua *class*

dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing *class*. *Pattern* yang paling dekat ini disebut sebagai *support vector*.

Optimal *hyperplane* menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua *class*, sedangkan titik merah dan biru yang berada di garis-garis putus adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM. Proses pembelajaran SVM adalah untuk menentukan *support vector*, kita hanya cukup mengetahui fungsi *kernel* yang dipakai, dan tidak perlu mengetahui wujud dari fungsi non-linear (Nugroho, Witarto, & Handoko, 2003). Menurut Bhavsar & Panchal (2012) terdapat empat *kernel* dasar yang dapat digunakan, sebagaimana persamaan 1, 2, 3, dan 4:

1. *Linear*

$$K(x_i, x_j) = x_i \cdot x_j + C \dots\dots\dots(1)$$

2. *Polynomial*

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d, \gamma > 0 \dots\dots\dots(2)$$

3. *Sigmoid*

$$K(x_i, x_j) = \tanh(K \langle x_i, x_j \rangle + r) \dots\dots\dots(3)$$

4. *Radial Basis Function (RBF)*, atau yang biasa dikenal dengan *Gaussian*

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \dots\dots\dots(4)$$

Dengan γ , r , dan d adalah parameter *kernel*. Penelitian ini menggunakan metode SVM dengan *kernel* yang digunakan ialah *kernel linear*, *gaussian*, dan *polynomial*.

2.8. Confusion Matrix

Menurut Rahman, Darmawidjadja, & Alamsyah (2017) *Confusion matrix* adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan. *Confusion matrix* bisa dilihat pada Tabel 3.

Tabel 3. Tabel *Confusion Matrix* (Kohavi & Provost, 1998)

<i>Correct Classification</i>	<i>Classified as</i>	
	<i>Actual "+"</i>	<i>Actual "-"</i>
<i>Predicted "+"</i>	TP (<i>True Positive</i>)	FP (<i>False Positive</i>)
<i>Predicted "-"</i>	FN (<i>False Negative</i>)	TN (<i>True Negative</i>)

Menurut Kohavi & Provost (1998) penjelasan tabel *Confusion Matrix* diatas, yaitu:

True Positive (TP) adalah jumlah data dengan nilai sebenarnya positif dan nilai prediksi positif

False Positive (FP) adalah jumlah data dengan nilai sebenarnya negatif dan nilai prediksi positif

False Negative (FN) adalah jumlah data dengan nilai sebenarnya positif dan nilai prediksi negatif

True Negative (TN) adalah jumlah data dengan nilai sebenarnya negatif dan nilai prediksi negatif

Menurut Kohavi & Provost (1998) nilai yang dihasilkan melalui metode *Confusion Matrix* adalah berupa evaluasi sebagai berikut :

Accuracy yaitu presentase seberapa akuratnya suatu model dalam mengklasifikasikan (prediksi) secara benar oleh algoritma. Fungsi dari *accuracy* dapat dilihat dari persamaan 5:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \times 100\% \dots\dots\dots (5)$$

Sensitivity merupakan kemampuan tes untuk menunjukkan individu mana yang menderita sakit dari seluruh populasi yang benar-benar sakit. *Sensitivity* dinyatakan dalam jumlah pengenalan entitas bernilai benar dibagi jumlah entitas yang dikenali sistem. Fungsi dari *sensitivity* dapat dilihat dari persamaan 6:

$$Sensitivity = \frac{TP}{(TP+FN)} \times 100\% \dots\dots\dots (6)$$

Specificity adalah kemampuan tes untuk menunjukkan individu mana yang tidak menderita sakit dari mereka yang benar-benar tidak sakit. Fungsi dari *specificity* dapat dilihat dari persamaan 7:

$$Specificity = \frac{TN}{(TN+FP)} \times 100\% \dots\dots\dots (7)$$

III. METODE PENELITIAN

3.1. Tempat dan Waktu Penelitian

3.1.1 Tempat Penelitian

Penelitian ini dilaksanakan di Laboratorium Rekayasa Perangkat Lunak (RPL) Gedung Matematika dan Ilmu Pengetahuan Alam Terpadu Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang beralamatkan di Jl. Soemantri Brojonegoro No. 1 Gedong Meneng, Bandar Lampung.

