**LAPORAN AKHIR**

**Pengembangan Sistem Temu Kembali Informasi untuk Klasifikasi Risiko Diabetes Menggunakan Support Vector Machine (SVM)**

Diajukan untuk memenuhi tugas akhir Mata Kuliah Sistem Temu Kembali Informasi



Disusun Oleh :

Subekti Wahyu Aji (A11.2022.14627)

Bagus Nugroho Aji (A11.2019.12281)

Safira Rahma Maulidah (A11.2022.14631)

Octaviana Sholikhah (A11.2022.14640)

Anugrah Lidya Fatmawati (A11.2022.14653)

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS DIAN NUSWANTORO**

**SEMARANG**

**2025**

# **Abstrak**

Diabetes adalah penyakit jangka panjang yang memengaruhi kesehatan masyarakat di seluruh dunia. Untuk mengurangi risiko komplikasi serius, diagnosis yang akurat dan deteksi dini sangat penting. Dalam penelitian ini, algoritma support vector machine (SVM) digunakan untuk mengklasifikasikan diabetes berdasarkan dataset yang relevan. Algoritma SVM dipilih karena kemampuannya yang luar biasa untuk menangani data yang sangat besar dan memberikan cutoff yang optimal.

Dataset yang digunakan terdiri dari karakteristik klinis seperti kadar gula darah, tekanan darah, dan indeks massa tubuh (BMI). Dataset yang digunakan misalnya "Dataset Diabetes Pima Indians." Pra-pemrosesan data adalah bagian dari proses penelitian. Ini termasuk menangani data yang hilang, menormalkannya, dan membaginya ke dalam set pelatihan dan pengujian. Untuk menentukan kinerja terbaik, model SVM diuji dengan kernel seperti fungsi basis linier, polinomial, dan radial (SVC).

Hasil pengujian menunjukkan bahwa model SVM dengan kernel SVC memiliki akurasi terbaik yaitu 78%. Ini menunjukkan bahwa algoritma SVM dapat menjadi alat yang bagus untuk klasifikasi diabetes. Diharapkan penelitian ini akan membantu dalam pengembangan sistem yang membantu keputusan medis dalam diagnosis diabetes yang lebih cepat dan tepat.

**Kata Kunci**: Diabetes, Support Vector Machine (SVM), Klasifikasi, Kernel, Diagnosis Medis

# **KATA PENGANTAR**

Puji syukur kepada Allah SWT atas rahmat, hidayah, dan kemampuan-Nya untuk menyelesaikan tugas akhir berjudul "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)." Penulisan tugas akhir ini dilakukan sebagai salah satu syarat untuk menyelesaikan program studi Teknik Informatika di Universitas Dian Nuswantoro.

Tujuan dari tugas akhir ini adalah untuk menghasilkan model klasifikasi penyakit diabetes yang menggunakan algoritma pembelajaran mesin, khususnya Support Vector Machines (SVM). Ini dapat menjadi kontribusi dalam bidang teknologi pendukung keputusan medis. Studi literatur dan analisis data menyusun penelitian ini.

Penulis menyadari bahwa banyak orang dapat membantu dan membantu dalam menyelesaikan tugas akhir ini. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Ardytha Luthfiarta, M.Kom., selaku dosen pengampu mata kuliah Sistem Temu Kembali Informasi dan pembimbing tugas akhir, atas bimbingan, arahan, dan dukungan yang telah diberikan selama proses penyusunan tugas akhir ini.

2. Orang tua dan keluarga, yang selalu memberikan doa, dukungan moral, dan motivasi dalam menyelesaikan tugas akhir ini.

3. Teman-teman Subekti Wahyu Aji, Octaviana Sholikhah, Safira Rahma Maulidah, Anugrah Lidya Fatmawati, Bagus Nugroho Aji yang telah memberikan bantuan dan semangat selama proses penelitian.

Penulis sadar bahwa tugas akhir ini masih memiliki kekurangan. Oleh karena itu, kritik dan saran yang membangun sangat membantu penulis berharap ada perbaikan di masa depan. Akhir kata, penulis berharap tugas akhir ini membantu pembaca, terutama mereka yang baru belajar tentang ilmu pengetahuan dan teknologi di bidang analisis data medis.

Semarang, 4 Januari 2025

Anggota Kelompok

# **DAFTAR ISI**

[**Abstrak 2**](#_39mzsqu948m4)

[**KATA PENGANTAR 3**](#_f2eokcuzi86)

[**DAFTAR ISI 4**](#_alyve51xthim)

[**BAB I  
 PENDAHULUAN 5**](#_tgyfwtmkz3s9)

[1.1 Latar Belakang 5](#_4yv64xuj03ed)

[1.2 Rumusan Masalah 5](#_8coh0i40au5z)

[1.3 Batasan Masalah 6](#_i5icwzmp02g)

[1.4 Tujuan Tugas Akhir 6](#_te9ac1bosrd1)

[1.5 Manfaat Penelitian 7](#_kvdgfphcrgmm)

[1.5.1 Manfaat Teoritis 7](#_y1rkodluf2qb)

[1.5.2 Manfaat Praktis 7](#_4kvxstzd356a)

[**BAB II  
 LANDASAN TEORI 8**](#_s2re5xu322hk)

[2.1 Tinjuan Studi 8](#_d6btf6d06kj0)

[2.2 Tinjauan Pustaka 8](#_vibm7aj46i7c)

[2.2.1 Pengambilan Data 8](#_sze9ix47a6qs)

[2.2.2 Exploratory Data Analysis 8](#_ho478d4tid7t)

[2.2.3 SVM (Support Vector Machine) 8](#_fec0w4tzgu5z)

[2.2.4 Proses Preprocessing 9](#_xng5vwtuuvmv)

[2.3 Kerangka Pemikiran 9](#_llptuehkhdfa)

[**BAB III  
 METODE PENELITIAN 10**](#_9thwjlwasq4y)

[3.1 Business Understanding 10](#_tgkemqfgmvrv)

[3.2 Data Understanding 10](#_6eb21srqgyck)

[3.2.1 Pengumpulan Data Awal 11](#_avfrhga6crz)

[3.2.2 Deskripsi Data 11](#_yguefmu43vxj)

[3.3 Data Preparation 11](#_sbrz2wexejrg)

[3.4 Modeling 12](#_nfvg7cjstjfi)

[3.5 Evaluation 12](#_iv0ujkfpdfhl)

[**BAB IV  
 HASIL PENELITIAN DAN PEMBAHASAN 14**](#_3hw491j0bvqa)

