



Classification of Diabetes Mellitus Using C4.5 Algorithm, Support Vector Machine (SVM) and Linear Regression

Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma C4.5, Support Vector Machine (SVM) dan Regresi Linear

Dwi Sri Rahayu^{1*}, Nursafika², Jihan Afifah³, Sri Intan⁴

^{1,2,3,4}Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarif Kasim Riau

E-Mail: ¹12050320393@students.uin-suska.ac.id, ²12050320473@students.uin-suska.ac.id,
³12050321639@students.uin-suska.ac.id, ⁴12050326125@students.uin-suska.ac.id

Corresponding Author: Dwi Sri Rahayu

Abstract

WHO predicts the number of diabetics in Indonesia will increase from 8 million in 2000 to around 21.3 million in 2030. The IDF also predicts the number of diabetics will increase from 7 million in 2009 to 12 million in 2030. Unknown, doctor believe that the disease is related to lifestyle and diet, which play an important role in the disease. Data mining can be used as a reference for predicting and diagnosing the nature of the disease with the right method. The purpose of this study is to carry out the process of classifying diabetes using the C4.5 algorithm, support vector machine and linear regression method. After the data is processed using the classification technique using the C4.5 data mining algorithm, the results obtained by dividing the data using the holdout method are obtained by dividing the training data by 90 test data by 10%, which results in an accuracy of 75.324%. The SVM algorithm obtained a result of 82.01% and for processing using Linear Regression the MSE result was 0.216. Then when compared, the SVM algorithm is better than the C4.5 algorithm in the data used in this study.

Keyword: C4.5, Classification, Diabetes, Linear Regression, Support Vector Machine (SVM)

Abstrak

WHO memprediksi jumlah penderita diabetes di Indonesia akan meningkat dari 8 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030. IDF juga memprediksi jumlah penderita diabetes akan meningkat dari 7 juta pada tahun 2009 menjadi 12 juta pada tahun 2030. Tidak dikenal, dokter percaya bahwa penyakit ini terkait dengan gaya hidup dan pola makan, yang berperan penting dalam penyakit tersebut. Data mining dapat dijadikan acuan untuk memprediksi dan mendiagnosa sifat penyakit dengan metode yang tepat. Tujuan dari penelitian ini adalah melakukan proses klasifikasi penyakit diabetes dengan menggunakan algoritma C4.5, support vector machine dan metode regresi linier. Setelah data diolah dengan teknik klasifikasi dengan algoritma data mining C4.5, hasil yang diperoleh dengan membagi data dengan metode holdout diperoleh dengan membagi data latih dengan 90 data uji dengan 10%, yaitu hasilnya dalam akurasi 75,324%. Pada algoritma SVM didapatkan hasil 82,01% dan untuk pengolahan menggunakan Regresi Linear didapat hasil MSE 0,216. Kemudian jika dibandingkan maka Algoritma SVM lebih baik dari pada algoritma C4.5 pada data yang digunakan pada penelitian kali ini.

Kata Kunci: C4.5, Diabetes, Klasifikasi, Regresi Linear, Support Vector Machine (SVM)

1. PENDAHULUAN

Di dunia global saat ini, banyak penyakit kronis menyebar ke seluruh dunia, dan penyakit serius seperti itu menyebar baik di negara berkembang maupun negara maju. Di antara penyakit serius tersebut, diabetes adalah penyakit kronis yang membunuh kaum muda dunia [1]. Diabetes adalah penyakit di mana

kemampuan tubuh untuk memproduksi atau merespon insulin terganggu, menyebabkan metabolisme karbohidrat yang tidak normal dan peningkatan kadar gula darah dan urin. Meski penyebab sebenarnya dari penyakit ini masih belum diketahui, dokter percaya bahwa penyakit ini berkaitan dengan gaya hidup dan pola makan [2]. Penderita diabetes pula berisiko mengalami perkara kesehatan sekunder misalnya penyakit jantung & kerusakan saraf [3].

