Indonesian Journal on Software Engineering (IJSE)
Vol. 7, No. 1, Juni 2021, hlm. 63-68 p-ISSN: 2461-0690
Akreditasi KEMENRISTEKDIKTI, No. 85/M/KPT/2020 e-ISSN: 2714-9935

Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes

Nurlaelatul Maulidah[*1], Riki Supriyadi[2], Dwi Yuni Utami[3], Fuad Nur Hasan[4], Ahmad Fauzi[5], Ade Christian[6]

[1,3,4,5]Universitas Bina Sarana Informatika, [2,6]Universitas Nusa Mandiri
Email: [1]nurlaelatul.nlt@bsi.ac.id, [2]riki.rsd@nusamandiri.ac.id, [3]dwi.dyu@bsi.ac.id, [4]fuad.fnu@bsi.ac.id, [5]ahmad.fzx@bsi.ac.id, [6]ade.adc@nusamandiri.ac.id

Abstrak

Diabetes melitus adalah penyakit metabolik yang ditandai terjadinya kenaikan gula darah yang disebabkan oleh terganggunya hormon insulin yang memiliki fungsi sebagai hormon dalam menjaga homeostatis tubuh menggunakan cara penurunan kadar gula darah (American Diabetes Association, 2017). World Health Organization (WHO) memperkirakan jumlah penderita diabetes melitus orang dewasa diatas 18 tahun dalam tahun 2014 berjumlah 422 juta (WHO, 2016:25). Prevalensi diabetes melitus Asia Tenggara sudah berkembang dalam tahun 1980 sebanyak 4,1% dan tahun 2014 menjadi sebanyak 8,6%. Menurut Riset Kementerian Kesehatan pada tahun 2018, Prevalensi diabetes Indonesia sebanyak 2,0%, sedangkan di Provinsi Jawa Timur sebanyak 2,6% pada penduduk umur diatas 15 tahun (KEMENKES RI, 2019). Penelitian ini dikembangkan melalui pengolahan data sekunder database kesehatan Dataset Diabetes yang diambil dari dataset Kaggle dan dapat diakses melalui https://www.kaggle.com/johndasilva/diabetes. Dimana datanya sendiri terdiri dari 2000 *record* dengan beberapa variabel prediktor medik (*Pregnancies/*Kehamilan*, Glucose/*Glukosa*, BloodPressure/*Tekanan Darah*, SkinThickness/*Ketebalan Kulit*, Insulin, BMI/*Indeks Masa Tubuh*, DiabetesPedigreeFunction/*Keturunan*, Age/*Umur *and Outcome/*Hasil). Kemudian data tersebut akan diolah dengan menggunakan metode *Support Vector Machine* dan metode *Naive Bayes* untuk mengetahui akurasi hasil diagnosa diabetes. Berdasarkan hasil dari penelitian yang sudah dilakukan metode *Support Vector Machine* memiliki nilai akurasi yang jauh lebih tinggi dibandingkan dengan menggunakan metode *Naive Bayes*. Nilai akurasi untuk model metode *Support Vector Machine* adalah 78,04% dan nilai akurasi untuk metode *Naive Bayes* 76,98%. Berdasarkan nilai ini, perbedaan akurasinya adalah 1,06%. Sehingga dapat disimpulkan bahwa penerapan metode *Support Vector Machine* mampu menghasilkan tingkat akurasi diagnosis diabetes yang lebih baik dibandingkan dengan menggunakan metode *Naive Bayes*.

Kata kunci: *Diabetes, Metode Support Vector Machine, Metode Naive Bayes*

Abstract

Diabetes mellitus is a metabolic disease characterized by an increase in blood sugar caused by disruption of the hormone insulin, which functions as a hormone in maintaining the body's homeostasis by decreasing blood sugar levels (American Diabetes Association, 2017). The World Health Organization (WHO) estimates that the number of adults with diabetes mellitus over 18 years in 2014 will be 422 million (WHO, 2016: 25). The prevalence of diabetes mellitus in Southeast Asia has grown in 1980 as much as 4.1% and in 2014 to as much as 8.6%. According to Ministry of Health Research in 2018, the prevalence of diabetes in Indonesia was 2.0%, while in East Java Province it was 2.6% among people over 15 years of age (KEMENKES RI, 2019). This research was developed through secondary data processing of the Diabetes Dataset health database which was taken from the Kaggle dataset and can be accessed through https://www.kaggle.com/johndasilva/diabetes. Where the data itself consists of 2000 records with several medical predictor variables (Pregnancies / Pregnancy, Glucose / Glucose, Blood Pressure / Blood Pressure, Skin Thickness / Skin Thickness, Insulin, BMI / Body Mass Index, Diabetes Pedigree Function / Heredity, Age / Age and Outcome / Results)). Then the data will be processed using the Support Vector Machine method and the Naive Bayes method to determine the accuracy of diabetes diagnosis results. Based on the results of research that has been done, the Support Vector Machine method has a much higher accuracy value than using the Naive Bayes method. The accuracy value for the Support Vector