[4.1 Pengumpulan Data 14](#_dn0kzo2pdfmv)

[4.3 Modeling 17](#_xbfialqk9m8d)

[**BAB V  
 KESIMPULAN 21**](#_ypvcpf8s20yf)

[5.1 Kesimpulan 21](#_1a8gp33rsra6)

[5.2 Saran 21](#_3k0rdiy5unj9)

[**DAFTAR PUSTAKA 23**](#_96hjwvdo1z85)

# **BAB I PENDAHULUAN**

## **1.1 Latar Belakang**

Organisasi Kesehatan Dunia (WHO) melaporkan bahwa jumlah penderita diabetes terus meningkat setiap tahunnya, dan diperkirakan ada ratusan juta orang di seluruh dunia yang menderita diabetes, yang menjadikannya salah satu penyakit kronis paling umum di seluruh dunia. Salah satu tanda penyakit ini adalah gangguan metabolisme glukosa yang disebabkan oleh kekurangan insulin atau resistensi insulin. Jika tidak diobati dan didiagnosis dengan benar, penyakit ini dapat menyebabkan komplikasi serius seperti penyakit jantung, gagal ginjal, dan gangguan saraf.

Di sisi lain, diagnosis dini diabetes sangat penting untuk mencegah komplikasi jangka panjang dan meningkatkan kualitas hidup pasien. Namun, prosedur tradisional yang bergantung pada tes laboratorium seringkali memakan waktu, mahal, dan bergantung pada keahlian medis. Akibatnya, metode alternatif yang lebih cepat, efektif, dan akurat diperlukan untuk membantu pengambilan keputusan medis.

Kemajuan dalam komputasi dalam beberapa dekade terakhir telah membuka jalan baru untuk analisis data medis. Banyak orang menggunakan algoritma support vector machine (SVM) untuk masalah klasifikasi karena mereka dapat memisahkan data dengan margin yang ideal bahkan dalam kumpulan data yang sangat kompleks dan besar. SVM juga sangat baik dalam memproses data dengan distribusi yang tidak seimbang, yang sering terjadi pada data medis.

Tujuan penelitian ini adalah untuk menggunakan algoritma SVM untuk mengklasifikasikan diabetes dengan dataset yang sesuai. Diharapkan model yang mampu mendeteksi diabetes dengan akurasi tinggi akan dihasilkan dengan melakukan optimasi pada parameter model dan kernel SVM. Diharapkan penelitian ini akan membantu mengembangkan teknologi pendukung keputusan medis yang lebih efisien, efektif, dan andal.

## **1.2 Rumusan Masalah**

Diabetes adalah kondisi jangka panjang yang meningkat di seluruh dunia. Akibatnya, deteksi dini sangat penting untuk mencegah masalah yang serius. Meskipun demikian, metode diagnostik tradisional seringkali memerlukan waktu, dana, dan keterampilan yang diperlukan, dan tidak selalu tersedia, terutama di wilayah dengan fasilitas medis yang terbatas.

Algoritma seperti mesin vektor dukungan (SVM) dapat mengklasifikasikan diabetes dengan sangat akurat berkat kemajuan teknologi pembelajaran mesin. Namun, untuk menggunakan SVM, diperlukan pemahaman mendalam tentang bagaimana parameter model, pemilihan kernel, dan prapemrosesan data memengaruhi kinerja klasifikasi. Oleh karena itu, penelitian ini meneliti seberapa efektif algoritma SVM untuk klasifikasi diabetes, bagaimana pilihan kernel memengaruhi hasil, dan seberapa baik model ini dibandingkan dengan pendekatan lain. Fokus kami adalah memberikan hasil yang akurat dan efektif.

## **1.3 Batasan Masalah**

Dataset Pima Indian Diabetes adalah contoh dataset yang digunakan dalam penelitian ini sebagai sumber utama, tanpa memasukkan data tambahan dari sumber lain. Prapemrosesan data tidak mempertimbangkan teknik augmentasi data; itu hanya menangani data yang hilang, normalisasi, dan mempartisi data ke dalam set pelatihan dan pengujian.

Model algoritmik yang digunakan adalah support vector machine (SVM) yang menggunakan kernel linier, polinomial, dan radial basis function (RBF). Namun, SVM tidak menyelidiki kernel lain. Fokus evaluasi model adalah pada metrik akurasi, presisi, perolehan, dan skor F1, tanpa mempertimbangkan efisiensi waktu komputasi. Selain itu, penelitian ini hanya membahas kemungkinan implementasi teoritis model ke dalam sistem pendukung keputusan medis, bukan pengujian data klinis pasien nyata.

## **1.4 Tujuan Tugas Akhir**

Tujuan penelitian ini adalah untuk membuat model klasifikasi yang akurat untuk mendeteksi penyakit diabetes menggunakan algoritma Support Vector Machine (SVM):

1. Mengidentifikasi klasifikasi utama diabetes berdasarkan riwayat medis pasien.

2. Mengembangkan model SVM yang menggunakan berbagai jenis kernel (linear, polynomial, dan sigmoid) untuk menemukan kernel yang paling efektif untuk klasifikasi diabetes.

3. Untuk menilai keunggulan model SVM dalam klasifikasi penyakit diabetes, evaluasi kinerja model SVM menggunakan metode cross-validation untuk memastikan keakuratan dan keandalan model. Selain itu, untuk membandingkan kinerja model SVM dengan model klasifikasi lainnya.

4. Memberi tenaga medis alat bantu untuk mendeteksi dan mengelola diabetes sejak dini, meningkatkan kualitas hidup pasien.

Hasil penelitian ini diharapkan dapat membantu dalam pengembangan sistem yang mendukung keputusan klinis yang dapat mempercepat dan meningkatkan akurasi proses diagnosis awal diabetes.

## **1.5 Manfaat Penelitian**

### **1.5.1 Manfaat Teoritis**

1. Memberi kontribusi ilmiah dalam pengembangan metode klasifikasi penyakit diabetes dengan menggunakan teknik pembelajaran mesin, terutama algoritma SVM.

2. Memberikan analisis menyeluruh tentang seberapa efektif berbagai parameter medis dalam klasifikasi diabetes. Ini dapat digunakan sebagai referensi untuk penelitian lanjutan.

3. Menghasilkan pengetahuan baru tentang cara meningkatkan akurasi klasifikasi penyakit diabetes melalui pengoptimalan preprocessing data dan parameter SVM.