WHO memperkirakan jumlah penderita diabetes pada Indonesia akan semakin tinggi berdasarkan 8 juta dalam tahun 2000 sebagai lebih kurang 21,3 juta dalam tahun 2030. International Diabetes Federation (IDF) pula memperkirakan jumlah penderita diabetes akan semakin tinggi berdasarkan 7 juta dalam tahun 2009 sebagai 12 juta dalam tahun 2030. Jumlah ini akan berlipat ganda atau 3 kali lipat dalam tahun 2030 [4]. Pada tahun 2035, jumlah ini diperkirakan akan meningkat menjadi 592 juta. Dari 382 juta ini, sekitar 175 juta tetap tidak terdeteksi dan mengembangkan komplikasi progresif yang tidak terdeteksi dan berisiko mengalami komplikasi yang tidak dapat dihindari [5]. Sementara itu, dari data yg dihimpun Survei Kesehatan Dasar (Riskesdas) 2013, diketahui 2.650 orang mengidap diabetes [6]. Karena prevalensi diabetes yang semakin meningkat, penyakit ini harus didiagnosis sejak dini untuk menghindari komplikasi dan mengurangi risiko masalah kesehatan [7]. Oleh sebab itu, maka diperlukan penggunaan teknologi data mining.

Dengan metode yang tepat, data mining dapat digunakan sebagai standar untuk memprediksi dan mendiagnosa penyakit. Diantara sekian banyak penyakit, diabetes merupakan salah satu penyakit degeneratif yang diprediksi dengan metode data mining [8]. Data mining punya beberapa fungsi yaitu: Fungsi estimasi (*estimation*), Fungsi deskripsi (*description*), Fungsi klasifikasi (*classification*), Fungsi prediksi (*prediction*), Fungsi pengelompokan (*cluster*), Fungsi asosiasi (*association*) [9]. Melakukan teknik pengolahan data mining yang salah satunya dapat diimplementasikan menggunakan algoritma klasifikasi. Klasifikasi adalah metode pengelompokan objek berdasarkan pola yang terkait dengan objek taksonomi dan menggunakan pola tersebut untuk mengklasifikasikan informasi baru [10].

Berdasarkan penelitian sebelumnya, banyak dari solusi ini mengambil pendekatan algoritmik yang berbeda, dimulai dengan pembelajaran yang mendalam, yaitu *Neural Network* [11], hingga menggunakan machine learning seperti C4.5 [12], Naive Bayes [13], *Logistic Regression* dan SVM Linear [14]. Telah ada penelitian sebelumnya yang menggunakan algoritma C4.5 untuk memprediksi diabetes berdasarkan faktor penyebab diabetes seperti jumlah wanita melahirkan, kadar gula darah, tekanan darah, insulin, *Body Massa Indeks*, usia, dan hasil kelas menghasilkan akurasi sebesar 70.32% [15]. Penelitian Dinyal Amru (2019) mengklasifikasikan data diabetes menggunakan Support Vector Machine (SVM) dalam sistem klasifikasi ini memberikan hasil 77,92% [16]. Untuk penelitian ini penulis menggunakan algoritma C4.5, Support Vector machine dan metode Regresi Linear untuk mengklasifikasi penyakit diabetes.

Algoritma C4.5 merupakan algoritma klasifikasi pohon keputusan yang banyak digunakan karena memiliki kelebihan yang besar dibandingkan algoritma lainnya. Salah satu keunggulan algoritma C4.5 adalah dapat menghasilkan pohon keputusan yang mudah dipahami, cukup akurat, dan efisien dalam menangani atribut numerik maupun diskrit [17]. SVM adalah metode pembelajaran terawasi untuk menganalisis data dan mengidentifikasi pola yang digunakan dalam klasifikasi [18]. SVM memiliki keuntungan untuk dapat mendefinisikan superclass yang berbeda yang memaksimalkan margin antara dua kelas yang berbeda [19]. Salah satu metode yang dapat digunakan untuk memprediksi diabetes adalah regresi linier. Metode ini merupakan metode peramalan yang menggunakan garis lurus untuk menggambarkan interaksi antara 2 variabel atau lebih [20]. Dalam penelitian ini penulis mengklasifikasi menggunakan tiga metode tersebut, kemudian membandingkan hasil klasifikasi antara Algoritma SVM dan algoritma C4.5.

Berdasarkan pemaparan di atas, maka perlu dipelajari klasifikasi penyakit diabetes dengan algoritma C4.5, support vector machine dan metode regresi linier.