3.1.2 Waktu Penelitian

Penelitian ini dilaksanakan mulai dari minggu ketiga bulan Februari 2020 sampai dengan akhir bulan Februari 2021. Adapun *timeline* waktu pelaksanaan penelitian dijabarkan pada Tabel 4.

Seperti yang terlihat pada Tabel 4 penelitian dimulai pada minggu ketiga dan keempat pada bulan Februari 2020. Peneliti mulai mengumpulkan data dan menganalisa *dataset* yang telah didapatkan dari *Diabetes 130-US Hospitals For Years 1999-2008 Dataset*. Minggu pertama bulan Maret sampai minggu kedua bulan September 2020 dimulai proses *preprocessing data* dengan cara melakukan *Cleaning Data* pada *dataset*, *Feature Selection* pada *dataset*, dan *Data Transformation* pada *dataset*.

Proses selanjutnya, pada minggu ketiga bulan September sampai minggu keempat bulan Februari 2021 dilakukan proses *k-fold cross validation* dan membagi data menjadi data *training* dan data *testing*. Selanjutnya, di minggu dan bulan yang sama juga dilakukan proses *training* SVM menggunakan model klasifikasi dan proses *Testing / Evaluasi*. Penelitian ini juga memuat alokasi waktu untuk studi literatur dan penulisan laporan bab 1-3 dilakukan pada minggu kedua bulan Februari sampai minggu keempat bulan Mei 2020 dan untuk penulisan laporan bab 4-5 dilakukan pada minggu pertama bulan Februari sampai minggu ketiga bulan Maret 2021.

3.2. Data dan Alat

3.2.1 Data

Data penelitian yang digunakan berupa data dari *Diabetes 130-US Hospitals For Years 1999-2008 Dataset* yang bersumber dari *UCI Machine Learning Repository*. *Dataset* yang digunakan dapat di akses melalui <https://archive.ics.uci.edu/ml/datasets/diabetes+130us+hospitals+for+years+1999-2008>. Data yang dipakai dalam penelitian ini terdiri dari 84900 *record* dengan 34 atribut dan 1 atribut untuk kelas, yaitu atribut *A1Cresult* yang bernilai atau berlabel *Norm* (DM kontrol baik), >7 (DM kontrol sedang), >8 (DM kontrol buruk) dan *None*. Berikut Tabel 5, contoh *dataset* yang digunakan dalam penelitian ini

Tabel 5. Dataset Penelitian

race	Gender	Age	time_in_hospital	diag_1	diag_2	diag_3	max_glu_serum	A1Cresult	change	diabetesMed	readmitted
Caucasian	Female	[10-20)	3	276	250.01	255	None	None	Ch	Yes	>30
Caucasian	Male	[40-50)	1	197	157	250	None	None	Ch	Yes	NO
Caucasian	Male	[50-60)	3	414	411	250	None	None	No	Yes	>30
Caucasian	Male	[70-80)	5	428	492	250	None	None	No	Yes	>30
Caucasian	Female	[90-100)	12	434	198	486	None	None	Ch	Yes	NO
AfricanAmerican	Female	[40-50)	9	250.7	403	996	None	None	No	Yes	>30
AfricanAmerican	Male	[60-70)	7	157	288	197	None	None	Ch	Yes	<30
Caucasian	Female	[40-50)	7	428	250.43	250.6	None	None	Ch	Yes	<30
Caucasian	Male	[80-90)	10	428	411	427	None	None	No	Yes	NO
AfricanAmerican	Female	[60-70)	1	518	998	627	None	None	No	Yes	>30
AfricanAmerican	Male	[60-70)	12	999	507	996	None	None	Ch	Yes	NO
AfricanAmerican	Male	[50-60)	4	410	411	414	None	None	Ch	Yes	<30
Caucasian	Female	[50-60)	3	682	174	250	None	None	No	Yes	NO
AfricanAmerican	Male	[70-80)	5	402	425	416	None	None	No	Yes	>30
AfricanAmerican	Female	[70-80)	2	410	401	582	None	None	No	No	NO
AfricanAmerican	Female	[70-80)	3	189	496	427	None	None	Ch	Yes	NO
Other	Female	[50-60)	1	786	401	250	None	None	No	Yes	NO
Caucasian	Male	[80-90)	6	427	428	414	None	>7	Ch	Yes	NO
Caucasian	Female	[50-60)	2	996	585	250.01	None	None	No	Yes	>30
Caucasian	Male	[20-30)	10	277	250.02	263	None	None	Ch	Yes	>30
Caucasian	Female	[80-90)	5	428	410	414	None	None	Ch	Yes	>30
AfricanAmerican	Female	[60-70)	6	584	496	250.42	None	None	Ch	Yes	NO
AfricanAmerican	Male	[70-80)	1	462	250.01	276	None	None	Ch	Yes	>30

