Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest

Duwi Cahya Putri Buani [1 *]

[1] Informatika, Universitas Nusa Mandiri
** duwi.dcp@nusamandiri.ac.id*

Abstract *- Diabetes is a deadly chronic disease according to the Institute for Health Metrics and Evaluation diabetes is the 3rd highest mortality disease in Indonesia so research for early detection of diabetes is needed, this aims to prevent the increase of diabetes in Indonesia. In this study using Knowledge Discovery in Database Process (KDD) which is a method that can be used for data mining from data selection, data cleaning, data transformation, data mining to the evaluation stage and generating knowledge. In this study using nine models with nine algorithms tested, the algorithms are Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), K-Nearest Neighbors (KNN), Ada Boost (AB), Gradient Boosting (GB) and XGBoost Classifier (XGB), Of the nine models, the model with the Random Forest (RF) algorithm has a high accuracy rate, the accuracy value of RF is 98.78% with an AUC value of 0.98 with an AUC value of 0.98, the classification level of the model with the Random Forest (RF) algorithm is Excellent.*

Keywords: Data Mining, KDD, Random Forest

Abstract - Diabetes merupakan penyakit kronis yang mematikan menurut Institude for Health Metrics and Evaluation diabetes merupakan penyakit kematian tertinggi ke 3 di Indonesia sehingga penelitian untuk deteksi dini penyakit diabetes sangat diperlukan, hal ini bertujuan untuk mencegah meningkatnya penyakit diabetes di Indonesia. Dalam penelitian ini menggunakan *Knowledge Discovery in Database Process* (KDD) yang merupakan suatu metode yang dapat digunakan untuk data mining dari data selection, data cleaning, data transformastion, data mining sampai dengan tahapan evaluasi dan menghasilkan pengetahuan. Pada penelitian ini menggunakan sembilan model dengan sembilan algoritma yang diuji, algoritma tersebut adalah *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Decision Tree* (DT), *K-Nearest Neighbors* (KNN), *Ada Boost* (AB), *Gradient Boosting* (GB) dan *XGBoost Classifier* (XGB), dari sembilan model, model dengan algoritma *Random Forest* (RF) yang memiliki tingkat akurasi tinggi, Nilai akurasi dari RF sebesar 98,78% dengan nilai AUC sebesar 0,98 dengan nilai AUC 0,98 maka tingkat klasifikasi dari model dengan algortiman *Random Forest* (RF) *Excellent.*

Keywords: Data Mining, KDD, Random Forest

1. Introduction

Diabetes Militus merupakan suatu penyakit gangguan metabolism kronis dengan multi etiologi dengan gejalan tingginya kadar gula darah disertai dengan gangguan metabolism karbohidrat, lipid, dan protein sebagai akibat insufisiensi fungsi insulin (P2PTM, 2022).

Diabetes Militus merupakan salah satu penyakit kronis yang menyebabkan kematian tertinggi di Indonesia, Menurut data dari Institude for Health Metrics and Evaluation bahwa diabetes merupakan penyakit penyebab kematian tertinggi ke 3 di Indonesia pada tahun 2019 yaitu sekitar 57,42 kematian per 100.000 penduduk, selain itu diambil dari Data International Diabetes Federation (IDF) tercatat bahwa jumlah penderita diabetes pada 2021 di Indonesia meningkat pesat dalam sepuluh tahun terakhir. Jumlah tersebut diperkirakan dapat mencapai 28,57 juta pada tahun 2045 atau lebih besar 47% dibandingkan pada tahun 2021 jumlah penderitas diabetes sebesar 19,47 juta (DITPUI, 2023; Rizki et al., 2022).

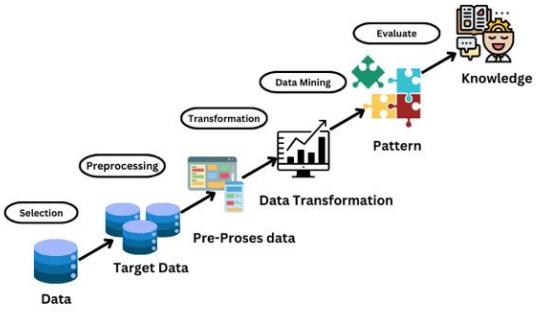
Dari uraian diatas maka perlu sekali melakukan deteksi dini penyakit diabetes mengingat diabetes merupakan salah satu penyakit kronis yang mematikan, serta Penyakit Diabetes di Indonesia semakin meningkat, salah satu cara untuk menghentikan peningkatan pasien diabetes maka dengan melakukan deteksi dini dengan melihat factor-faktor yang menyebabkan penyakit diabetes (Agustina et al., 2021).

Untuk melakukan deteksi dini penyakit diabetes dapat menggunakan data mining salah satunya adalah dengan melakukan prediksi. Data Mining adalah bagian dari tahap proses *Knowledge Discovery in Database* (KDD) Penambangan data memungkinkan untuk melakukan klasifikasi, prediksi, memperkirakan, dan mengekstrak informasi yang berguna dari data yang besar (Mardi, 2016).