2. METODOLOGI PENELITIAN

Adapun Metodologi yang digunakan pada penelitian ini dapat dilihat pada Gambar 1 yang dijabarkan sebagai berikut.

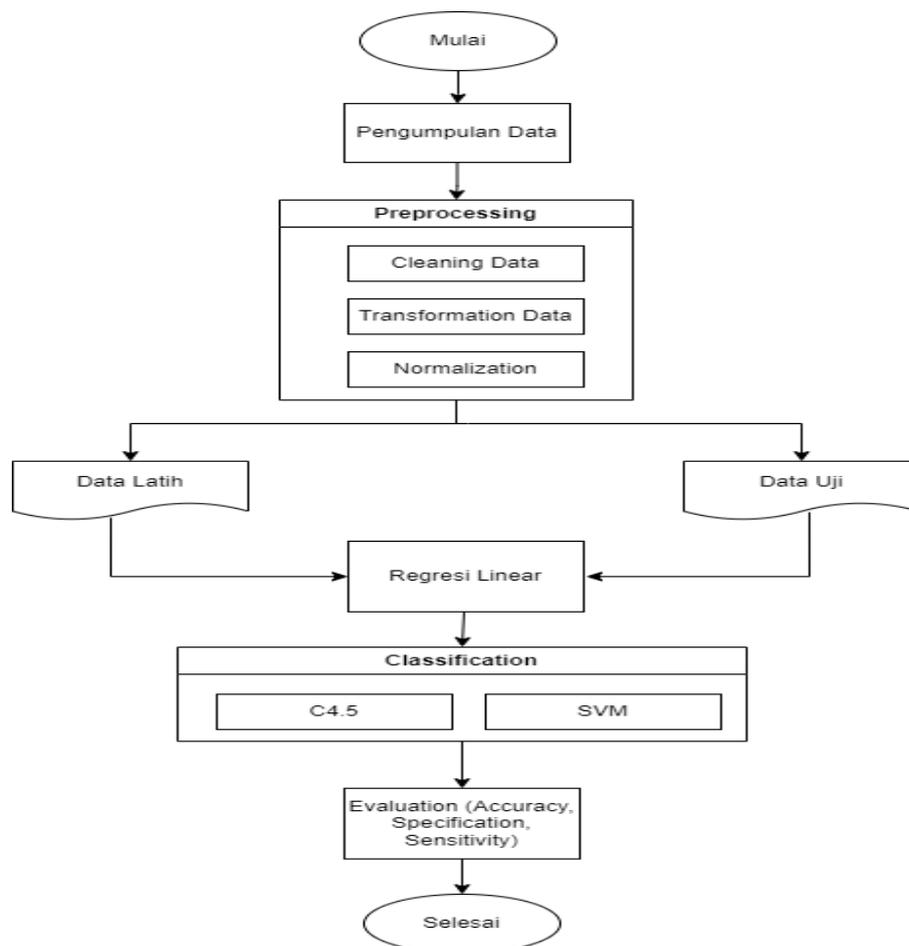
a) Pengumpulan Data

Dalam penelitian ini dataset yang digunakan adalah data mengenai penyakit diabetes. Dataset ini bersumber dari Kaggle dataset repository (UCI Pima Indians Diabetes Database), dataset penyakit diabetes memiliki 768 sample dan 9 atribut yang terdiri dari Pregnant, Glucose, Bp, Skin, Insulin, BMI, Pedigree, Age dan Diabetes.

b) Data Preprocessing

Data Preprocessing merupakan kegiatan pemrosesan data untuk mengubah atau menyandikan data sehingga dapat dengan mudah diinterpretasikan oleh mesin. Tahapannya adalah cleaning data, transformasi data dan normalisasi data.