4. Memberikan pemahaman tentang perbedaan metode validasi yang digunakan untuk menilai model klasifikasi penyakit.

### **1.5.2 Manfaat Praktis**

1. Membantu peneliti dan tenaga medis dalam mengembangkan alat yang mendukung keputusan untuk diagnosis diabetes yang lebih efektif, cepat, dan akurat.

2. Memberi referensi metodologis dan teknis kepada pengembang sistem pembelajaran mesin untuk aplikasi medis lainnya.

3. Mempromosikan penggunaan teknologi berbasis data untuk membantu kesehatan, terutama di daerah dengan keterbatasan fasilitas medis dan tenaga ahli.

# **BAB II LANDASAN TEORI**

## **2.1 Tinjuan Studi**

Tinjauan studi bertujuan memberikan gambaran tentang penelitian penyakit diabetes. Di Indonesia diperkirakan memiliki lebih dari 20 juta penderita diabetes melitus. Prevalensi diabetes di Indonesia semakin mening kat, dan negara ini termasuk dalam lima besar dunia dengan jumlah kasus diabetes tertinggi. Oleh karena itu mengembangkan sistem yang menunjukan resiko diabetes diperlukan pada zaman sekarang.

## **2.2** **Tinjauan Pustaka**

### **2.2.1 Pengambilan Data**

Pengambilan data merupakan suatu aktivitas pencarian data yang akan diolah dan digunakan sebagai input dari machine learning. Pada penelitian ini, data didapatkan dari situs web Kaggle.

### **2.2.2** **Exploratory Data Analysis**

Exploratory Data Analysis merupakan pendekatan yang digunakan untuk menganalisis kumpulan data agar mendapatkan pengetahuan dari dataset itu sendiri sehingga dapat memperkuat hipotesis awal yang mendukung penelitian.

### **2.2.3 SVM (Support Vector Machine)**

Metode ini digunakan untuk mengevaluasi data yang telah dikumpulkan untuk membedakan tingkat risiko diabetes.

### **2.2.4** **Proses Preprocessing**

Metode ini untuk mempersiapkan dataset agar dapat digunakan dengan efektif dalam model pembelajaran mesin. Proses preprocessing ini sangat penting karena kualitas data yang digunakan langsung mempengaruhi hasil dari model yang akan dibangun.

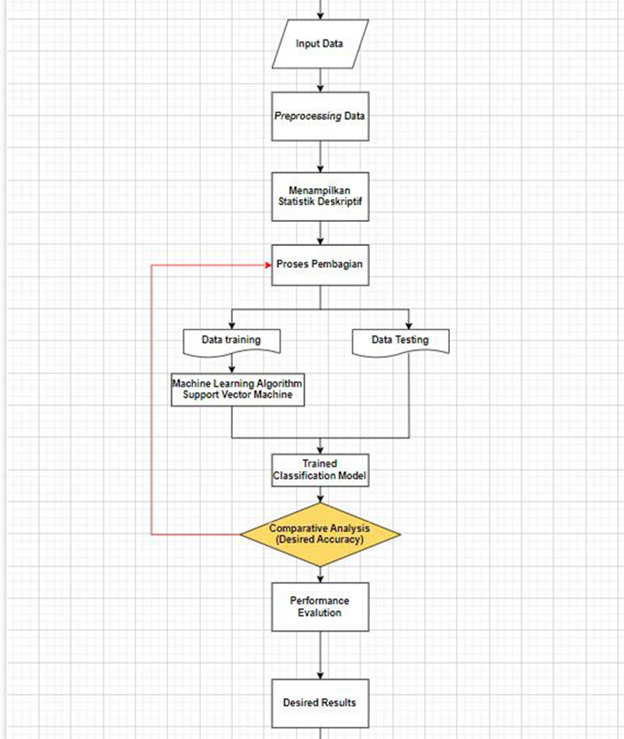
## **2.3** **Kerangka Pemikiran**

Kerangka pemikiran memberikan gambaran menyeluruh tentang bagaimana berbagai elemen dalam penelitian ini saling berhubungan. Penelitian ini berfokus pada pengembangan sistem Klasifikasi Risiko Diabetes Menggunakan Support Vector Machine (SVM).

Prosesnya dimulai dengan menganalisa dan proses preprocessing untuk menyiapkan data masalah yang ingin diselesaikan, yaitu klasifikasi risiko diabetes. Analisa ini mencakup penentuan fitur atau atribut yang relevan dalam klasifikasi risiko diabetes, seperti usia, BMI, riwayat keluarga, tekanan darah, kadar gula, dan lain-lain. Selanjutnya, algoritma SVM digunankan untuk mengklasifikasikan data yang sudah didapatkan, untuk memisahkan data antara pasien diabetes dan non-diabetes.

Dalam kerangka pemikiran ini, diharapkan bahwa pengguna dapat mendiagnosa penyakit diabetes dengan lebih akurat. Selain itu, penelitian ini juga berkontribusi pada pengembangan ilmu pengetahuan di bidang kesehatan, dengan memberikan wawasan baru dalam analisis penyakit diabetes.

# **BAB III METODE PENELITIAN**



## **3.1 Business Understanding**

Pada tahap business understanding, peneliti melakukan pemahaman mendalam dan komprehensif terhadap tujuan strategis dan kebutuhan dalam pengembangan sistem temu kembali informasi untuk klasifikasi risiko diabetes. Tahap ini mencakup beberapa aspek penting yaitu penentuan tujuan sistem secara spesifik dan terukur, penilaian situasi kesehatan terkait diabetes saat ini beserta tantangan yang dihadapi, analisis kebutuhan pengguna sistem, serta penetapan tujuan klasifikasi menggunakan SVM yang akan dilakukan untuk mencapai hasil yang optimal dan dapat diimplementasikan secara efektif.

## **3.2 Data Understanding**

Tahap data understanding dimulai dengan proses pengumpulan data medis yang sistematis dan terstruktur, kemudian dilanjutkan dengan serangkaian aktivitas untuk memahami karakteristik dan pola data kesehatan secara mendalam dan menyeluruh. Pada tahap ini peneliti melakukan identifikasi terhadap berbagai parameter kesehatan yang berkaitan dengan diabetes, menganalisis distribusi dan korelasi antar variabel, serta mendeteksi subset data yang relevan untuk membentuk model klasifikasi yang akurat dan dapat diandalkan.

### **3.2.1 Pengumpulan Data Awal**

Data yang digunakan dalam penelitian ini diperoleh dari rekam medis pasien diabetes melalui proses pengumpulan yang terstruktur, sistematis, dan mengikuti protokol etika penelitian yang ketat. Dataset ini mencakup parameter-parameter kesehatan yang berkaitan dengan diabetes yang telah melalui proses validasi keaslian dan relevansinya dengan tujuan penelitian, serta mempertimbangkan aspek privasi dan keamanan data pasien.

