Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol.3, No.1 Februari 2024
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 77-83
DOI: https://doi.org/10.55606/juprit.v3i1.3151

Klasifikasi Penyakit Diabetes Menggunakan Metode SVM Dan KNN

Aswin Ardiansyah*[1], Enos C.O. Telaumbanua [2], Aron S. Gultom[3], Angelita A. S. M. Limbong[3]
[1,2,3,4] Universitas Negeri Medan
*E-mail: *[1] aswinardiansyah06@gmail.com , [2] enoztel@gmail.com ,*
[3] aron_santoso@mhs.unimed.ac.id , [4] angelitalimbong03@gmail.com

Abstract *Diabetes is a disease caused by high blood sugar levels and impaired insulin production in the body. Although it is not a contagious disease, in fact, many Indonesians suffer from diabetes. In fact, according to the North Sumatra Health Department, the prevalence of diabetes in Indonesia is estimated to reach 21.3 million people by 2030. As technology develops, machine learning has helped many health practitioners in dealing with diabetes, one of which is modeling with SVM and KNN. The application of this algorithm aims to create a model that is able to classify diabetes in patients based on data of diabetes factors such as age, weight, blood pressure, blood sugar levels, etc. The model that has been built is then evaluated for its performance with a confusion matrix, with the evaluation results of the SVM model being better than KNN with an accuracy of 100% for the SVM model and an accuracy of 96% for the KNN model.*

Keywords : *Classification, Diabetes, SVM, KNN, Confusion Matrix*

Abstrak Diabetes merupakan penyakit yang disebabkan oleh tingginya kadar gula dan terhambatnya produksi insulin pada tubuh. Meskipun bukan penyakit menular, namun faktanya banyak sekali masyarakat Indonesia menderita penyakit diabetes. Bahkan menurut dinas kesehatan Sumatera Utara, prevalensi diabetes di Indonesia diperkirakan akan mencapai 21,3 juta orang pada tahun 2030. Seiring perkembangan teknologi, machine learning banyak membantu para praktisi kesehatan dalam menangani penyakit diabetes ini, salah satunya adalah pemodelan dengan SVM dan KNN. Penerapan algoritma ini bertujuan untuk membuat model yang mampu mengklasifikasikan penyakit diabetes pada pasien berdasarkan data faktor-faktor penyakit diabetes seperti usia, berat badan, tekanan darah, kadar gula, dll. Model yang telah dibangun selanjutkan dievaluasi performanya dengan confusion matrix, dengan hasil evaluasi model SVM lebih baik dibanding KNN dengan akurasi sebesar 100% untuk model SVM dan akurasi sebesar 96% untuk model KNN.

Kata Kunci : Klasifikasi, Diabetes, SVM, KNN, Confusion Matrix

PENDAHULUAN

Diabetes adalah kondisi kekurangan hormon insulin yang menyebabkan kadar gula darah (glukosa) tinggi secara kronis. Insulin adalah hormon yang membantu penyerapan glukosa dalam sel-sel tubuh untuk diubah menjadi energi. Insulin juga membantu menyimpan sebagian glukosa sebagai cadangan energi. Kondisi ini juga dikenal sebagai penyakit gula atau kencing manis. Gula dalam darah seharusnya diserap oleh sel-sel tubuh untuk diubah menjadi energi [1]. Diabetes adalah penyakit kronis yang tidak menular, namun memiliki tingkat kematian yang tinggi. Penyakit ini dapat menyebabkan berbagai komplikasi, yang dapat menurunkan produktivitas pasien dan kualitas hidupnya. Menurut dinas kesehatan Sumatera Utara, prevalensi diabetes di Indonesia diperkirakan akan mencapai 21,3 juta orang pada tahun 2030 [2].