- c) **Pembagian Data**
Pembagian data merupakan teknik yang digunakan untuk membagi data menjadi data latih dan data uji. Teknik yang dipakai adalah *Hold Out* yaitu dengan membagi data latih sebesar 90% dan data uji sebesar 10%.
- d) **Pengolahan Menggunakan C4.5**
Pengolahan menggunakan algoritma C4.5 untuk mengolah data menggunakan algoritma C4.5 setelah data di *processing*.
- e) **Evaluasi**
Evaluasi merupakan teknik yang dilakukan untuk menentukan skor akurasi dari suatu algoritma.
- f) **Regresi Linear**
Regresi linier adalah metode prediksi yang menggunakan garis lurus untuk menggambarkan hubungan antara dua variabel atau lebih. Regresi linier dengan satu variabel independen disebut regresi linier sederhana, sedangkan regresi linier dengan lebih dari satu variabel bebas regresi linier berganda [20].
- g) **Google Colab**
Google Colab adalah format yang dimodifikasi dari *Notebook Jupyter* yang disediakan untuk Google dan platform ini banyak digunakan untuk Machine Learning maupun *Deep Learning* [21].
- h) **Pengolahan Menggunakan SVM**
Pengolahan menggunakan algoritma SVM merupakan kegiatan untuk mengolah data menggunakan algoritma SVM setelah data di *processing*.
- i) **Evaluasi**
Evaluasi merupakan teknik yang dilakukan untuk menentukan skor akurasi dari suatu algoritma.



Gambar 1. Metodologi Penelitian

2.1 Data

Data menurut *Webster New World Dictionary* adalah segala sesuatu yang diketahui atau dicurigai. Apa yang terjadi adalah fakta yang diketahui (bukti). Data dapat memberikan gambaran tentang suatu keadaan atau masalah. Data juga dapat diartikan sebagai kumpulan informasi atau nilai data yang diperoleh dari pengamatan (pengamatan) terhadap suatu objek berupa angka dan simbol atau fungsi [22]. Data dapat berupa nilai (angka, string, atau simbol). Pengertian data diperluas : fakta berguna, teks, grafik, audio-video dalam konteks [23].

2.2 Data Mining

Data mining adalah sebuah proses menemukan hubungan yang bermakna, pola, dan tren dengan memeriksa dalam kumpulan data yang besar disimpan dalam stok dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [24]. *Supervised learning* merupakan model prediktif yang mekanisme kerjanya adalah melatih set data dengan variabel target dan kemudian mengklasifikasikannya sesuai dengan tingkat kemiripan antar data. Model prediktif sering digunakan untuk memprediksi fenomena masa depan berdasarkan kumpulan data masa lalu. Algoritma data mining termasuk dalam jenis pembelajaran terawasi yaitu klasifikasi [25].

2.3 Klasifikasi

Klasifikasi adalah teknik yang digunakan untuk memprediksi kelas atau atribut dari setiap instance data. Klasifikasi memetakan data ke dalam kelompok kelas yang telah ditentukan sebelumnya. Klasifikasi juga dikenal sebagai pembelajaran terawasi karena kelas data sudah ditentukan sebelumnya [26]. Beberapa Algoritma klasifikasi yang umum digunakan adalah NBC, K-Nearest Neighbor, Decision Tree Induksi, inferensi kasus, algoritma genetika, dan mesin vektor pendukung. Algoritma klasifikasi bekerja berdasarkan 4 komponen utama, yaitu:

1. Label adalah variabel target yang mewakili objek taksonomi.
2. Prediktor adalah variabel independen (atribut) dari data yang diklasifikasikan.
3. Data latih (*training set*) adalah kumpulan data latih untuk menjalankan fungsi algoritma.
4. Kumpulan data uji (*test suite*) untuk menghitung akurasi uji [27].

2.4 Algoritma C4.5

Merupakan salah satu algoritma yang digunakan untuk membuat pohon keputusan. Metode Decision Tree (pohon keputusan) mengubah kejadian yang sangat besar menjadi *Decision Tree* yang mewakili aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami [28]. Algoritma C4.5 dibuat oleh Ross Quinlan, yang merupakan evolusi dari ID3 yang juga dibuat oleh Quinlan. Beberapa pengembangan yang dilakukan pada C4.5 adalah dapat memperbaiki missing value, dapat memperbaiki data persisten dan truncation. Ada beberapa langkah untuk membuat pohon keputusan dengan algoritma C4.5, yaitu:

1. Siapkan data pelatihan. Data pelatihan biasanya diambil dari data historis yang terjadi sebelumnya dan telah dikelompokkan ke dalam kelas-kelas tertentu.
2. Identifikasi akar pohon. Sebelum menghitung gain atribut, terlebih dahulu hitung nilai entropy menggunakan Rumus 1 sebagai berikut:

$$Entropy(s) = - \sum_{i=1}^n \frac{s_i}{s} - p_i * \log_2 p_i \quad (1)$$

Keterangan:

- S : himpunan kasus
- A : atribut
- n : jumlah partisi S
- pi : proporsi dari Si terhadap S

3. Menentukan nilai gain dengan metode informasi gain:

$$Gain(S, A) = Entropy(s) - \sum_{i=1}^n \frac{s_i}{s} * entropy(S_i) \quad (2)$$

Keterangan:

- S : himpunan kasus
- A : atribut
- n : jumlah partisi atribut A
- |Si| : jumlah kasus pada partisi ke-i
- |S| : jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua kasusterpartisi.
5. Proses partisi pohon keputusan akan berhenti saat:
 - a) Semua kasus dalam node N mendapat kelas yang sama.
 - b) Tidak ada atribut di dalam kasus yang dipartisi lagi.
 - c) Tidak ada kasus di dalam cabang yang kosong

2.5 Algoritma Support Vector machine

Pengertian *Support Vector Machine* yaitu sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi – fungsi linier dalam sebuah fitur yang berdimensi tinggi dan dilatih dengan menggunakan algoritma pembelajaran yang didasarkan pada teori optimasi [29]. Tingkat akurasi pada model yang akan dihasilkan oleh proses peralihan dengan SVM sangat bergantung terhadap fungsi kernel dan parameter yang digunakan.

2.6 Metode Regresi Linear

Langkah-langkah prediksi dapat dilakukan dengan menggunakan perhitungan rumus regresi linier. Terdapat dua jenis rumus regresi sebagai langkah dalam proses analisis prediktif, yaitu regresi linier berganda dan regresi linier sederhana. Regresi linier adalah suatu cara untuk mengukur hubungan korelasi antara dua variabel atau lebih yang digunakan untuk data prediktif melalui garis lurus [30]. Ada 2 tipe variabel dalam regresi linier yaitu: variabel pemberi pengaruh dan variabel terpengaruh. Variabel pemberi pengaruh digunakan sebagai sebab, sedangkan variabel terpengaruh sebagai akibat [31]. Perhitungan Regresi Linear dapat dilakukan dengan persamaan berikut.

$$\beta_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} \quad (3)$$

Keterangan:

- y : data uji
- x : persentase produktif (P.E) rasio
- n : total jumlah terjadinya variabel

3. HASIL DAN PEMBAHASAN

3.1 Data

Dataset yang digunakan pada penelitian ini diambil melalui website Kaggle.com dimana data tersebut berjumlah 768 baris.

Tabel 1. Dataset

No	Pregnant	Glucose	Bp	Skin	insulin	bmi	pedigree	age	diabetes
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
....
768	1	93	70	31	0	30.4	0.315	23	0

Pada data diatas telah dilakukan normalisasi dengan atribut *Glucose, Diabetes Pedigree, Blood Pressure, Skin, Pregnant, Age, BMI, Insulin*, diabetes. Penjelasannya adalah sebagai berikut:

1. *Pregnant* adalah Berapa kali hamil.
2. *Glucose* adalah konsentrasi plasma gula dalam 2 jam pada tes glukosa oral.
3. *Blood Pressure* (Bp) adalah Tekanan darah dengan satuan mm Hg.
4. *Skin* adalah Ketebalan lipatan kulit trisep dengan satuan mm.
5. *Insulin* adalah Pemberian serum insulin setiap 2 jam dengan satuan mu U/ml.
6. *BMI* (*Body Mass Index*) adalah pengukuran presentase lemak di dalam tubuh.
7. *Pedigree* adalah indikasi diabetes turunan.
8. *Age* usia (Tahun).
9. diabetes (0 atau 1) dimana 1 merupakan kelas terindikasi terkena diabetes dan 0 merupakan kelas terindikasi tidak terkena diabetes.