Berikut deskripsi mengenai atribut penelitian yang digunakan terdapat pada Tabel 6.

Tabel 6. Deskripsi Atribut Penelitian

Nama Atribut	Deskripsi
<i>Race</i>	Menjelaskan tentang jenis-jenis ras, yaitu <i>Caucasian, Asian, African American, Hispanic, dan Other</i>
<i>Gender</i>	Menjelaskan tentang jenis kelamin, yaitu Perempuan dan Laki-laki
<i>Age</i>	Menjelaskan tentang umur
<i>Time in Hospital</i>	Menjelaskan tentang jumlah / lama pasien di rawat di rumah sakit
Diag1	Menjelaskan hasil diagnosa utama
Diag2	Menjelaskan hasil diagnosa kedua
Diag3	Menjelaskan hasil diagnosa tambahan yang mendukung diagnosis kedua
<i>Glucose Serum Test Result</i>	Menjelaskan hasil dari tes serum glukosa
<i>A1c Test Result</i>	Menjelaskan hasil dari tes <i>hemoglobin A1c</i>
<i>Change Of Medication</i>	Menjelaskan jika ada perubahan dalam pengobatan (baik dosis obat / obat yang dipakai
<i>Diabetes Medication</i>	Menjelaskan jika ada pengobatan diabetes lain yang diresepkan
<i>Readmitted</i>	Menjelaskan tentang apakah pasien kembali di rawat inap atau tidak
<i>24 Features for Medications</i>	Nama-nama obat seperti <i>metformin, repaglide, nateglinide, chlorpropamide</i>

3.2.2 Alat

Untuk melakukan penelitian ini, dibutuhkan *hardware* dan beberapa *software* sebagai berikut:

A. Hardware

Laptop dengan spesifikasi *processor* Intel® Core™ i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), *Installed Physical Memory* (RAM) 8,00 GB, *system type* x64-based PC.