Menurut data dari Federasi Diabetes Internasional (IDF), lebih dari 90% kasus diabetes disebabkan oleh gangguan sekresi insulin dan/atau sensitivitas insulin [3]. Diabetes

Received: Oktober 29, 2023; Accepted: Desember 16, 2023; Published: February 28, 2024
* Aswin Ardiansyah, *aswinardiansyah06@gmail.com*

dapat disebabkan oleh berbagai faktor seperti usia, tekanan darah tinggi, kadar gula darah tinggi, obesitas, riwayat keluarga, kadar insulin, dan pola makan. Faktor-faktor tersebut akan digunakan dalam penelitian ini untuk membangun sistem kalsifiaksi yang dapat memprediksi diabetes. Tingginya prevalensi diabetes di Indonesia, yang merupakan negara berkembang dengan populasi yang besar, membuat sulit bagi masyarakat untuk berkonsultasi dengan tenaga medis untuk pemeriksaan.

Kemajuan teknologi memungkinkan dokter spesialis untuk memperoleh data yang valid dari rekam medis dan uji laboratorium. Data tersebut kemudian digunakan untuk mendiagnosis apakah pasien mengidap diabetes. Model matematis dapat dibentuk berdasarkan data tersebut, yaitu dengan menggunakan metode klasifikasi [4]. K-Nearest Neighbors (KNN) adalah algoritma klasifikasi yang bekerja dengan cara mencari data pembelajaran yang paling mirip dengan data baru, kemudian mengklasifikasikan data baru tersebut berdasarkan label dari data pembelajaran yang paling mirip tersebut [5]. Sedangkan Support Vector Machine (SVM) adalah algoritma klasifikasi yang menghasilkan hasil klasifikasi dengan akurasi tertinggi dengan cara menemukan garis pemisah optimal antara dua kelas data. Garis pemisah optimal adalah garis yang memiliki margin terluas antara data dari dua kelas [6].

Penelitian terkait yang dilakukan oleh Nita [7] pada penelitian klasifikasi penyakit diabetes dengan algoritma SVM dan *Logistic Regression* didapatkan hasil pada pemodelan dengan SVm menghasilkan nilai akurasi sebesar 88,77% dengan teknik SMOTE yang menunjukkan bahwa performa model sangat baik dalam melakukan klasifikasi. Pada penelitian lainnya oleh Refa [8] dengan klafikasi penyakit diabetes dengan Logistic Regression menggunakan faktor faktor penyakit seperti tes toleransi glukosa, tekanan darah, ketebalan lipatan kulit, BMI, riwayat keturunan hingga usia menjadi parameter dalam pemodelan ini. Hasil akurasi yang didapatkan dari penelitian ini yaitu sebesar 76%. Lalu menurut Galih [5] pada penelitiannya mengenai identifikasi faktor faktor terjadinya diabetes dengan KNN, menunjukkan bahwa faktor yang berpengaruh secara signifikan pada penyakit diabetes adalah usia yaitu pada rentang usia 20 – 40 tahun dengan akurasi terbaik terdapat pada K=19 sebesar 76%.

Penelitian ini menggunakan SVM dan KNN untuk menghasilkan model prediksi yang bertujuan untuk mengklasifikasikan seseorang sebagai penderita diabetes atau tidak penderita diabetes. Performa model tersebut kemudian diukur untuk mengetahui sejauh mana kemampuannya dalam mengklasifikasi penyakit diabetes.

METODE PENELITIAN

Pengumpulan Data

Tahap pengumpulan data adalah tahap dimana data diperoleh dari berbagai cara, seperti survey, wawancara, *data mining*, dsb [9]. Data yang diperoleh nantinya akan dikumpulkan menjadi sebuah dataset yang nantinya akan diolah untuk tahap pra-pemrosesan data.

Pra-pemrosesan Data

Pra-pemrosesan bertujuan untuk mempersiapkan data agar mudah untuk diproses. Pada tahap ini, data akan dieksplorasi dengan cara melakukan *preprocessing* dengan memeriksa data duplikat, *missing value,* dan *one-hot-encoding* pada data kategorikal. Missing value dapat terlihat dari nilai 0 dalam atribut yang seharusnya tidak boleh memiliki nilai 0 [10].

