

IMPLEMENTASI METODE SVM (SUPPORT VECTOR MACHINE) DALAM DATA PENDERITA PENYAKIT DIABETES

Queena Aurora Batubara¹, Puti Windrahmatullah², Mario Desendi Manurung³,
Muhammad Nabil Azizi⁴, Nabilah Wahyu Hafizhah⁵

Program Studi Sains Data Institut Teknologi Sumatera Jl. Terusan Ryacudu, Way Huwi,
Kec. Jati Agung, Kabupaten Lampung Selatan, Lampung
35365

E-mail : puti.120450070@student.itera.ac.id, queena.120450062@student.itera.ac.id,
mario.120450082@student.itera.ac.id, muhammad.120450090@student.itera.ac.id,
nabilah.120450106@student.itera.ac.id

Abstrak

Diabetes melitus (DM) merupakan penyakit kronis yang mematikan. Kenaikan gula darah yang tidak terkontrol, dirasakan, dan tidak terduga dengan cepat menyebabkan komplikasi. Data mining telah menggunakan banyak pendekatan untuk memprediksi penyakit, salah satunya adalah penggunaan Metode Support Vector Machine. Tujuan dari penelitian ini adalah untuk mengidentifikasi seseorang yang terkena Diabetes Melitus, agar masyarakat lebih sadar agar akan bahaya penyakit tersebut, metode SVM ini sangat cocok untuk digunakan karena SVM memiliki akurasi yang cukup bagus untuk sebuah klasifikasi. Dengan adanya program ini seseorang akan mengetahui seberapa persen kemungkinan untuk seseorang tersebut terkena Diabetes Melitus dan tidak terkena Diabetes Melitus, yang akan diketahui melalui persentase yang akan keluar setelah program dijalankan.

Kata Kunci: *Data Mining, Support Vector Machine, Diabetes*

1. PENDAHULUAN

1.1 Latar Belakang

Kesehatan adalah aspek yang sangat penting dalam kehidupan. Gangguan pada kesehatan manusia dapat mempengaruhi aktivitas sehari-hari. Salah satu penyakit yang dapat mengganggu aktivitas manusia adalah diabetes. Diabetes adalah penyakit kronis serius yang terjadi ketika pankreas tidak memproduksi cukup insulin, hormon yang mengatur kadar gula darah, atau ketika tubuh tidak dapat menggunakan insulin dengan efektif.

Diabetes terdiri dari beberapa tipe: Tipe 1, di mana sistem kekebalan tubuh menyerang dan menghancurkan sel-sel pankreas yang memproduksi insulin; Tipe 2, yang disebabkan oleh sel-sel tubuh yang kurang sensitif terhadap insulin, sehingga insulin yang diproduksi tidak dapat digunakan dengan baik; dan diabetes gestasional, yang terjadi pada ibu hamil akibat perubahan hormon dan kadar gula darah setelah persalinan.

Penting untuk memahami gejala awal diabetes guna mencegah penyakit ini sejak dini, karena deteksi lebih awal meningkatkan peluang untuk sembuh. Salah satu metode yang dapat digunakan untuk klasifikasi adalah Support Vector Machine (SVM). Penelitian sebelumnya menunjukkan bahwa SVM dapat memberikan skor akurasi klasifikasi yang tinggi. Dengan kemajuan dalam metode klasifikasi, metode yang baik dan efektif semakin banyak digunakan. Diharapkan bahwa penambahan metode smoothing pada SVM non-linear dapat meningkatkan kemampuan klasifikasi.

1.2 Tujuan dan Problem

Penelitian ini mengembangkan sistem machine vision untuk mendiagnosis prediksi apakah pasien menderita diabetes berdasarkan pengukuran diagnostik tertentu. Dataset yang digunakan mencakup data dari perempuan dengan rata-rata usia 21 tahun, dan mencakup variabel-variabel prediktor medis.