3.2 Hasil Pada C4.5

Berikut ini adalah hasil dan pembahasan pada proses pengolahan data menggunakan C4.5, dimana data dibagi menggunakan teknik *hold out* dengan 90% data latih dan 10% data testing, dan dataset yang digunakan berjumlah 768 data yang sudah dilakukan proses cleaning, transformasi, dan normalisasi.

Tabel 2. Hasil Akurasi menggunakan algoritma C4.5

hasil	precision	recall	f1-score	support
0	0,75	0,89	0,82	47
1	0,76	0,53	0,63	30
accuracy			0,75	77
macro avg	0,76	0,71	0,72	77
weighted avg	0,75	0,75	0,74	77

Berdasarkan tabel diatas, hasil akurasi klasifikasi diabetes menggunakan algoritma C4.5 dengan menggunakan tools google colab yaitu 75,324%.

3.3 Hasil Algoritma SVM

Hasil akurasi pada proses pengolahan data menggunakan SVM, dimana data dibagi menggunakan teknik *hold out* juga dengan 90% data latih dan 10% data testing, dan dataset yang digunakan berjumlah 768 data yang sudah dilakukan proses cleaning, transformasi, dan normalisasi adalah sebagai berikut:

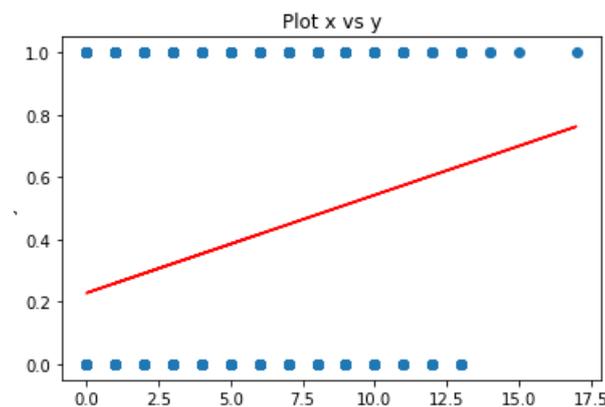
Tabel 3. Hasil Akurasi menggunakan algoritma SVM

hasil	precision	recall	f1-score	support
0	0,81	0,91	0,86	47
1	0,83	0,67	0,74	30
accuracy			0,82	77
macro avg	0,82	0,79	0,8	77
weighted avg	0,82	0,82	0,81	77

Berdasarkan tabel diatas, hasil akurasi klasifikasi diabetes menggunakan algoritma SVM dengan menggunakan tools google colab yaitu 82,01%.

3.4 Hasil Menggunakan Regresi Linear

a. Hasil Klasifikasi



Gambar 2. Grafik hasil Klasifikasi diabetes menggunakan Regresi Linear

b. Hasil MSE

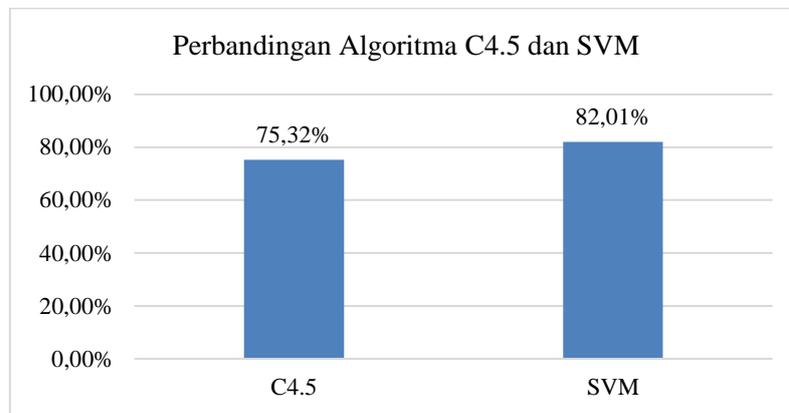
Tabel 4. Hasil Mean Square Error menggunakan Regresi Linear

MSE	0,216000031
R2	0,04923879

Mean Squared Error (MSE) adalah Rata-rata Kesalahan kuadrat antara nilai aktual dan nilai peramalan. Metode Mean Squared Error secara umum digunakan untuk mengecek estimasi berapa nilai

kesalahan pada peramalan. Nilai Mean Squared Error yang rendah atau nilai mean squared error mendekati nol menunjukkan bahwa hasil peramalan sesuai dengan data aktual dan bisa dijadikan untuk perhitungan peramalan di periode mendatang. Pada penelitian ini prediksi diabetes menggunakan Regresi linear mempunyai nilai MSE 0,216 yang dimana nilainya mendekati angka 0, maka dapat diambil kesimpulan bahwa hasil peramalan sesuai dengan data aktual dan bisa dijadikan untuk perhitungan peramalan di periode mendatang.