Machine method model is 78.04% and the accuracy value for the Naive Bayes method is 76.98%. Based on this value, the difference in accuracy is 1.06%. So it can be concluded that the application of the Support Vector Machine method is able to produce a better level of diabetes diagnosis accuracy than using the Naive Bayes method.

Keywords: *Diabetes, Support Vector Machine Method, Naive Bayes Method*

1. PENDAHULUAN

Diabetes melitus adalah penyakit metabolik yang ditandai terjadinya kenaikan gula darah yang disebabkan oleh terganggunya hormon insulin yang memiliki fungsi sebagai hormon dalam menjaga homeostatis tubuh menggunakan cara penurunan kadar gula darah (American Diabetes Association, 2017).

World Health Organization (WHO) memperkirakan jumlah penderita diabetes melitus orang dewasa diatas 18 tahun dalam tahun 2014 berjumlah 422 juta (WHO, 2016:25). Prevalensi diabetes melitus Asia Tenggara sudah berkembang dalam tahun 1980 sebanyak 4,1% dan tahun 2014 menjadi sebanyak 8,6%. Menurut Riset Kementerian Kesehatan pada tahun 2018, Prevalensi diabetes Indonesia sebanyak 2,0%, sedangkan di Provinsi Jawa Timur sebanyak 2,6% pada penduduk umur diatas 15 tahun (KEMENKES RI, 2019). Penyakit diabetes atau sering disebut dengan penyakit kencing manis ini disebut juga sebagai *sillent killer*. Hal ini memicu peningkatan jumlah penderita diabetes melitus setiap tahun. diabetes melitus memiliki potensi merusak tubuh secara perlahan, sehingga apabila tidak segera mendapat penanganan dapat menimbulkan komplikasi/penyakit lain. Penderita diabetes melitus 2 kali lebih berisiko terkena penyakit kardiovaskular dan sekitar 75% diabetes melitus menyebabkan kematian.Maka sebab itu padapenelitian inimencobamenerapkansuatumetodeklasifikasidalammemprediksiapakah seseorang mengidap penyakit diabetes atau tidak. untuk penelitian prediksi penyakit diabetes sudah beberapa kali dilakukan, penelitian sebelumnya berguna bagi penulis untukselanjutnya dijadikan pedoman serta pegangan penelitian yang akan penulis lakukan sehingga nantinya dengan adanya penelitian sebelumnya dapat membantu dan memudahkanpenulisdalam melakukan penelitiannya sesuai dengan tema dan membuat sistem yang baru dan bermanfaat. untuk penelitian terkait tentang penyakit diabetes yang dijadikan referensi utama oleh penulis adalah penelitian dari Noviandi yang membuat penelitian dengan judul Implementasi Algoritma *Decision Tree* C4.5 Untuk Prediksi Penyakit Diabetes. Model prediksi dibentuk dengan menggunakan data *Pima Indians Diabetes Databases* (PPID) yang bersumber dari UCI *Machine Learning Repository*. Model prediksi dengan metode *decision tree* C4.5 memiliki akurasi 70.32 persen dengan hasil 9 *rule*, denganjumlah *class* sebanyak4*rule* dan5*rule class* untuk melakukan prediksi penyakit DM (Noviandi, 2018).

Pada penelitian sebelumnya yang sudah dlakukan oleh (Dewi Rahma Ente et al., 2020) dengan judul "Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5". Penelitian ini dilakukan dengan tujuan mengidentifikasi serta melihat hubungan antara faktor-faktor yang mempengaruhi penyakit DM. Metode yang digunakan dalam penelitian ini yaitu algoritma C4.5. Hasil yang diperoleh menunjukkan ada empat faktor yang mempengaruhi prediksi status pasien DM yaitu *Fasting Blood Glucose*, LDL Kolesterol, Trigliserida, dan Berat Badan. Sedangkan pada penelitian yang sudah dilakukan oleh (Faizal Aris dan Benyamin, 2019) yang berjudul "Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi". Penelitian ini menggunakan perhitungan metode C.45 dengan menggunakan aplikasi *rapid miner*. Pada pengujian penelitian ini menggunakan beberapa atribut klasifikasi yaitu atribut Jenis Kelamin, berat badan,Usia, Perokok, kadar gula darah, dan Tipe penyakit diabetes.