1.3 Data

Dataset yang digunakan adalah dataset diabetes yang diperoleh dari <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset> jumlah dataset terdiri dari 768 record dan 9 columns yang mana terdiri dari :

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome (Class)
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.228	33	1
...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Gambar 1 Dataset

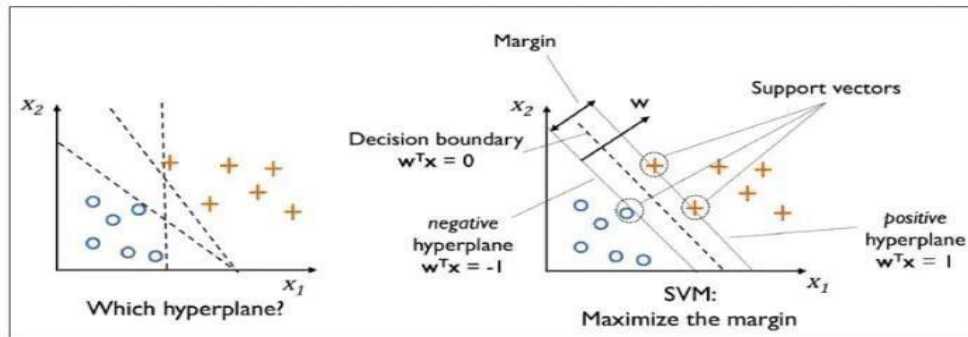
Keterangan dataset :

1. Pregnancies: Untuk menyatakan jumlah kehamilan
2. Glucose: Untuk mengekspresikan tingkat Glukosa dalam darah
3. Blood Pressure: Untuk menyatakan pengukuran tekanan Darah dalam (mmHg)
4. Skin Thickness: Untuk menyatakan ketebalan kulit (mm)
5. Insulin : Untuk mengekspresikan tingkat Insulin dalam darah (mu U/ml)
6. BMI(Body Mass Index): Untuk mengekspresikan indeks massa tubuh
7. DiabetesPedigreeFunction: Untuk menyatakan persentase Diabetes (berat dalam kg / tinggi dalam m) dikali 2
8. Age : Untuk menyatakan umur
9. Outcome(Class): Untuk menyatakan hasil akhir 1 adalah Ya dan 0 adalah Tidak

2. METODE

2.1 Gambaran Umum

Support Vector Machines (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi (Santosa, 2007). Pada ruang berdimensi tinggi, akan dicari hyperplane (fungsi pemisah) yang dapat memaksimalkan jarak (margin) antara dua kelas data. Klasifikasi menggunakan SVM dapat dijelaskan secara sederhana yaitu usaha untuk mendapatkan garis sebagai fungsi pemisah terbaik yang dapat memisahkan dua kelas yang berbeda (+1,-1) pada ruang