3.5 Perbandingan hasil Akurasi Algoritma C.45 dan SVM



Gambar 3. Perbandingan Hasil Akurasi C4.5 dan SVM

Dari gambar diatas dapat dilihat SVM memiliki akurasi lebih tinggi dibandingkan C.45, dimana akurasi SVM yaitu 82,01% sedangkan C.45 yaitu 75,324%. Dari hasil tersebut SVM lebih baik dibandingkan C.45 terhadap data yang digunakan saat ini.

4. KESIMPULAN

Setelah menerapkan Algoritma C4.5, SVM dan metode Regresi Linear untuk mengklasifikasikan tingkat akurasi penyakit diabetes, diketahui bahwa kinerja dari algoritma SVM mempunyai hasil akurasi yang lebih tinggi daripada Algoritma C4.5. terbukti Setelah dilakukan pengolahan data menggunakan teknik *Hold Out* dengan membagi data training 90% dan data testing sebanyak 10% dalam penerapan algoritma C4.5 didapatkan lah hasil 75,324% dan untuk pengolahan menggunakan SVM didapatkan hasil 82,01%. Kemudian dengan menggunakan metode Regresi linear dihasilkan nilai MSE 0,216. Tentunya hal ini dipengaruhi oleh jumlah data yang digunakan dan preprocessing yang dilakukan pada penelitian ini.

REFERENSI

- [1] M. Alehegn, R. Joshi, and P. Mulay, "Analysis and prediction of diabetes mellitus using machine learning algorithm," *Int. J. Pure Appl. Math.*, vol. 118, no. 9, pp. 871–878, 2018.
- [2] H. S. W. Hovi, A. I. Hadiana, and F. R. Umbara, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," *Informatics Digit. Expert*, vol. 4, no. 1, pp. 40–45, 2022.
- [3] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: review and case study," *Appl. Sci.*, vol. 9, no. 21, p. 4604, 2019.
- [4] D. R. Ente, S. A. Thamrin, S. Arifin, H. Kuswanto, and A. Andreza, "Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4. 5," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 80–88, 2020.
- [5] B. Bertalina and P. Purnama, "Hubungan lama sakit, pengetahuan, motivasi pasien dan dukungan keluarga dengan kepatuhan diet pasien Diabetes Mellitus," *J. Kesehat.*, vol. 7, no. 2, pp. 329–340, 2016.
- [6] F. I. Kurniadi and V. K. Putri, "Penggunaan Heaviside Activation Function pada Regresi Linear untuk Klasifikasi Diabetes," *Ultim. J. Tek. Inform.*, vol. 10, no. 1, pp. 7–10, 2018.
- [7] J. J. Purnama, S. Rahayu, S. Nurdiani, T. Haryanti, and N. A. Mayangky, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Diabetes," *Indones. J. Comput. Inf. Technol. Retrieved from <http://ejournal.bsi.ac.id/ejurnal/index.php/ijcit>*, 2019.
- [8] F. Elfaladonna and A. Rahmadani, "Analisa Metode Classification-Decision Tree dan Algoritma C. 45 untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner," *SINTECH (Science Inf. Technol. J.)*, vol. 2, no. 1, pp. 10–17, 2019.
- [9] S. Haryati, A. Sudarsono, and E. Suryana, "implementasi data mining untuk memprediksi masa studi mahasiswa menggunakan algoritma c4. 5 (studi kasus: universitas dehasen bengkulu)," *J. Media*