Pemodelan dan Evaluasi Performa

Proses pemodelan adalah proses pembentukan model klasifikasi dilakukan dengan menerapkan algoritma atau metode yang diinginkan. Tahap pemodelan juga bertujuan untuk meneentukan parameter yang ingin diatur sehingga mampu memaksimalkan kinerja model [11]. Pada penelitian ini, algoritma yang akan diimplementasikan adalah SVM dan KNN. Setelah implementasi algoritma selesai, selanjutnya model akan dievaluasi dengan confusion matrix untuk mengetahui performa tolak ukur model yang telah dibuat [12].

HASIL DAN PEMBAHASAN

Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data faktor-faktor penyebab penyakit diabetes yang berasal dari *platform* Kaggle dalam format .csv. Berikut adalah karakteristik atribut dataset yang digunakan pada penelitian ini.

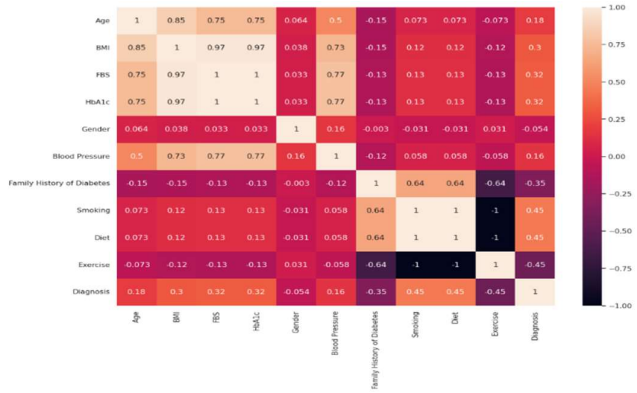
Tabel 1. Karakterisitik Dataset

Atribut	Deskripsi Atribut	Jenis Atribut
Age	Usia Pasien	*Numeric*
Gender	Jenis Kelamin Pasien	*String*
BMI	Indeks Massa Tubuh Pasien	*Numeric*
Blood Pressure	Tekanan Darah Pasien	*String*
FBS	Puasa Glukosa Darah	*Numeric*
HbA1c	Hemoglobin Pasien	*Numeric*
Family History of Diabetes	Keturunan Keluarga	*Boolean*
Smoking	Perokok	*Boolean*
Diet	Pola Makan	*Boolean*

Exercise	Olahraga	*Object*
Diagnosis	Diagnosa	*Boolean*

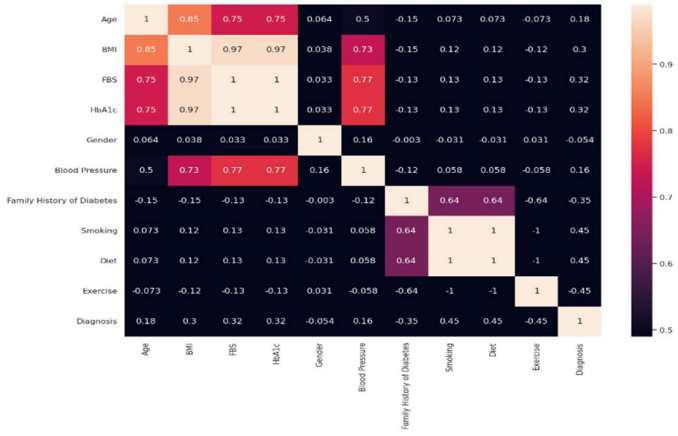
Pra-pemrosesan Data

Pada tahap ini data akan di bersihkan dan dirapikan seperti menghapus data duplikat, mengecek *missing value* hingga melakukan *one-hot-encoding* pada data yang bersifat kategorikal. Tujuan dilakukannya *encoding* adalah untuk memudahkan model dalam mengolah data dalam bentuk numerik [13]. Lalu setelah melakukan pra-pemrosesan data, selanjutnya melihat korelasi masing-masing faktor pada data. Hasil korelasi dapat dilihat pada grafik berikut.



Gambar 1. *Heatmap* korelasi faktor

Pada grafik *heatmap* diatas, terlihat bahwa persebaran korelasi antar faktor sangat beragam, dan seperti sulit untuk mengidentifikasi faktor mana saja yang sangat berkolerasi. Maka dari itu diperlukan *filtering* pada nilai korelasi yang mana pada penelitian ini, peneliti menentukan nilai korelasi antara 0,49 – 0,99. Berikut adalah grafik *heatmap* setelah dilakukannya *filtering*.