2. PENDAHULUAN

2.1. *Dataset*

Dataset diabetes yang digunakan adalah data yang diperoleh dari database kesehatan Diabetes *Dataset* yang dapat diakses melaluihttps://www.kaggle.com/johndasilva/diabetes. Dimana datanyasendiri terdiri dari 2000 *record* dengan beberapa variabel atau atribut prediktor medis *(Pregnancies*/Kehamilan, Glucose/Glukosa, BloodPressure/Tekanan Darah, Skin*Th*ickness/Ketebalan Kulit, Insulin, BMI/Indeks Masa Tubuh, DiabetesPedigree*Function*/Keturunan*, Age*/Umur dan *Outcome*/Hasil*),*yangkemudiandiolah dengan menggunakan tools Python*.*

2.2. Diabetes

Diabetes adalah penyakit kronis serius yang terjadi karena pankreas tidak menghasilkan cukup insulin (hormon yang mengatur gula darah atau glukosa), atau ketika tubuhtidak dapatsecaraefektifmenggunakan

insulin yang dihasilkannya. WHO memberitahukan bahwa angka kejadian penyakit yang tidak menular pada tahun 2004 mencapai 48,30 persen sedikit lebih tinggi dari angka kejadian penyakit yang menular, yaitu sebanyak 47,50 persen(KEMENKES RI, 2019).

Toharin, et al memberitahukan bahwa Diabetes disebut juga *mother of disease* karena dapat menjadi penyebab munculnya penyakit komplikasi seperti hipertensi, penyakit jantung dan pembuluh darah, stroke, gagal ginjal, dan kebutaan, serta kerusakan jangka panjang meliputi gangguan dan kegagalan fungsi berbagai organ terutama mata, ginjal, syaraf, jantung, dan pembuluh darah (Kantono et al,. 2019).

2.3. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) pertama kali dikenalkan oleh Vapnik pada tahun 1992 sebagai salah satu metode *machine learning* yang bekerja dengan prinsip *Structural Risk Minimization*/SRM yang bertujuan untuk menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*. Metode ini menggunakan hipotesis berupa fungsi linier dalam sebuah ruang fitur yang berdimensi tinggi, dengan mengimplementasikan *learning* bisa yang berasal dari teori pembelajaran statistik(Parapat et al,. 2018).

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi (Darmayanti et al,. 2018). *Support Vector Machine* (SVM) memiliki prinsip dasar *linier classifier* yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun *Support Vector Machine* (SVM) telah dikembangkan agar dapat bekerja pada *problem non- linier* dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan jarak (*margin*) antara kelas data.

2.4. *Naive Bayes*

Algoritma *Naive Bayes* merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. *Naive Bayes* yaitu sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan (Saleh, 2015).

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan *Naive* dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya (Bustami, 2014).

Menurut Bustami dalam Saleh (2015:209) persamaan dari teorema *Bayes* adalah:

$$\mathrm{P(H\backslash X)} = \frac{P(X\backslash H).P(H)}{P(X)} \quad (1)$$

Dimana:

X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class yang spesifik
P(H\X) : Probabilistik hipotesis H berdasar kondisi X (posteriori probabilistik)
P(H) : Probabilistik hipotesis H (prior probabilitas)
P(X\H) : Probabilistik hipotesis X berdasar kondisi pada hipotesis H
P(X) : Probabilitas X