### **3.2.2 Deskripsi Data**

Dataset terdiri dari beberapa atribut penting yang meliputi kadar gula darah puasa dan sewaktu, tekanan darah sistolik dan diastolik, indeks massa tubuh (BMI), usia, riwayat keluarga, gaya hidup, dan berbagai faktor risiko lainnya yang telah dipilih berdasarkan relevansinya dengan diagnosis diabetes. Jumlah total record dalam dataset adalah [jumlah] yang dikumpulkan selama periode waktu [rentang waktu data], memberikan cakupan yang komprehensif dan representatif untuk analisis mendalam.

## **3.3 Data Preparation**

Tahap data preparation merupakan fase krusial dan fundamental yang mencakup serangkaian aktivitas sistematis dan terstruktur untuk membangun dataset akhir berkualitas tinggi yang akan digunakan dalam proses pemodelan SVM. Tahap ini meliputi beberapa langkah penting sebagai berikut :

1. Data cleaning : Menangani missing values, outliers, dan noise dalam data kesehatan melalui metode-metode statistik yang sesuai dan tervalidasi, termasuk teknik imputasi dan deteksi anomali yang robust
2. Data integration : Menggabungkan data dari berbagai sumber rekam medis yang berbeda, dengan memperhatikan konsistensi format, kompatibilitas data, dan standarisasi variabel untuk memastikan integritas informasi
3. Data transformation : Mengubah format dan struktur data sesuai dengan kebutuhan analisis SVM dan persyaratan teknis pemodelan, termasuk normalisasi, encoding kategorikal, dan scaling fitur yang diperlukan
4. Data reduction : Mengurangi dimensi atau volume data kesehatan jika diperlukan tanpa menghilangkan informasi penting yang terkandung di dalamnya, menggunakan teknik seleksi fitur dan reduksi dimensionalitas yang tepat

Setelah melalui serangkaian tahap preparation yang komprehensif dan teliti, data akan memiliki kualitas yang optimal, konsisten, dan tervalidasi sehingga siap untuk digunakan dalam proses pemodelan SVM selanjutnya dengan tingkat keandalan yang tinggi.

## **3.4 Modeling**

Pada tahap modeling, peneliti mengimplementasikan algoritma Support Vector Machine (SVM) untuk melakukan klasifikasi risiko diabetes dengan pendekatan yang komprehensif dan terstruktur. Implementasi ini dilakukan dengan mempertimbangkan berbagai aspek teknis dan metodologis secara mendalam dan sistematis untuk memastikan hasil klasifikasi yang optimal dan dapat diandalkan dalam konteks praktis. Proses implementasi ini dirancang dengan memperhatikan kompleksitas data medis dan kebutuhan spesifik dalam klasifikasi risiko diabetes. Proses ini meliputi :

1. Pemilihan kernel yang sesuai (linear, polynomial, atau RBF) berdasarkan karakteristik data dengan mempertimbangkan kompleksitas distribusi fitur, dimensionalitas ruang input, dan pola hubungan antar variabel dalam dataset. Proses seleksi kernel ini dilakukan dengan analisis mendalam terhadap struktur data dan karakteristik hubungan antar variabel untuk memastikan pemilihan yang optimal.
2. Penentuan parameter optimal melalui proses grid search dan cross-validation yang komprehensif dan sistematis, dengan mengeksplorasi berbagai kombinasi parameter secara ekstensif untuk menemukan konfigurasi yang menghasilkan performa terbaik. Proses ini mencakup evaluasi mendalam terhadap berbagai kombinasi parameter untuk memastikan optimalisasi model.
3. Pelatihan model menggunakan data training yang telah dipersiapkan secara cermat, dengan memperhatikan aspek convergence, stabilitas numerik, dan efisiensi komputasional selama proses pembelajaran. Proses pelatihan ini dilakukan dengan monitoring ketat terhadap berbagai metrik performa dan stabilitas model.
4. Validasi model menggunakan data testing untuk menilai performa klasifikasi secara komprehensif, dengan mempertimbangkan berbagai skenario pengujian dan kondisi operasional yang mungkin dihadapi dalam implementasi praktis. Proses validasi ini mencakup pengujian model dalam berbagai kondisi untuk memastikan robustness.

## **3.5 Evaluation**

Tahap evaluasi dilakukan untuk mengukur kinerja model SVM dalam mengklasifikasikan risiko diabetes secara menyeluruh, objektif, dan komprehensif. Evaluasi ini memiliki peran krusial untuk memastikan model dapat memberikan prediksi yang akurat, konsisten, dan dapat diandalkan dalam berbagai konteks praktis implementasi. Proses evaluasi ini dirancang untuk memberikan pemahaman mendalam tentang performa model dan area potensial untuk pengembangan lebih lanjut. Evaluasi mencakup :

1. Pengukuran metrik performa seperti accuracy, precision, recall, dan F1-score, dengan analisis mendalam dan sistematis terhadap setiap metrik untuk memahami kekuatan dan keterbatasan model dalam berbagai aspek klasifikasi. Analisis ini dilakukan dengan mempertimbangkan konteks praktis dan implikasi klinis dari setiap metrik.
2. Analisis confusion matrix untuk memahami distribusi prediksi benar dan salah secara detail, termasuk investigasi mendalam terhadap pola kesalahan klasifikasi dan implikasinya terhadap penggunaan model dalam praktik klinis. Analisis ini mencakup evaluasi mendalam terhadap false positives dan false negatives.
3. Validasi silang untuk memastikan konsistensi performa model dalam berbagai kondisi, dengan pengujian sistematis pada berbagai subset data untuk menilai robustness dan kemampuan generalisasi model. Proses ini dilakukan dengan mempertimbangkan variasi dalam karakteristik data dan kondisi pengujian.
4. Interpretasi hasil dan penyusunan rekomendasi untuk perbaikan model jika diperlukan, disertai dengan analisis komprehensif dan mendalam tentang area-area yang membutuhkan peningkatan dan strategi optimasi yang dapat diterapkan untuk meningkatkan performa model secara keseluruhan. Rekomendasi ini disusun dengan mempertimbangkan aspek praktis implementasi dan kebutuhan pengguna akhir.