Gambar 2. *Heatmap* setelah *filtering*

Terlihat jelas faktor-faktor yang masuk kedalam filter korelasi yang baik, yaitu pada faktor usia, BMI, FBS, hBA1c, tekanan darah. Maka dari itu faktor-faktor tersebut akan digunakan pada penelitian kali ini. Pada kasus faktor perokok dan pola makan pada penelitian ini tidak akan dipertimbangkan karena tidak berkolerasi satu sama lain.

Pemodelan dan Evaluasi Performa

Pemodelan pada penelitian ini akan menggunakan algoritma SVM dan KNN, pemodelan dimulai dari pembagian dataset pada model SVM dengan porsi 80% data *training* dan 20% data *test*, lalu penentuan nilai K pada model KNN sebanyak 6 iterasi. Penentuan porsi data sangat penting mengingat model harus mempelajari data *training* yang banyak agar model mampu mengklasifikasikan data *test* dengan tepat nantinya [14]. Penentuan jumlah K juga tak kalah penting pada pemodelan KNN agar model dapat menghasilkan klaster yang baik [15].

Setelah model dilatih, selanjutnya yaitu pengujian model dan evaluasi performa. Berikut adalah tabel perbandingan evaluasi performa algoritma SVM dan KNN.

Tabel 2. Perbandingan evaluasi performa SVM dan KNN

	Akurasi	**Presisi**	***Recall***	***F1-Score***
SVM	100%	100%	100%	100%
KNN	96%	100%	86%	92%

Berdasarkan tabel diatas, dapat disimpulkan bahwa hasil pemodelan dengan algoritma SVM lebih baik dibandingkan dengan algoritma KNN dalam mengklasifikasikan penyakit diabetes dengan akurasi yang sangat baik pada algoritma SVM. Hal ini dapat terjadi mengingat dataset merupakan data kategorikal yang mana sangat efektif bagi SVM mengklasifikasikan data tersebut. Faktor lainnya adalah jumlah data yang tidak terlalu banyak menyebabkan model mampu mengenali data dengan sangat baik tanpa ada error sedikitpun pada model.

KESIMPULAN

Penelitian ini memanfaatkan data faktor penyebab diabetes yang selanjutnya akan digunakan untuk memprediksi penyakit diabetes pada pasien. Faktor-faktor tersebut meliputi usia, jenis kelamin, tekanan darah, berat badan, kadar gula,dll. Model prediksi dibangun dengan menggunakan algoritma SVM dan KNN dan didapatkan hasil pemodelan dengan evaluasi performa SVM sebesar 100% akurasi, 100% presisi, 100% *recall* dan 100% *f1-score* sedangkan KNN sebesar 96% akurasi, 100% presisi, 86% *recall* dan 92% *f1-score.* Hal ini menunjukkan bahwa algoritma SVM bekerja lebih baik dalam mengklasifikasikan data kategorikal dibandingankan dengan algoritma KNN dalam klastering data.

SARAN

Saran untuk penelitian ini adalah parameter tuning yang lebih spesifik dalam pemodelan data serta volume data sebaiknya lebih besar untuk menghasilkan model yang lebih baik dalam mengklasifikasikan untuk jangka panjang.

DAFTAR PUSTAKA

[1] A. M. Widodo, Y. S. Anggraeni, N. Anwar, A. Ichwani, and B. A. Sekti, "Performansi K-NN, J48, Naive Bayes dan Regresi Logistik sebagai Algoritma Pengklasifikasi Diabetes," *Pros. SISFOTEK*, vol. 5, no. 1, pp. 27–33, 2021, [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=FOwZ8hUAAAAJ&pagesize=100&citation_for_view=FOwZ8hUAAAAJ:a3BOlSfXSfwC.

[2] N. K. Hasibuan, S. Dur, and I. Husein, "Faktor Penyebab Penyakit Diabetes Melitus dengan Metode Regresi Logistik," *G-Tech J. Teknol. Terap.*, vol. 6, no. 2, pp. 257–264, 2022, doi: 10.33379/gtech.v6i2.1696.