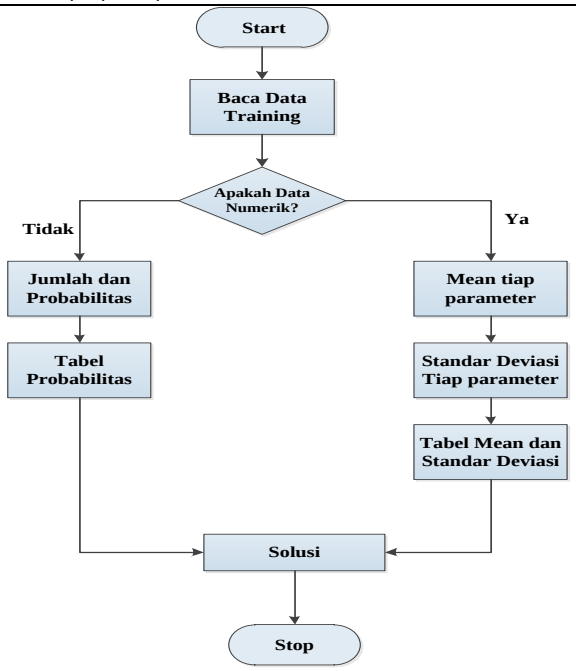
Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naive Bayes* di atas disesuaikan sebagai berikut:

$$P(C\backslash F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn} \quad (2)$$

Dimana variable *C* merepresentasikan kelas, sementara variable *F1…Fn* merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sample berkarakteristik tertentu kedalam kelas *C* (*Posterior*) adalah peluang munculnya kelas *C* (sebelum masuk sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas *C* (disebut juga *likelihood*), dibagi dengan kemunculan karakteristik-karakteristik sampel secara global (disebut juga *evidence*). Karena itu rumus diatas dapat ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (3)$$

Berikut adalah alur dari metode *Naive Bayes* menurut Saleh (2015:211):



Sumber: Saleh (2015:211)

Gambar 1. Alur Metode *Naive Baiyes*

Adapun keterangan dari gambar diatas sebagai berikut:

1. Baca *data training*
2. Hitung jumlah dan probabilitas, namun apabila data numerik maka:
   a. Cari nilai *mean* dan standar deviasi dari masing-masing parameter yang merupakan data numerik. Adapun persamaan yang digunakan untuk menghitung nilai rata-rata hitung (*mean*) dapat dilihat sebagai berikut:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (4)$$

   Atau

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \quad (5)$$

   di mana:
   μ : rata-rata hitung (*mean*)
   x_i : nilai sample ke i
   n : jumlah sampel
   Dan persamaann untuk menghitung simpangan baku (standar deviasi) dapat dilihat sebagai berikut:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \quad (6)$$

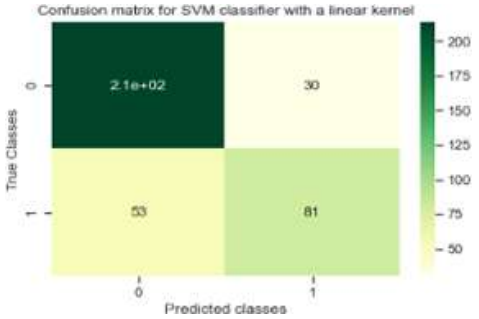
   Dimana:
   σ : standar deviasi
   x_i: nilai *x* ke –*i*
   μ : rata-rata hitung
   n : jumlah sampel
   b. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.
3. Mendapatkan nilai dari tabel *mean*, standar deviasi dan probabilitas.
4. Solusi kemudian dihasilkan.

3. Hasil dan Pembahasan

3.1. Pengujian

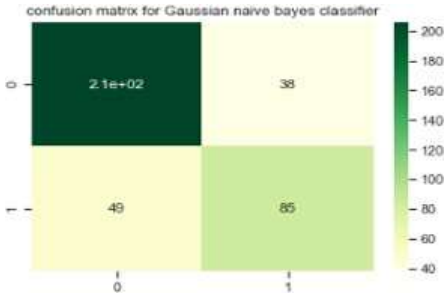
Pengujian ini bertujuan untuk mencari nilai akurasiserta kinerja mana yang lebih baik diantara metode *Support Vector Machine* dan *Naive Bayes* dalam mengklasifikasikan data ke dalam kelas yang telah ditentukan. Pada uji coba ini, tahapan untuk mendapatkan nilai akurasi yaitu eksplorasi data, *invalid data*, Memvisualisasikan kolom yang berbeda pada kelas dan Pemilihan Data dan Model *Fitting*.

3.2. Penjelasan Confusion Matrix Algoritma *Support Vector Machine*



Gambar 1. *Confusion Matrix* Algoritma *Support Vector Machine*

Untuk nilai *Confusion Matrix* yang dihasilkan dari metode atau algoritma *Support Vector Machine*, yaitu: *True* Positif bernilai 2 1e+02 dimana untuk kelas *Outcome*, *True* Negatif bernilai81dimanauntukkelas *Outcome*,*False* Positifbernilai 30dimanauntukkelas*Outcome*,dan*False* Negatifbernilai53 dimana untuk kelas*Outcome*.