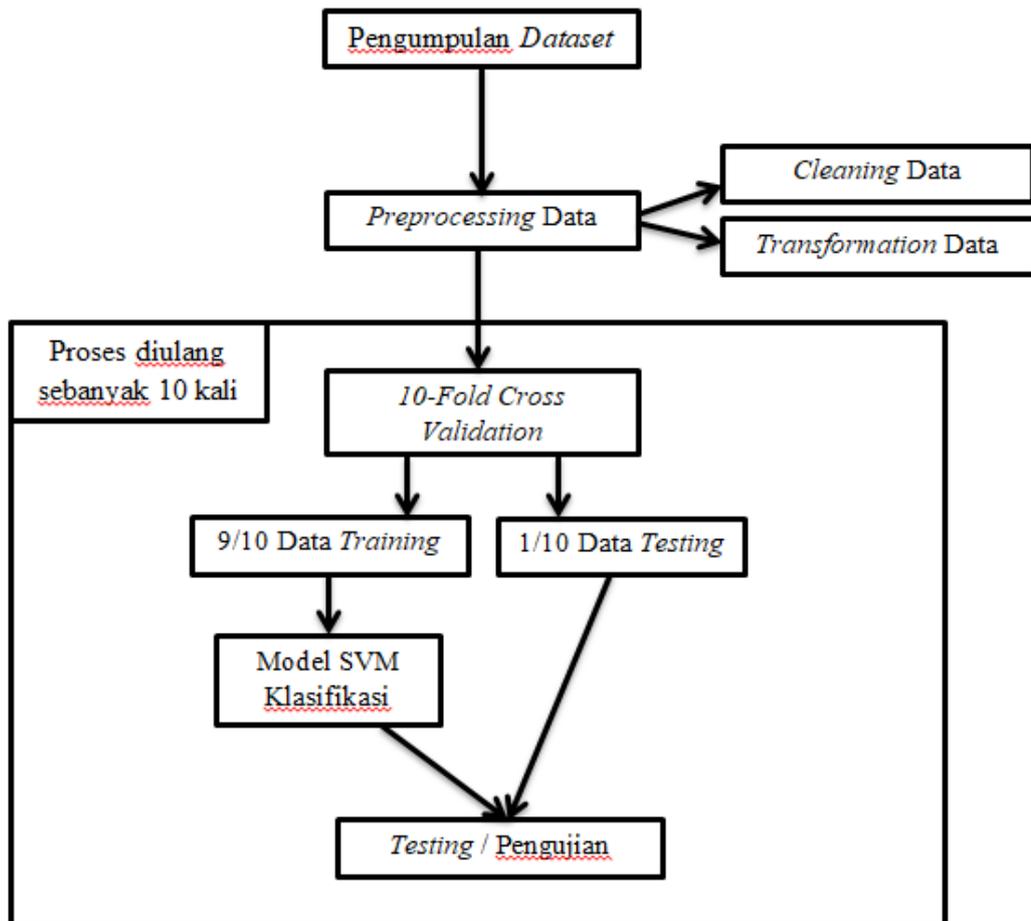
B. Software

Software yang digunakan pada penelitian ini, yakni:

- sistem operasi *Windows 10 Home Single Language*,
- Microsoft Excel 2013 untuk menampilkan data,
- *R studio for Windows* versi 3.5.1 sebagai bahasa pemrograman untuk mengolah dan menganalisis data penelitian. Adapun package pada R Studio yang digunakan penelitian ini adalah sebagai berikut:
 - *Package readr* versi 1.4.0 (Wickham, 2020), digunakan untuk mengimport data *file*
 - *Package e1071* versi 1.7-7 (Meyer, 2021), digunakan untuk melakukan klasifikasi dengan SVM.
 - *Package caret* versi 6.0-88 (Kuhn, 2021), (*Classification And Regresseion Training*) digunakan untuk menghitung kinerja algoritma klasifikasi menggunakan fungsi *confusion matrix*
 - *Package Hmisc* versi 4.5-0 (Harrel , 2021), digunakan untuk menampilkan proses analisis korelasi fitur.
 - *Package Shiny* versi 1.6.0 (Chang, 2021), digunakan untuk membangun web apps yang interkatif menggunakan kode R.

3.3. Metode

Metode yang dilakukan dalam melakukan penelitian tentang klasifikasi penderita penyakit *diabetes mellitus* menggunakan metode SVM dilakukan dengan beberapa tahap yang ditunjukkan pada Gambar 7, sebagai berikut:



Gambar 7. Tahapan Metode Penelitian

Penjelasan dari tahapan metode penelitian pada Gambar 5 adalah sebagai berikut:

1. Pengumpulan *Dataset* penyakit *diabetes mellitus*

Proses awal dalam melakukan penelitian ini yaitu dengan melakukan pengumpulan data yang berkaitan dengan penelitian. *Dataset* yang digunakan pada penelitian ini adalah data dari *Diabetes 130-US Hospitals*