- Infotama*, vol. 11, no. 2, 2015.
- [10] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, and A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil: Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 68–75, 2022.
- [11] S. I. Ayon and M. M. Islam, "Diabetes prediction: a deep learning approach," *Int. J. Inf. Eng. Electron. Bus.*, vol. 12, no. 2, p. 21, 2019.
- [12] Y. Cheng-Hong *et al.*, "Prediction of mortality in the hemodialysis patient with diabetes using support vector machine," *Rev. Argentina Clínica Psicológica*, vol. 29, no. 4, p. 219, 2020.
- [13] A. Dagliati *et al.*, "Machine learning methods to predict diabetes complications," *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 295–302, 2018.
- [14] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.
- [15] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, and A. Pangestu, "Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 7, no. 4, pp. 1265–1270, 2022.
- [16] M. D. Purbolaksono, M. I. Tantowi, A. I. Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021.
- [17] R. Rusbandi, F. A. Ma'ruf, A. Pratama, I. Sholihin, and A. R. Rinaldi, "Penerapan Model Prediksi Menggunakan Algoritma C. 45 Untuk Prediksi Kelulusan Siswa SMK Wahidin," *J. DATA Sci. Inform.*, vol. 1, no. 1, pp. 16–20, 2021.
- [18] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization," *Procedia Eng.*, vol. 53, pp. 453–462, 2013.
- [19] J.-S. Chou, M.-Y. Cheng, Y.-W. Wu, and A.-D. Pham, "Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification," *Expert Syst. Appl.*, vol. 41, no. 8, pp. 3955–3964, 2014.
- [20] M. Masrurroh, "Perbandingan Metode Regresi Linear Dan Neural Network Backpropagation Dalam Prediksi Nilai Ujian Nasional Siswa Smp Menggunakan Software R," *Joutica*, vol. 5, no. 1, pp. 331–336, 2020.
- [21] D. F. Sengkey, F. D. Kambey, S. P. Lengkong, S. R. Joshua, and H. V. F. Kainde, "Pemanfaatan Platform Pemrograman Daring dalam Pembelajaran Probabilitas dan Statistika di Masa Pandemi COVID-19," *J. Tek. Inform.*, vol. 15, no. 4, pp. 257–264, 2020.
- [22] S. H. Situmorang, I. Muda, M. Doli, and F. S. Fadli, *Analisis data untuk riset manajemen dan bisnis*. USUpress, 2010.
- [23] B. Hermanto, "Sistem Informasi Manajemen Keuangan pada PT. Hulu Balang Mandiri Menggunakan Framework Laravel," *J. Komputasi*, vol. 7, no. 1, pp. 17–26, 2019.
- [24] A. Darmawan, N. Kustian, and W. Rahayu, "Implementasi data mining menggunakan model svm untuk prediksi kepuasan pengunjung taman tabebuaya," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.)*, vol. 2, no. 3, pp. 299–307, 2018.
- [25] S. Suryani and M. Mustakim, "Estimasi Keberhasilan Siswa dalam Pemodelan Data Berbasis Learning Menggunakan Algoritma Support Vector Machine," *Bull. Informatics Data Sci.*, vol. 1, no. 2, pp. 81–88, 2022.
- [26] P. Mayadewi and E. Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," *SESINDO 2015*, vol. 2015, 2015.
- [27] T. M. Rahayu, B. A. Ningsi, I. Isnurani, and I. Arofah, "KLASIFIKASI KETEPATAN WAKTU KELULUSAN MAHASISWA DENGAN METODE NAÏVE BAYES," *MEDIA BINA Ilm.*, vol. 15, no. 8, pp. 4993–5000, 2021.
- [28] B. Novianti, T. Rismawan, and S. Bahri, "Implementasi Data Mining Dengan Algoritma C4. 5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak)," *Coding J. Komput. dan Apl.*, vol. 4, no. 3, 2016.
- [29] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi penyakit gigi dan mulut menggunakan metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, p. 964X, 2018.
- [30] T. I. Andini, W. Witanti, and F. Renaldi, "Prediksi Potensi Pemasaran Produk Baru dengan Metode Naïve Bayes Classifier dan Regresi Linear," 2016.
- [31] A. A. Suryanto and A. Muqtadir, "Penerapan metode mean absolute error (MEA) dalam algoritma regresi linear untuk prediksi produksi padi," *Saintekbu*, vol. 11, no. 1, pp. 78–83, 2019.