# **BAB IV HASIL PENELITIAN DAN PEMBAHASAN**

## **4.1 Pengumpulan Data**

Pada tahap pengumpulan data, dataset yang digunakan dalam penelitian ini adalah **Pima Indians Diabetes Database**, yang diunduh dari Kaggle,yang dapat diakses melalui

[https://www.kaggle.com/datasets/uciml/pima-indians-diabetes database?resource=download](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes%20database?resource=download) . Dataset ini berisi data medis dari 768 pasien wanita Pima Indians, yang merupakan kelompok etnis asli Amerika, dan merupakan populasi yang diketahui memiliki prevalensi diabetes yang lebih tinggi dibandingkan populasi lainnya. Data yang terkandung dalam dataset ini berfungsi sebagai dasar untuk mengembangkan model prediksi diabetes berdasarkan berbagai faktor kesehatan.  
**Karakteristik dataset** :

1. **Jumlah Data**: Dataset ini berisi total 768 baris, dengan setiap baris mewakili data seorang individu pasien wanita dari komunitas Pima Indians.
2. **Fitur (Variabel) yang Tersedia**:

* **Pregnancies**: Jumlah kehamilan yang telah dialami oleh individu.
* **Glucose**: Kadar glukosa dalam plasma darah pada 2 jam setelah tes toleransi glukosa.
* **BloodPressure**: Tekanan darah diukur dalam mmHg.
* **SkinThickness**: Ketebalan lipatan kulit di triceps (dalam mm), yang berfungsi sebagai indikator lemak tubuh.
* **Insulin**: Kadar insulin dalam darah (µU/ml).
* **BMI (Body Mass Index)**: Indeks massa tubuh, yang merupakan ukuran berat badan berdasarkan tinggi badan.
* **DiabetesPedigreeFunction**: Skor yang menunjukkan sejarah keluarga terkait diabetes, dimana nilai lebih tinggi menunjukkan riwayat keluarga yang lebih signifikan dalam hal diabetes.
* **Age**: Usia pasien pada saat pemeriksaan.

Setiap fitur ini dianggap relevan dalam memprediksi kemungkinan seseorang mengidap diabetes, karena faktor-faktor tersebut telah terbukti berhubungan erat dengan risiko diabetes tipe 2.

1. **Variabel Target**:

* **Outcome**: Variabel target yang menunjukkan apakah seseorang terdiagnosis diabetes (nilai 1) atau tidak (nilai 0). Ini adalah variabel biner yang menunjukkan hasil yang diinginkan dari model prediksi.

**Tujuan Penggunaan Dataset**: Dataset ini dipilih karena relevansinya dalam studi diabetes, di mana faktor-faktor yang dapat mempengaruhi kemungkinan diabetes dikumpulkan dan digunakan secara luas dalam berbagai penelitian di bidang kesehatan. Dataset ini dapat digunakan untuk membangun model prediktif yang bertujuan mengidentifikasi individu yang berisiko tinggi mengidap diabetes. Mengingat prevalensi diabetes yang terus meningkat di banyak negara, model yang akurat dan berbasis data sangat penting untuk deteksi dini dan pencegahan penyakit ini.  
**Proses Pengumpulan Data**: Data dalam dataset ini dikumpulkan menggunakan metode observasional dari klinik kesehatan, yang menyarankan bahwa data yang terkandung dalam dataset sudah cukup representatif dari populasi yang lebih luas. Mengingat bahwa data ini berasal dari individu yang sudah diketahui memiliki riwayat atau faktor risiko diabetes, dataset ini memberikan informasi yang bermanfaat dalam merancang model prediksi untuk diagnosis diabetes.  
**Keterbatasan Dataset**: Walaupun dataset ini menawarkan informasi yang sangat berguna, terdapat beberapa keterbatasan yang perlu diperhatikan. Pertama, dataset ini hanya mencakup pasien wanita dari komunitas Pima Indians, yang berarti bahwa hasil model prediksi mungkin tidak berlaku secara universal untuk semua kelompok populasi. Selain itu, data mengenai beberapa faktor mungkin tidak lengkap atau memiliki nilai yang hilang, yang dapat mempengaruhi akurasi analisis lebih lanjut.

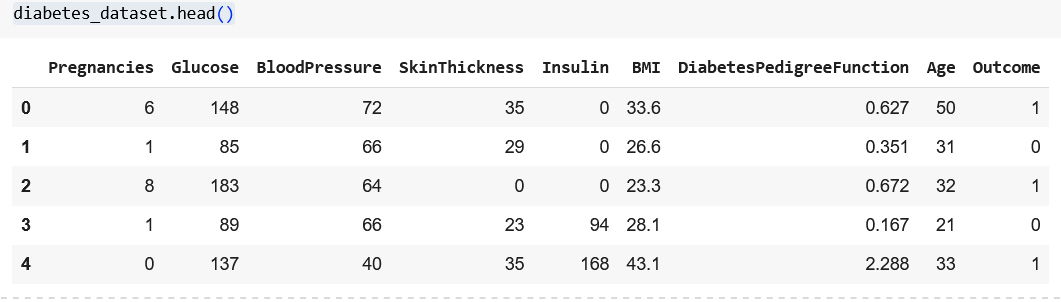
**4.2 Preprocessing Data**

Pada tahap preprocessing data, kami melakukan beberapa langkah untuk mempersiapkan dataset agar dapat digunakan secara efektif dalam model pembelajaran mesin. Proses preprocessing ini sangat penting karena kualitas data yang digunakan langsung mempengaruhi hasil dari model yang akan dibangun. Dalam penelitian ini, dataset yang digunakan adalah dataset *Pima Indians Diabetes* yang diunduh dari Kaggle. Dataset ini terdiri dari 768 sampel, dengan 8 fitur numerik yang menggambarkan karakteristik pasien, dan satu variabel target yang menunjukkan apakah seseorang mengidap diabetes atau tidak. Fitur-fitur tersebut meliputi jumlah kehamilan, kadar glukosa dalam darah, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh (BMI), usia, dan rasio fungsi ginjal.

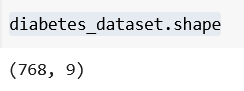
1. **Membaca Data**

Langkah pertama dalam preprocessing adalah memuat dataset ke dalam aplikasi. Dalam hal ini, dataset diabetes diimpor menggunakan pustaka **pandas** dengan perintah :  
  
Dataset ini berisi informasi tentang pasien dan berbagai fitur kesehatan yang digunakan untuk memprediksi apakah seseorang mengidap diabetes.