For Years 1999-2008 Dataset. Data ini terdiri dari 84900 *record* dengan 1 label atau kelas dan 34 atribut lainnya yakni *Race, Gender, Age, Time In Hospital, Diag_1, Diag_2, Diag_3, Max_glu_serum, Metformin, Repaglinide, Nateglinide, Chlorpropamide, Glimepiride, Acetohexamide, Glipizide, Glyburide, Tolbutamide, Pioglitazone, Rosiglitazone, Acarbose, Miglitol, Troglitazone, Tolazamide, Examide, Citoglipton, Insulin, Glyburide-metformin, Glipizide-metformin, Glimepiride-pioglitazone, Metformin-Rosiglitazone, Metformin-pioglitazone, DiabetesMed, Change, Readmitted.*

2. *Preprocessing Data*

Preprocessing data dilakukan dengan 2 tahap yaitu, dengan cara penghapusan data / *Cleaning Data* dan *Transformation Data* / *Data Transformasi*. Setelah data dikumpulkan, *cleaning data* dilakukan dengan menghapus data yang tidak sesuai atau tidak diperlukan dalam penelitian, seperti menghapus data kosong yang tidak memiliki nilai pada data penelitian dan mengambil data penting yang akan diteliti secara manual. Setelah itu, melakukan transformasi data dengan cara mengubah data dari *categorical* menjadi numerik. Jumlah data yang dihasilkan pada tahap *cleaning data* menjadi 84900 data dan jumlah atribut / variabel yang dihasilkan pada tahap *transformation data* menjadi 85 atribut / variabel.

3. *K-Fold Cross Validation*

Proses selanjutnya yaitu melakukan *k-fold cross validation* dengan membagi data penelitian menjadi data *testing* dan data *training*. Pembagian data meliputi 1/10 data sebagai data *testing*, dan 9/10 data sebagai data *training*. Uji validitas menggunakan *10-fold cross validation* yang membagi data secara acak menjadi 10 iterasi. Kemudian sesuaikan model menggunakan iterasi $k - 1$ (k minus 1) dan validasikan model menggunakan iterasi ke- k yang tersisa, lalu catat skor/kesalahan (*error*). Ulangi proses sampai setiap iterasi k yang berfungsi sebagai set uji dan mengambil rata-rata skor yang

tercatat. Rata-rata skor yang tercatat akan menjadi metrik performa untuk model.

4. Pembuatan model SVM Klasifikasi

Tahap ini, hal yang terlebih dahulu dilakukan yaitu dengan menentukan tipe *kernel* (*linear*, *polynomial* dan *Gaussian*) serta menentukan nilai parameter. Selanjutnya *training* SVM mulai dilakukan pada data untuk model SVM klasifikasi. Jika model terbaik pada data telah terpilih maka sudah bisa untuk dilakukan proses klasifikasi.

5. *Testing*/Pengujian

Tahap terakhir pada penelitian ini adalah dengan melakukan *testing* atau pengujian. Model SVM yang telah didapatkan pada tahapan sebelumnya digunakan untuk menguji data *testing* dari proses pembagian data penelitian. Pengujian yang digunakan menggunakan *confusion matrix* untuk mendapatkan nilai *accuracy*, *sensitivity*, dan *specificity*.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Dari penelitian yang dilakukan, didapatkan hasil dan dapat diambil simpulan sebagai berikut:

- a. Model prediksi untuk klasifikasi dapat diimplementasikan kedalam sebuah sistem yang dapat digunakan dalam penanganan kejadian pada penyakit *diabetes*.
- b. Analisis korelasi fitur pada penelitian ini didapatkan hasil korelasi seluruhnya ada yang bernilai positif, negatif dan nol. Hasil analisis korelasi dengan jumlah terbanyak didapatkan dengan nilai hasil korelasi yaitu 0 yang berarti terdapat variabel-variabel tersebut yang tidak memiliki korelasi antar variabel.
- c. Hasil klasifikasi menggunakan metode *support vector machine* (SVM) didapatkan hasil rata-rata *accuracy* terbesar pada *kernel gaussian* sebesar 82,76%, dengan hasil rata-rata *sensitivity* sebesar 25,18%, dan hasil rata-rata *specificity* sebesar 75,13%. Selanjutnya didapatkan hasil rata-rata *accuracy* terbesar kedua setelah *kernel gaussian* yaitu pada *kernel linear* sebesar 72,48%, dengan hasil rata-rata *sensitivity* sebesar 27,47%, dan hasil rata-rata *specificity* sebesar 76,02%. Untuk hasil rata-rata *accuracy* terkecil didapatkan oleh *kernel polynomial* sebesar 39,56%, dengan hasil rata-rata *sensitivity* sebesar 26,24%, dan hasil rata-rata *specificity* sebesar 75,05%. Adapun untuk hasil rata-rata *sensitivity* dan *specificity* terbesar ada pada *kernel linear*.