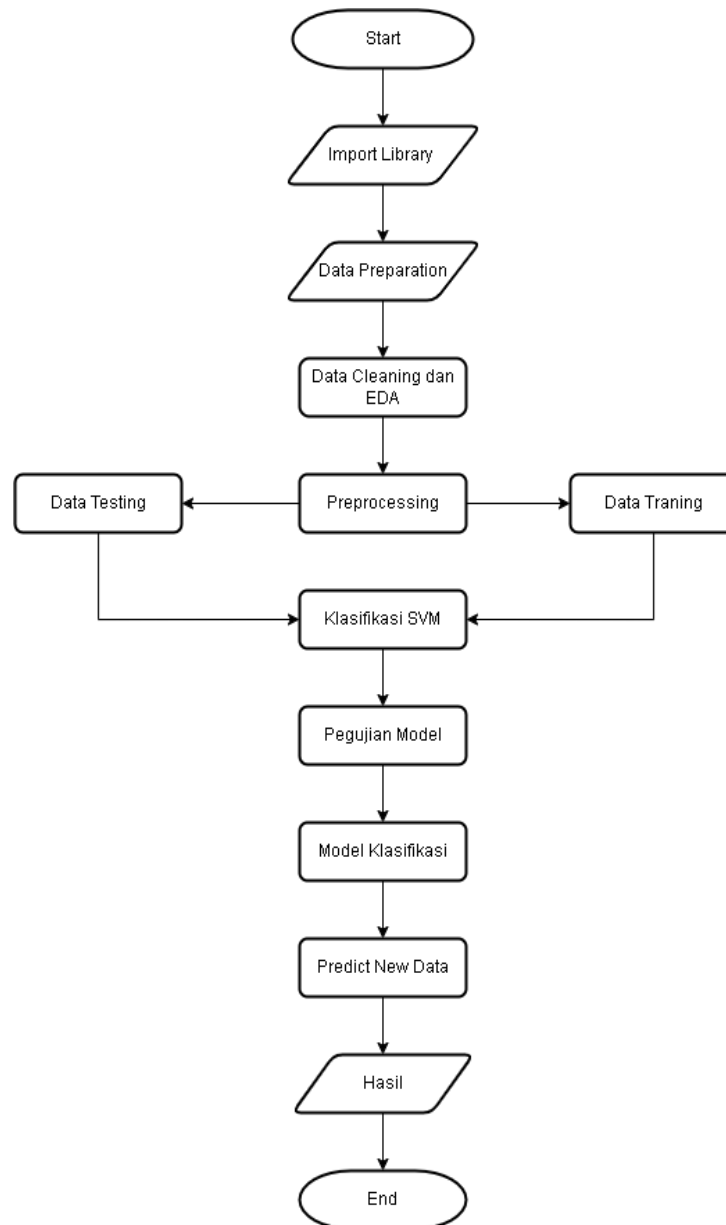


Gambar 2 Cara Kerja SVM

Hyperplane yang ditemukan oleh SVM seperti terlihat pada Gambar , posisinya berada di antara dua kelas, artinya jarak antara hyperplane dengan objek data berbeda dengan kelas tetangga (luar), yang ditunjukkan dengan lingkaran kosong dan positif.

Dalam SVM, objek data terluar yang paling dekat dengan hyperplane disebut support vector. Objek yang disebut support vector adalah yang paling sulit untuk diklasifikasikan karena hampir tumpang tindih dengan kelas lain. Karena sifatnya yang kritis, hanya support vector ini yang dianggap dapat menemukan hyperplane paling optimal menggunakan SVM.

2.2 Flowchart Sistem



Gambar 3 Flowchart

2.3 Metrics Evaluation

Matriks evaluasi digunakan untuk mengukur kualitas statistik dan machine learning, mengevaluasi model algoritma dalam model apapun sangat penting ada banyak model untuk mengevaluasi suatu data termasuk akurasi. dikarenakan suatu model mungkin bekerja dengan baik menggunakan matriks evaluasi, metrik evaluasi sangat penting untuk memastikan apakah model yang digunakan sudah benar dan optimal.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Gambar 4 Matriks Evaluasi

Keterangan:

1. Accuracy: tingkat keakuratan antara data prediksi benar positif yang diminta dengan hasil prediksi yang diberikan oleh model
2. Recall: keberhasilan model dalam menemukan kembali sebuah informasi.

2.4 Data Splitting

Pada data splitting dilakukan pembagian data dengan membagi data menjadi 2 bagian data test dan train dengan membagi sebesar 20% dari data, data splitting ini digunakan untuk membuat model berdasarkan data karena semakin besar nilai data training maka dapat memastikan data keseluruhan dengan karakteristik yang berbeda.