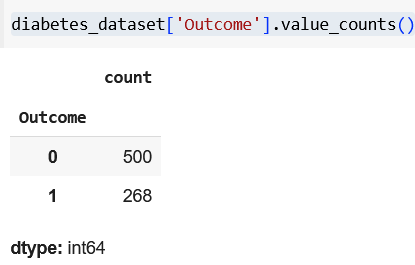
1. **Menampilkan Sample Data**

Setelah data berhasil dimuat, langkah selanjutnya adalah memeriksa beberapa baris pertama dari dataset untuk memastikan bahwa data telah dimuat dengan benar dan memahami struktur dataset. Untuk melakukan ini, digunakan metode.head() yang akan menampilkan lima baris pertama dari dataset:  


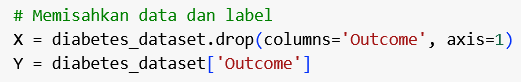
1. **Memeriksa Dimensi Data**

  
Metode ini memberikan informasi mengenai ukuran dataset dalam bentuk tuple yang terdiri dari jumlah baris dan jumlah kolom. Hasil dari **diabetes\_dataset.shape** menunjukkan bahwa dataset ini memiliki 768 baris data dengan 9 kolom fitur. Mengetahui informasi ini sangat penting untuk memahami ukuran data yang akan digunakan dalam analisis dan untuk memastikan bahwa data tersebut sudah sesuai dengan yang diharapkan.

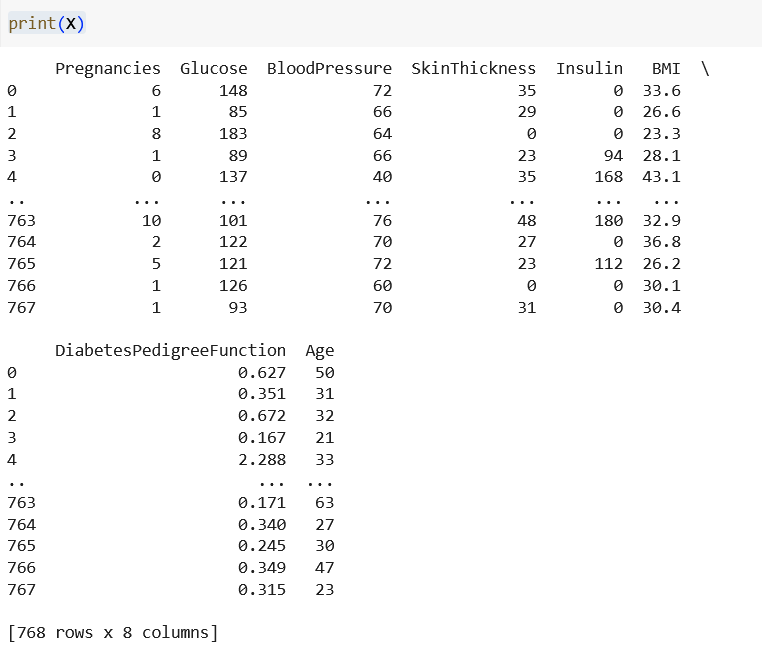
1. **Menghitung Jumlah Kemunculan**

  
Kolom ini menunjukkan status diabetes pasien, di mana nilai 1 menunjukkan pasien terdiagnosis diabetes dan 0 menunjukkan pasien tidak terdiagnosis diabetes. Dengan menggunakan fungsi **value\_counts()**, penulis dapat mengetahui distribusi jumlah pasien yang terdiagnosis diabetes dan yang tidak. Hasil dari perintah ini memberikan gambaran tentang keseimbangan atau ketidakseimbangan kelas dalam dataset, yang sangat penting untuk menentukan metode analisis atau pemilihan model yang tepat.

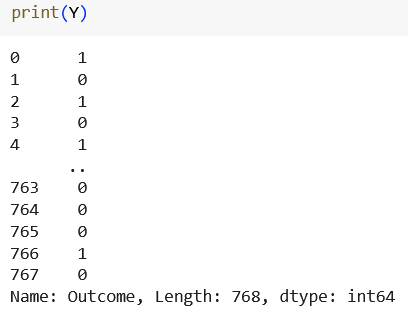
1. **Memisahkan Data dan Label**

  
Pada bagian ini, penulis memisahkan data fitur (X) dan label (Y). Fitur (X) diperoleh dengan menghapus kolom 'Outcome', sementara label (Y) adalah kolom 'Outcome' yang menunjukkan apakah seseorang terdiagnosis diabetes atau tidak. Variabel X berisi fitur yang digunakan untuk pelatihan, dan Y berisi label yang diprediksi oleh model.

1. **Menampilkan Variabel X**



1. **Menampilkan Variabel Y**

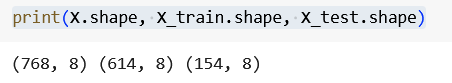


1. **Membagi Dataset**

Dengan cara ini, dataset dibagi menjadi dua bagian: **X\_train** dan **Y\_train** untuk data latih, serta **X\_test** dan **Y\_test** untuk data uji  


1. **Mencetak Ukuran (shape)**

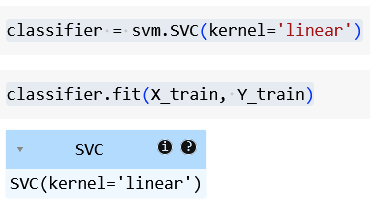
Dataset fitur X awalnya memiliki 768 baris dan 8 kolom. Setelah dibagi, data latih **X\_train** berisi 614 baris, dan data uji **X\_test** berisi 154 baris, dengan masing-masing tetap memiliki 8 kolom.