5.2 Saran

Adapun saran yang dapat diberikan untuk penelitian selanjutnya mengenai Klasifikasi penderita penyakit *diabetes mellitus* menggunakan metode *support vector machine* (SVM) adalah sebagai berikut:

- a. Mencoba menggunakan metode lain dalam penelitian klasifikasi selanjutnya seperti *neural network*, *logistic regression*, dan lain-lain supaya hasil klasifikasi dapat diketahui sehingga dapat dibandingkan dengan hasil dari metode lainnya.
- b. Mencoba menggunakan data *diabetes* lain untuk melakukan klasifikasi *binary-class* ataupun *multi-class* untuk mengetahui hasil evaluasi kinerja klasifikasi dari metode SVM.

DAFTAR PUSTAKA

- Afrianti, E., Bahri, N., Yuniarti, & Elmatris, S. 2012. EFEK HIDROTERAPI PADA PENURUNAN KADAR GULA DARAH SESAAT (KGDS) TERHADAP PENDERITA DIABETES MELITUS TIPE 2. *Majalah Kedokteran Andalas*, 36.
- Ahmad, A. 2017. Mengenal Artificial Intelligence, Machine Learning, Neural Network dan Deep Learning. *Teknologi Informasi*.
- Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. 2012. application of data mining diabetes health care in young and old patients. 25, 127-136.
- American Diabetes Association. 2010. Position statement: Standards of Medical Care in Diabetes 2010. *Diab Care*.
- Astuti, C. C. 2017. Analisis Korelasi untuk Mengetahui Keeratan Hubungan antara Keaktifan Mahasiswa dengan Hasil Belajar Akhir. *Journal of Information and Computer Technology Education*, 1-7.
- Azis, H., Purnawansyah, Fattah, F., & Putri, I. P. 2020. Performa Klasifikasi K-NN dan Cross-validation Pada Data Pasien Pengidap Penyakit Jantung. *ILKOM Jurnal Ilmiah*, 81-86.
- Bhavsar, H., & Panchal, M. H. 2012. A Review On Support Vector Machine For Data Classification . *International Journal of Advanced Research In Computer Engineering and Technology*, 2278-1328.
- Chang, W. 2021. Web Application Framework for R. *Package 'shiny'*. <https://cran.r-project.org/web/packages/shiny/index.html>
- Cortes, C., & Vapnik, V. 1995. Support-Vector Networks . 273-297.
- Direktur Pengendalian PTM. 2008. *Pedoman Pengendalian Diabetes Mellitus dan Penyakit Metabolik*. Jakarta: Departemen Kesehatan Republik Indonesia.
- Fu, L. 1994. Neural Network In Computer Intelligent.