2.5 Cross - Validation

Pada Tahap ini Pengujian sistem dilakukan dengan melihat jumlah nilai dari setiap variabel serta parameter-parameter lainnya, yang dimana parameter tersebut akan di proses dan kemudian akan di kalsifikasikan dalam dua kelas yaitu terkena diabetes dan yang tidak terkena diabetes

3. HASIL DAN ANALISIS

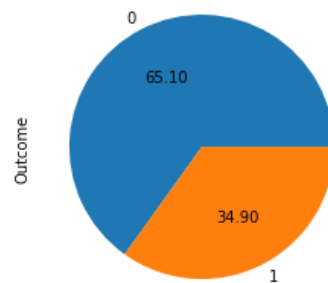
3.1 Hasil Data cleaning dan EDA

Tahap awal yang dilakukan sebelum ke processing data adalah proses pembersihan data

3.2 Hasil Preprocessing

Hasil penelitian yang telah diuji pada dataset diabetes pada tahap ini menggunakan metode klasifikasi SVM. sebelum dilakukan tahap preprocessing dengan memeriksa dataset dimana semua kolom sudah memiliki tipe yang sesuai dan mengubah data pada kolom class menjadi 0 dan 1, berikut hasil dari dataset yang sudah dilakukan preprocessing.

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome (Class)
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.228	33	1
...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0



Gambar 5 Dataset yang telah dilakukan preprocessing

3.3. Tahap Hasil Pengujian

```
[ ]
svc=svm.SVC(kernel='rbf')
svc.fit(X_train,y_train)

SVC()

[ ] #akurasi dari prediksi data Train

X_train_prediction = svc.predict(X_train)
train_data_accuracy = accuracy_score(X_train_prediction, y_train)
print('Accuracy Score:\n',
      train_data_accuracy)

Accuracy Score:
0.7638436482884691

[ ] #akurasi dari prediksi data Test

X_test_prediction = svc.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, y_test)
print('Accuracy Score:\n',
      test_data_accuracy)

Accuracy Score:
0.7857142857142857
```

Gambar 6 Pelatihan Model

Pada langkah pertama, model SVM dilatih menggunakan data train. Setelah pelatihan, model ini diuji untuk memprediksi data train tersebut, dan didapatkan nilai akurasi sebesar 0.76. Nilai ini menunjukkan seberapa baik model SVM dapat mengklasifikasikan data yang telah dilatih.

Selanjutnya, model SVM yang telah dilatih diuji kembali menggunakan data test yang terpisah dari data train. Dari hasil prediksi terhadap data test, model ini berhasil mencapai akurasi sebesar 0.78. Ini menunjukkan bahwa model SVM dengan kernel RBF yang digunakan memiliki performa yang lebih baik saat diuji pada data yang belum pernah dilihat sebelumnya dibandingkan dengan saat diuji pada data latihannya.

Secara keseluruhan, hasil pengujian menunjukkan bahwa model SVM dengan kernel RBF memiliki kemampuan yang cukup baik dalam memprediksi data baru, meskipun ada sedikit perbedaan akurasi antara data train dan data test.

```
[ ] y_pred = svc.predict(X_test)

cm = confusion_matrix(y_test, y_pred)
print(cm)
print(f1_score(y_test, y_pred))
print(accuracy_score(y_test, y_pred))

[[95  5]
 [28 26]]
0.6117647058823529
0.7857142857142857
```