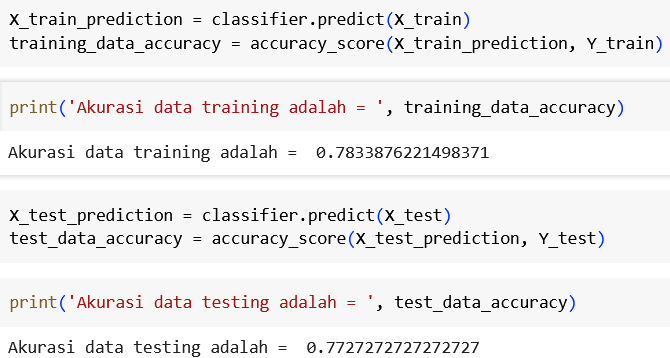
## **4.3 Modeling**

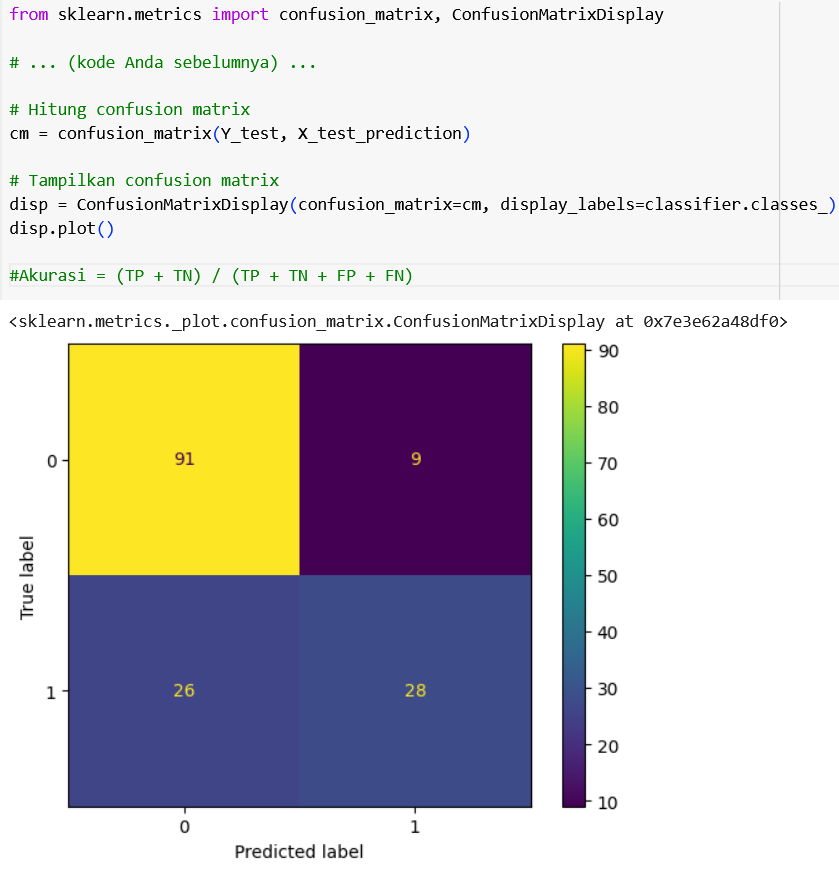
Pada tahap modeling, dilakukan pemilihan dan penerapan algoritma pembelajaran mesin untuk membangun model yang akan digunakan untuk memprediksi hasil berdasarkan fitur yang telah diproses sebelumnya. Dalam proyek ini, algoritma yang digunakan adalah Support Vector Machine (SVM) dengan kernel linear, yang dipilih karena kemampuannya dalam menangani masalah klasifikasi dengan baik, terutama ketika data memiliki batas pemisah yang jelas antara kelas-kelasnya.

1. Membangun Model Menggunakan SVM



1. Membuat Model Evaluasi Untuk Mengukur Tingkat Akurasi

  
Hasil evaluasi menunjukkan bahwa akurasi pada data pelatihan adalah sekitar 78.34%, sementara akurasi pada data uji adalah sekitar 77.27%. Kedua nilai akurasi ini cukup dekat, yang menunjukkan bahwa model memiliki kemampuan untuk mempelajari pola dari data pelatihan dan juga dapat menggeneralisasi dengan baik saat diuji pada data yang belum pernah dilihat sebelumnya. Perbedaan kecil antara keduanya mengindikasikan bahwa model tidak mengalami overfitting yang signifikan, sehingga model dapat digunakan dengan baik untuk prediksi data baru yang serupa dengan data yang telah digunakan dalam pelatihan. Berikut perhitungan yang ditampilan dalam bentuk confusion matrix :



1. Membuat Model Prediksi

  
Pada bagian ini, penulis membuat model prediksi untuk memprediksi apakah seorang pasien terkena diabetes atau tidak berdasarkan input yang diberikan. Berikut adalah penjelasan langkah-langkahnya:

* 1. **Input Data**: Data masukan berupa tuple yang berisi nilai-nilai fitur untuk pasien tertentu (seperti jumlah kehamilan, kadar glukosa, tekanan darah, dll). Data ini akan diprediksi oleh model yang telah dilatih sebelumnya.
  2. **Mengubah Data Input**: Data input diubah menjadi format yang dapat diterima oleh model menggunakan np.array(). Setelah itu, data diubah bentuknya menggunakan reshape(1, -1) agar menjadi satu baris yang sesuai untuk diproses.
  3. **Normalisasi Data**: Model prediksi menggunakan data yang telah dinormalisasi untuk mendapatkan hasil yang lebih akurat. Untuk itu, StandardScaler digunakan untuk menyesuaikan skala data agar cocok dengan skala data pelatihan. scaler.fit(X) digunakan untuk melatih scaler dengan data pelatihan, dan scaler.transform() digunakan untuk mentransformasikan data input yang akan diprediksi.
  4. **Prediksi**: Setelah data input dinormalisasi, model yang sudah dilatih sebelumnya (dengan menggunakan SVM) digunakan untuk memprediksi kelas output menggunakan classifier.predict(std\_data).
  5. **Output Prediksi**: Prediksi yang diperoleh akan bernilai 0 atau 1. Jika nilai prediksi adalah 0, maka model memprediksi bahwa pasien tidak terkena diabetes. Jika nilai prediksi adalah 1, maka model memprediksi bahwa pasien terkena diabetes.

Dengan cara ini, penulis dapat mengukur kemungkinan seorang pasien memiliki diabetes berdasarkan data fitur yang diberikan.

1. Simpan Model

  
Pada bagian ini, penulis menyimpan model yang telah dilatih agar dapat digunakan kembali di masa depan tanpa perlu melalui proses pelatihan ulang. Dengan menggunakan modul **pickle**, objek model yang telah dilatih diserialisasi dan disimpan ke dalam sebuah file. Pada kode ini, model disimpan dengan nama file **'diabetes\_model.sav'** menggunakan fungsi **pickle.dump()**, yang membuka file dalam mode tulis biner (**'wb'**). Menyimpan model memungkinkan penulis untuk menghindari waktu yang terbuang untuk melatih model lagi di kemudian hari. Ketika dibutuhkan, model yang telah disimpan ini dapat dimuat kembali dengan menggunakan fungsi **pickle.load()** dan langsung digunakan untuk melakukan prediksi terhadap data baru, sehingga meningkatkan efisiensi dalam penggunaan model yang telah dilatih.