- Gradiana, R. D., & Irhamah. 2014. Klasifikasi Pasien Penderita Diabetes Mellitus Tipe 2 Menggunakan Metode Analisis Diskriminan Hybrid Algoritma Genetika. *JURNAL SAINS DAN SENI POMITS*, 3.
- Harrel , F. E. 2021. Harrell Miscellaneous. *Package 'Hmisc'*. <https://cran.r-project.org/web/packages/Hmisc/index.html>
- International Diabetes Federation. 2019. Global and Regional Diabetes Prevalence Estimates For 2019 and Projections For 2030 and 2045: Results From The International Diabetes Federation Diabetes Atlas, 9th Edition. *Diabetes Research and Clinical Practice*.
- Kohavi, R., & Provost, F. 1998. Glossary of terms: Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Machine Learning*, 271-274.
- Kuhn, M. 2021. Classification and Regression Training. *Package 'caret'*. <https://cran.r-project.org/web/packages/caret/index.html>
- Kumari, V. A., & Chitra, R. 2013. Classification Of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications*, 3(2).
- Maimon, O., & Rokach, L. 2010. Data mining and knowledge discovery handbook.
- Mamase, S., & Buliali, J. L. 2016. Metode Hibrida K-Means dan Generalized Regression Neural Network Untuk Prediksi Arus Lalu Lintas. *Jurnal Buana Informatika*, 7, 159-168.
- Meyer, D. 2021. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. *Package 'e1071'*. <https://cran.r-project.org/web/packages/e1071/index.html>
- Muttaqin, F. A., & Bachtiar, A. M. 2019. IMPLEMENTASI TEKS MINING PADA APLIKASI PENGAWASAN PENGGUNAAN INTERNET ANAK "DODO KIDS BROWSER". *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*.
- Nugroho, A. S., Witarto, A. B., & Handoko, D. 2003. Support Vector Machine Teori dan Aplikasi dalam Bioinformatik.
- Nurhayati, Busman, & Iswara, R. P. 2019. Pengembangan Algoritma Unsupervised Learning Technique Pada Big Data Analysis Di Media Sosial Sebagai Media Promosi Online Bagi Masyarakat. *Teknik Informatika*, 12.

- Pandie, E. 2012. Implementasi Algoritma Data Mining K Nearest Neighbour (KNN) Dalam Pengambilan Keputusan Pengajuan Kredit.
- PERKENI. 2011. *Konsensus Pengelolaan dan Pencegahan Diabetes Mellitus Tipe 2 Di Indonesia*. Jakarta: Perkeni.
- Prasatya, A., Siregar, R. R., & Arianto, R. 2020. Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes. *Pengkajian dan Penerapan Teknik Informatika*, 13.
- Puspitasari, A. M., Ratnawati, D. E., & Widodo, A. W. 2018. Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 802-810.
- Qi, Z., Tian, Y., & Shi, Y. 2013. Robust Twin Support Vector Machine for Pattern Classification. 305– 316.
- Que, A., Yasa, I., & Lestari, A. 2013. Gambaran Hasil Pemeriksaan Kadar HbA1C pada Penderita Diabetes Mellitus di Laboratorium Rumah Sakit Umum Surya Husada Tahun 2013.
- Rahman, M. F., Darmawidjadja, M. I., & Alamsyah, D. 2017. KLASIFIKASI UNTUK DIAGNOSA DIABETES MENGGUNAKAN METODE BAYESIAN REGULARIZATION NEURAL NETWORK (RBNN). *Jurnal Informatika*, 11.
- Refaeilzadeh, P., Tang, L., & Liu, H. 2008. Cross Validation.
- Santoso, B. 2007. *Data Mining, Teori dan Aplikasi*. Jakarta: Graha Ilmu.
- Sulistria, Y. 2013. Tingkat Self Care Pasien Rawat Jalan Diabetes Mellitus Tipe 2 Di Puskesmas Kalirungkut Surabaya. *Jurnal Ilmiah Mahasiswa Universitas Surabaya*, 2.
- Tim Riskesdas. 2019. *Laporan Nasional RISKESDAS 2018*. Jakarta: Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan (LPB).
- Trevor, H., Robert, T., & Friedman, J. 2008. The Elements Of Statistical Learning Data Mining, Interface, and Prediction.
- Wickham, H. 2020. Read Rectangular Text Data. *Package 'readr'*. <https://cran.r-project.org/web/packages/readr/index.html>
- Witten, I. H., Frank, E., & Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd edition*. Burlington: Elsevier.

- Yunita, F. 2016. SISTEM KLASIFIKASI PENYAKIT DIABETES MELLITUS MENGGUNAKAN METODE K-NEAREST NEIGHBOR (K-NN). *Jurnal BAPPEDA*, 2.
- Yusa, M., Utami, E., & Luthfi, E. T. 2016. Analisis Komparatif Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes. *Buana Informatika*, 7, 293-302.