Gambar 2. *Confusion Matrix* Algoritma *Naive Bayes*

Untuk nilai*Confusion Matrix* yang dihasilkan dari metode atau algoritma *Naive Bayes*, yaitu: *True* Positif bernilai 2 1e+02 dimana untuk kelas *Outcome*, *True* Negatif bernilai 85 dimana untuk kelas *Outcome*, *False* Positifbernilai38dimana untuk kelas *Outcome* dan *False* Negatif bernilai 49 dimana untuk kelas *Outcome*.

3.3. Penjelasan Hasil Algoritma *Support Vector Machine* dan Algoritma *Naive Bayes*

Berdasarkan dari penelitian yang sudah dilakukan untuk menentukan nilai akurasi dalam memprediksi penyakit diabetes, didapatkan hasil sebagai berikut:

Tabel 1. Tabel Perbandingan Hasil Akurasi

Algoritma	Hasil Akurasi
Support Vector Machine	78,04 %
Naive Bayes	76,98 %

Sumber: (Maulidah et al, 2021)

Dari hasil pengujian, dengan dilakukan evaluasi baikse- cara *confusion matrix* maupun nilai akurasi terbukti bahwa pengujian yang dilakukan dengan algoritma *Support Vector Machine* memilikinilaiakurasi yang lebih tinggi dibandingkan dengan menggunakan algoritma *Naive Bayes*. Nilai akurasi untuk model algoritma*Support Vector Machine* sebesar78,04 %dannilai akurasi algoritma*Naive Bayes* sebesar76,98%. Berdasarkan nilai tersebut diperoleh selisih akurasi sebesar 1,06%.

4. KESIMPULAN

Dari hasil penelitian yang telah dilakukan algoritma *Sup- port Vector Machine* memiliki nilai akurasi yang lebihtinggi dibandingkan dengan menggunakan algoritma *Naive Bayes*. Nilai akurasi untuk model algoritma *Support Vector Machine* sebesar 78,04 persen dan nilai akurasi algoritma *Naive Bayes* sebesar 76,98 persen. Berdasarkan nilai tersebut diperoleh selisih akurasi sebesar 1,06 persen. Sehingga dapat disim- pulkan bahwa penerapan algoritma *Support Vector Machine* mampu menghasilkan tingkat akurasi diagnosis penyakit dia- betes yang lebih baik dibandingkan menggunakan algoritma yaitu algoritma *Naive Bayes*.

REFERENSI

A. Kantono, I. Y. Purbasari, and F. T. Anggraeny. “Penerapan pruning pada algoritma c5.0 untuk mendiagnosis penyakit diabetes melitus 1”. no. September, pp. 184–189, 2019.

American Diabetes Association. “Standards of Medical Care in Diabetes 2017”. Vol. 40. USA: ADA, 2017.

Aris, Faizal dan Benyamin. “Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi”. Vol 1 No 1. Hal 01-06. Desember, 2019.

Bustami. “Penerapan Algoritma Naive Bayes Untuk Nasabah Asuransi”. J. Inform., vol. 8, no. 1, pp.884–898, 2014.

Ente, Dewi Rahma, Sri Astuti Thamrin, Hedi Kuswanto, Samsul Arifin Dan Andreza. “Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5”. Vol 4 No 1. 80-88, 2020.

E. Y. Darmayanti, D. S. Budi, and A. B. Fitra. “Particle Swarm Optimization Untuk Optimasi Bobot Extreme Learning Machine Dalam Memprediksi Produksi Gula Kristal Putih Pabrik Gula”. J. Pengemb. Teknol. Inf. dan IlmuKomput., vol. 2, no. 11, pp. 5096–5104, 2018.

KEMENKES RI. “Hari Diabetes Sedunia Tahun 2018”. Pus. Data dan Inf. Kementrian Kesehat. RI, pp. 1–8, 2019

Noviandi. ”Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes”. J. Inohim. Vol. 6, No.1. 2018.

Parapat, Furqon dan Sutrisno. ”Penerapan Metode Sup- port Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak”. J. PTIIK. Vol. 2, No.10. 2018.

Saleh, A. ”Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Besarnya Penggunaan LIstrik Rumah Tangga”. CItec Jurnal, p. 2. 2015.