Model klasifikasi diabetes telah berhasil dibangun menggunakan algoritma Support Vector Machine (SVM) dengan kernel linear. Setelah proses pelatihan, model ini dievaluasi menggunakan data pelatihan dan data pengujian. Hasil akurasi menunjukkan performa yang cukup baik dengan akurasi data pelatihan sebesar 78,34% dan akurasi data pengujian sebesar 77,27%. Selain itu, confusion matrix memberikan gambaran lebih rinci mengenai kinerja model dalam mengklasifikasikan pasien yang terkena diabetes atau tidak, dengan hasil yang cukup seimbang antara true positives, true negatives, false positives, dan false negatives. Model ini kemudian dapat digunakan untuk melakukan prediksi terhadap data baru, dan disimpan dalam bentuk file untuk keperluan penggunaan di masa depan tanpa harus melatih ulang.

# **BAB V KESIMPULAN**

## **5.1 Kesimpulan**

Penelitian ini telah berhasil merancang model klasifikasi risiko diabetes dengan memanfaatkan algoritma Support Vector Machine (SVM) berbasis kernel linear. Model ini dikembangkan menggunakan dataset Pima Indians Diabetes yang diperoleh dari Kaggle, mencakup berbagai parameter kesehatan seperti kadar glukosa darah, tekanan darah, indeks massa tubuh (BMI), dan jumlah kehamilan.

Proses persiapan data melibatkan langkah-langkah penting, termasuk pembersihan data, penanganan nilai yang hilang, normalisasi fitur, serta pembagian dataset menjadi data pelatihan dan data uji. Prosedur ini dilakukan untuk memastikan bahwa data yang digunakan dalam proses pelatihan model memiliki kualitas yang baik dan dapat meningkatkan keakuratan prediksi.

Hasil evaluasi menunjukkan bahwa model SVM yang dibangun memiliki tingkat akurasi sebesar 78,34% pada data pelatihan dan 77,27% pada data pengujian. Perbedaan yang kecil antara kedua hasil tersebut menunjukkan bahwa model mampu melakukan generalisasi dengan baik tanpa mengalami overfitting. Evaluasi model dilakukan menggunakan berbagai metrik, termasuk accuracy, precision, recall, F1-score, serta analisis confusion matrix untuk memahami pola prediksi yang benar maupun salah. Selain itu, validasi silang digunakan untuk memastikan bahwa model tetap konsisten dalam berbagai kondisi pengujian.

Model yang telah dilatih kemudian disimpan menggunakan pustaka pickle untuk memudahkan penggunaan di masa mendatang tanpa harus melakukan pelatihan ulang. Penelitian ini mengindikasikan bahwa algoritma SVM cukup efektif dalam memprediksi risiko diabetes berdasarkan data kesehatan yang tersedia.

Namun, terdapat beberapa keterbatasan dalam penelitian ini, khususnya terkait dengan cakupan dataset yang hanya mencakup data dari populasi wanita suku Pima Indians. Untuk meningkatkan kemampuan generalisasi model, penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih beragam dan mempertimbangkan penggunaan algoritma lain atau kombinasi algoritma untuk meningkatkan performa prediksi.

Secara keseluruhan, model prediksi yang dikembangkan dalam penelitian ini memiliki potensi untuk digunakan dalam mendukung deteksi dini dan pencegahan diabetes. Dengan adanya model ini, tenaga medis dapat lebih mudah mengidentifikasi individu yang berisiko tinggi terkena diabetes, sehingga dapat memberikan penanganan yang lebih cepat dan efektif untuk mengurangi dampak penyakit di masyarakat.

## **5.2 Saran**

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat diberikan untuk pengembangan lebih lanjut:

1. Peningkatan Dataset:

Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih beragam, mencakup data dari berbagai kelompok demografi, seperti usia, jenis kelamin, dan latar belakang etnis yang berbeda, agar model dapat memiliki kemampuan generalisasi yang lebih baik.

1. Penggunaan Algoritma Hybrid:

Untuk meningkatkan akurasi prediksi, disarankan untuk mengkombinasikan algoritma SVM dengan algoritma lain, seperti Random Forest atau Neural Network, guna memperoleh hasil yang lebih optimal.

1. Penerapan Feature Engineering:

Penambahan atau transformasi fitur yang relevan dapat meningkatkan kualitas data dan akurasi model. Misalnya, mempertimbangkan faktor gaya hidup, pola makan, dan aktivitas fisik dalam dataset.

1. Implementasi di Sistem Medis:

Pengembangan model lebih lanjut dapat diarahkan pada implementasi dalam aplikasi berbasis web atau perangkat lunak yang dapat digunakan oleh tenaga medis untuk memprediksi risiko diabetes secara langsung.

1. Pengujian di Lingkungan Nyata:

Model yang telah dikembangkan sebaiknya diuji di lingkungan nyata dengan data pasien yang lebih luas untuk memvalidasi keakuratan dan efektivitas model dalam mendukung diagnosis medis.

# **DAFTAR PUSTAKA**

1. Septiana, T., Muda, M. A., Budiyanto, D., Pratama, M., & Jaya, W. P. (2024). Analisis Penggunaan Support Vector Machine pada Deteksi Dini Penyakit Diabetes Melitus. *Jurnal Penelitian Inovatif*, *4*(3), 1631-1640.  
   <https://www.jurnal-id.com/index.php/jupin/article/view/643>
2. **Putri, R. A., & Nugroho, A. S.** (2023). *Analisis penggunaan support vector machine pada deteksi dini diabetes*. **Jurnal Pengembangan Informatika dan Komputer, 5**(3), 123–132.<https://jurnal-id.com/index.php/jupin/article/view/643>
3. Putra, M. A. F., & Wardhani, A. R. (2023). *Performance comparison of the kernels of support vector machine algorithm for diabetes mellitus classification*. *International Journal of Advanced Computer Science and Applications, 14*(2), 1–7.  
   <https://thesai.org/Publications/ViewPaper?Code=IJACSA&Issue=2&SerialNo=26&Volume=14>
4. **Universitas Perjuangan.** (n.d.). *Prediksi penyakit diabetes menggunakan algoritma support vector machine*. **Jurnal Informatika Universitas Perjuangan**. Diakses dari<https://e-journal.unper.ac.id/index.php/informatics/article/view/895>
5. Karatsiolis, S., & Schizas, C. N. (2012, November). Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset. In *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)* (pp. 139-144). IEEE.  
   <https://ieeexplore.ieee.org/abstract/document/6399663/>
6. Pujari, P. (2022). Classification of Pima Indian diabetes dataset using support vector machine with polynomial kernel. In *Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics* (pp. 55-67). CRC Press.

<https://www.taylorfrancis.com/chapters/edit/10.1201/9780367548445-5/classification-pima-indian-diabetes-dataset-using-support-vector-machine-polynomial-kernel-pujari>