[3] International of Diabetic Ferderation, "IDF Diabetes Atlas. Eighth Edition.," *International of Diabetic Ferderation*, 2017. www.idf.org/diabetesatlas (accessed Nov. 25, 2023).

[4] S. Innassuraiya, T. Widiharih, and I. T. Utami, "ANALISIS KLASIFIKASI MENGGUNAKAN METODE REGRESI LOGISTIK BINER DAN BOOTSTRAP AGGREGATING CLASSIFICATION AND REGRESSION TREES (BAGGING CART) (Studi Kasus: Nasabah Koperasi Simpan Pinjam Dan Pembiayaan Syariah (KSPPS))," *J. Gaussian*, vol. 11, no. 2, pp. 183–194, 2022, doi: 10.14710/j.gauss.v11i2.35458.

[5] G. Mahalisa and N. Arminarahmah, "Diabetes Classification Analysis Using the Euclidean Distance Method Based on the K-Nearest Neighbors Algorithm," *JTKSI (Jurnal Teknol. Komput. dan Sist. Informasi)*, vol. 5, no. 3, p. 178, 2022, doi: 10.56327/jtksi.v5i3.1249.

[6] D. Septhya *et al.*, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.

[7] N. Fitriyani, D. R. Amalia, H. H. Handayani, A. Fitri, and N. Masruriyah, "Aplikasi Berbasis Web Berdasarkan Model Klasifikasi Algoritma SVM dan Logistic Regression Terhadap Data Diabetes," *Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 7, pp. 1762–1771, 2023, [Online]. Available: http://doi.org/10.33395/remik.v7i4.13001%0D.

[8] Q. R. Cahyani *et al.*, "Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK," *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.

[9] S. S. Dewi, R. Resmawan, and L. O. Nashar, "Analisis Regresi Logistik Multinomial dengan Metode Bayes untuk Identifikasi Faktor-Faktor Terjadinya Diabetes Melitus," *J. Math. Theory Appl.*, vol. 5, no. 2, pp. 51–60, 2023, doi: 10.31605/jomta.v5i2.2520.

[10] A. Damayunita, R. S. Fuadi, and C. Juliane, "Comparative Analysis of Naive Bayes,

K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) Algorithms for Classification of Heart Disease Patients," *J. Online Inform.*, vol. 7, no. 2, pp. 219–225, 2022, doi: 10.15575/join.v7i2.919.

[11] P. Widiarti, "Perbandingan metode regresi logistik biner dan classification and regression trees (CART) untuk klasifikasi diagnosa penyakit diabetes mellitus (DM)," no. Dm, 2020, [Online]. Available: http://digilib.uinsby.ac.id/43882/%0Ahttp://digilib.uinsby.ac.id/43882/3/Puji Widiarti_H72216062.pdf.

[12] A. P. Wicaksono, T. Badriyah, and A. Basuki, "Data Mining Studi Perbandingan Prediksi Penyakit Diabetes dengan menggunakan Logistic Regression dan Decision Trees," *J. Semnaskit*, pp. 66–69, 2015.

[13] W. Purba, Yessy, and R. N. Gulo, "Application of Data Mining To Identify Diabetes Mellitus Using the Support Vector Machine (Svm) Algorithm and Knn," *J. Infokum*, vol. 10, no. 2, pp. 994–1000, 2022, [Online]. Available: http://infor.seaninstitute.org/index.php/infokum/index.

[14] A. R. S. Darwanto, Taza Luzia Viarindita, and Yekti Widyaningsih, "Analisis Regresi Logistik Binomial dan Algoritma Random Forest pada Proses Pengklasifikasian Penyakit Ginjal Kronis," *J. Stat. dan Apl.*, vol. 5, no. 1, pp. 1–14, 2021, doi: 10.21009/jsa.05101.

[15] M. Bagoes Pakarti, "Sistem Prediksi Lama Studi Kuliah Menggunakan Metode Naive Bayes," *J. Inform. Komput. dan Bisnis*, vol. 2, no. 1, 2022, [Online]. Available: https://jurnal.itbaas.ac.id/index.php/jikombis.