Gambar 7 Hasil Confusion Matriks

Dapat dilihat pada gambar 7, hasil evaluasi performa model SVM dengan kernel RBF pada data uji, yang mencakup matriks kebingungan (confusion matrix), F1-score, dan akurasi. Matriks kebingungan menunjukkan bahwa dari total data uji, model berhasil mengklasifikasikan 95 sampel positif dengan benar (True Positives) dan 26 sampel negatif dengan benar (True Negatives). Namun, terdapat 5 sampel positif yang salah diklasifikasikan sebagai negatif (False Negatives) dan 28 sampel negatif yang salah diklasifikasikan sebagai positif (False Positives). Berdasarkan perhitungan F1-score yang diperoleh sebesar 0.6117, model memiliki keseimbangan yang cukup baik antara precision dan recall, meskipun masih terdapat ketidakseimbangan yang perlu diperbaiki. Akurasi model tercatat sebesar 0.7857, yang menunjukkan bahwa model berhasil memprediksi dengan benar sekitar 78.57% dari total data uji. Secara keseluruhan, hasil ini mengindikasikan bahwa model SVM dengan kernel RBF memiliki performa yang baik,

```

#Membuat Prediksi Data Baru
scaler = StandardScaler()
scaler.fit(X)

####dengan Pregnancies 0
#### Glucose 120
#### BloodPressure 65
#### SkinThickness 19
#### Insulin 175
#### BMI 25.8
#### DiabetesPedigreeFunction 0.287
#### Age 20

datapasi_baru = (0,120,65,19,175,25.8,0.287,20)

#Merubah data list menjadi array
data_array = np.asarray(datapasi_baru)

#mengubah Array
data_reshape = data_array.reshape(1,-1)

#standarisasi data
data_standard = scaler.transform(data_reshape)

prediction = svc.predict(data_standard)

```

```

if(prediction[0] == 0):
    print('Pasien Dengan Data di atas Terhindar Dari DIABETES');
else:
    print('Pasien Dengan Data di atas TERJANGKIT DIABETES');

```

Pasien Dengan Data di atas Terhindar Dari DIABETES

Gambar 8 Prediksi dengan data baru

Gambar 8 menunjukkan proses prediksi diagnosis diabetes pada pasien baru menggunakan model SVM yang telah dilatih sebelumnya. Langkah pertama adalah normalisasi data pasien baru dengan StandardScaler untuk memastikan data memiliki skala yang sama dengan data pelatihan. Data pasien baru mencakup variabel seperti jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh (BMI), DiabetesPedigreeFunction, dan usia. Setelah data distandardisasi, model SVM digunakan untuk memprediksi kemungkinan pasien terkena diabetes. Hasil prediksi menunjukkan bahwa pasien dengan data tersebut terhindar dari diabetes, yang diindikasikan oleh keluaran `prediction[0] == 0`

4. KESIMPULAN

Penelitian ini dilakukan dikembangkan untuk memprediksi seseorang apakah ia terkena Diabetes Melitus atau tidak. Berdasarkan penelitian yang dilakukan menggunakan metode *SVM* dapat ditarik kesimpulan dari sebagai berikut :

- Penerapan metode SVM ini untuk memprediksi penyakit diabetes sangat membantu untuk mengetahui apakah seseorang tersebut terkena diabetes atau tidak yang dimana diukur dari parameter-parameter yang ditentukan
- Model yang dibuat yaitu dengan memperhitungkan semua variabel yang ada seperti Glucose,Bloodplessure,Skin Thickness,Insulin,BMI,DiabetesPedigreeFunction,Age,sehingga dapat dihasilkan output dari hasil data baru bahwa pasien dengan parameter tersebut tidak terjangkit diabetes sehingga dapat disimpulkan bahwa penelitian ini berhasil yang dilihat dari tingkat akurasi klasifikasinya sehingga metode SVM

ini dapat diterapkan untuk menentukan apakah seseorang terkena diabetes atau tidak

REFERENSI

Nugroho, R.A., Tarno, T. and Prahutama, A., 2017. Klasifikasi Pasien Diabetes Mellitus Menggunakan Metode Smooth Support Vector Machine (SSVM). *Jurnal Gaussian*, 6(3), pp.439-448.

Mucholladin, A.W., Bachtiar, F.A. and Furqon, M.T., 2021. Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, 2548, p.964X.

Kumari, V.A. and Chitra, R., 2013. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2), pp.1797-1801.

LAMPIRAN

Berikut kami lampirkan code dan dataset yang telah dibuat pada percobaan kali ini :

Code :

https://colab.research.google.com/drive/1eIZxPmfKs2T8jXcsFaWjZn8J_Lzle5Pq?usp=sharing

Dataset :

<https://drive.google.com/file/d/18bbJPpw1IC2hjwfGhlaNbkq8TDJA4rxR/view?usp=sharing>