Penelitian ini bukanlah penelitian yang pertama dilakukan penelitian sebelumnya dengan judul Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes dengan hasil akurasi sebesar 78,04% dan 76.98% (Maulidah et al., 2021), penelitian lain dengan judul Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes dengan hasil prediksi 70.32% (Noviandi, 2018), penelitian lainnya dengan judul Prediksi Penyakit Diabetes Dengan Naive Bayes dengan hasil akurasi sebesar 95% (Wardana & Sari, 2023). Dari beberapa penelitian sebelumnya masih belum mendapatkan akurasi dengan hasil maksimal, dan belum ada penelitian yang menggunakan Random Forest dalam penelitian ini penulis menggunakan algoritma Random forest, Random Forest adalah teknik pembelajaran mesin yang biasa digunakan untuk memecahkan masalah regresi dan klasifikasi, Random Forest merupakan kumpulan dari pohon keputusan. Algoritma ini merupakan kombinasi masing-masing pohon dari pohon keputusan yang kemudian digabungkan menjadi satu model (algorit, 2022).

2. Materials and Methods

Knowledge Discovery in Database Process (KDD) merupakan salah satu metode yang dapat digunakan untuk melakukan data mining, KDD merupakan proses yang dilakukan untuk menggali, menganalisa dan mengekstrak data sehingga menjadi informasi-informasi yang berguna (Arta et al., 2019)**.**



Sumber: (Binus, 2021)

Gambar 1. Proses Knowledge Discovery in Database Process (KDD)

Berikut ini adalah penjelasan dari gambar 1:

A. Data Original

Merupakan data asli yang diperoleh dari pemilihan data, tanpa dilakukan proses seleksi data dan pembersihan serta transformasi data (Aprianti et al., 2022) .

B. Data Selection

Data Selection merupakan tahapan melakukan identifikasi data, Pemilihan data yang relevan agar dapat dijadikan data untuk melakukan penelitian, sehingga model yang terbentuk lebih optimal (Arta et al., 2019).

C. Data Cleaning

Data cleaning merupakan proses membuang data yang duplikasi, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data, seperti kesalahan penulisan, dan memberikan nilai atau menghapus data yang kosong (Arta et al., 2019).

D. Data Transformation

Proses transformasi data menjadi data tertentu yang dapat digunakan atau dapat dibaca oleh Algoritma. Contohnya adalah merubah data dengan Type Object menjadi Numbering (Arta et al., 2019).

E. Data Mining

Data Mining adalah Proses penentuan Pola atau penentuan Model yang terbaik sehingga hasil dari proses KDD akurat serta dapat dipercaya (Arta et al., 2019).

F. Evaluasi

Tahapan evaluasi bertujuan untuk melakukan evaluasi terhadap model atau pola yang telah terbentuk dari proses sebelumnya yaitu proses data mining sehingga hasil dari pola/model tersebut memiliki hasil akurasi yang dapat dipercaya (Arta et al., 2019).

3. Results and Discussion

A. Data Original

Dalam penelitian ini penulis menggunakan Kumpulan data yang terdiri dari tanda dan gejala penting dari individu yang menunjukkan tanda-tanda awal

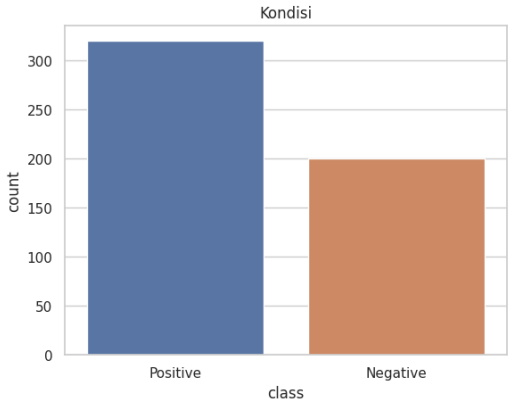
diabetes atau berisiko terkena diabetes. Variabel-variabel yang dimasukkan dalam kumpulan data memberikan wawasan berharga mengenai indikator-indikator potensial timbulnya diabetes. Kumpulan data tersebut mencakup beragam informasi, mulai dari rincian demografi hingga gejala spesifik yang terkait dengan diabetes.

B. Data Selection

Setelah memiliki data Original/Asli Langkah selanjutnya adalah melakukan seleksi data, seleksi data bertujuan untuk menentukan data yang relevan yang dapat digunakan untuk penelitian. Data yang digunakan terdapat 17 Atribut, dan dari 17 atribut tersebut adalah atribut Class/Label jumlah data sebanyak 520 data. Penjelasan dari Attribut dapat dilihat pada table 1:

Attribut	Keterangan
Usia	Merupakan Rentang Usia Individu (1-20 hingga 65 Th)
Jenis Kelamun	Informasi Gender (Laki-laki atau Perempuan)
Polyuria	Adanya Buang Air Kecil Berlebihan
Polydipsia	Rasa haus yang berlebihan
Penurunan Berat Badan Mendadak	Penurunan Berat Badan Secara Tiba-tiba
Kelemahan	Kelemahan Umum
Polyphagia	Rasa lapar yang berlebihan.
Genital Thrush	Adanya sariawan pada alat kelamin.
Penglihatan kabur	Penglihatan Kabu
Gatal	Adanya Rasa Gatal
Iritabilitas	Mudah Terkena Iritasi
Penyembuhan Tertunda	Penyembuhan Luka Yang Lama/Tertunda dari Biasanya
Paresis Parsial	Kurangnya kontrol penuh atas tubuh
Kekakuan Otot	Adanya Otot Yang Kaku
Alopecia	Rambut Rontok
Obesitas	Adanya Obesitas
Class	Klasifikasi Diabetes

Berikut adalah Perbandingan Class Negativ dan Positiv pada data Diabeter:



Gambar 2. Grafik Perbandingan Jumlah Class Positive dan Negative

Gambar 2 menunjukan bahwa jumlah individu yang Positive Diabetes adalah 320 dan yang negative Diabetes sebanyak 200 dari data tersebut menunjukan bahwa banyak Individu yang Positive Diabetes sehinggan Deteksi Dini Penyakit Diabetes sangat diperlukan, hal ini bertujuan untuk mencegah semakin meningkat pasien Diabetes mengingat bahwa penyakit diabetes merupakan salah satu penyakit kronis yang mematikan.

C. Data Cleaning

Langkah selanjutnya adalah melakukan proses pembersihan data atau *Data Cleaning*

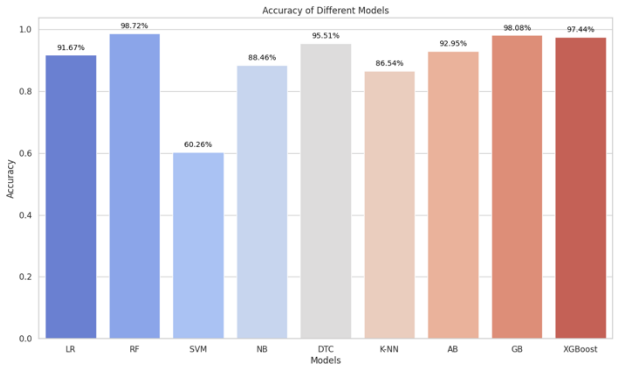
Pembersihan data dilakukan dengan membuang data yang tidak diperlukan dan melengkapi data yang kosong serta menghapus data duplikast, dari data Diabetes yang digunakan tidak ada data kosong dan data duplikat.

D. Data Transformation

Langkah selanjutnya adalah melakukan Transformasi Data, Transformasi Data digunakan untuk merubah data agar dapat dibaca oleh algoritma. Data Diabetes yang digunakan memiliki Type Data Text atau Object sehingga harus di Trasformasikan dulu menjadi Numerik agar dapat dibaca oleh algoritma.

E. Data Mining/Penerapan Algoritma

Dalam tahapan ini penulis menggunakan 8 Algoritma untuk melakukan Prediksi Deteksi dini Diabetes, Algoritma yang digunakan adalah *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM), *Naïve Bayes* (NB), Decision Tree (DT), *K-Nearest Neighbors* (KNN), *Ada Boost* (AB), *Gradient Boosting* (GB) dan *XGBoost Classifier* (XGB). Hasil dari 8 Algoritma tersebut dapat dilihat pada gambar 2 berikut:



Gambar 3. Grafik Hasil Akurasi Model Algoritma untuk melakukan Deteksi dini Diabetes

Gambar 3 merupakan hasil eksperimen yang dilakukan dari sembilan Algoritma yang diuji, Sembilan algoritma tersebut adalah *Gradient Boosting* (GB) memiliki akurasi sebesar 98,08%, Logistic Regression (LR) sebesar 91,67%, Random Forest (RF) sebesar 98,72%, Support Vector Machine (SVM) Sebesar 60,26%, Naïve Bayes (NB) sebesar 88,46% , Decision Tree (DT) sebesar 95,51%, K-Nearest Neighbors (KNN) sebesar 86,54%, Ada Boost (AB) sebesar 92,95%, dan XGBoost Classifier (XGB) sebesar 97,44%. Dari hasil eksperimen maka model terbaik yang dapat digunakan untuk deteksi dini penyakit diabetes adalah model dengan algoritma *Random Forest (RF)* dengan hasil akurasi 98,72%

F. Evaluasi

Tahapan Evaluasi adalah melakukan Evaluasi terhadap hasih akurasi dengan menggunakan *Receiver Operating Characteristic (ROC) Curve* atau Kurva ROC. Selain melihat Kurva ROC Evaluasi

juga dilakukan dengan melihat *Confusion Matrix* dengan mengecek nilai dari *precision recall, f1-score,* dan *Accuracy.*

1) Menentukan *precision, recall, f1-score,* dan *Accuracy.*

$$Akurasi = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

$$Recall = \frac{TP}{(TP + FP)}$$

$$Precision = \frac{(\mathrm{TP})}{(\mathrm{TP} + \mathrm{FP})}$$

$$f1\ score = \frac{(2 * (Recall * Precission)}{(Recall + Precission)}$$

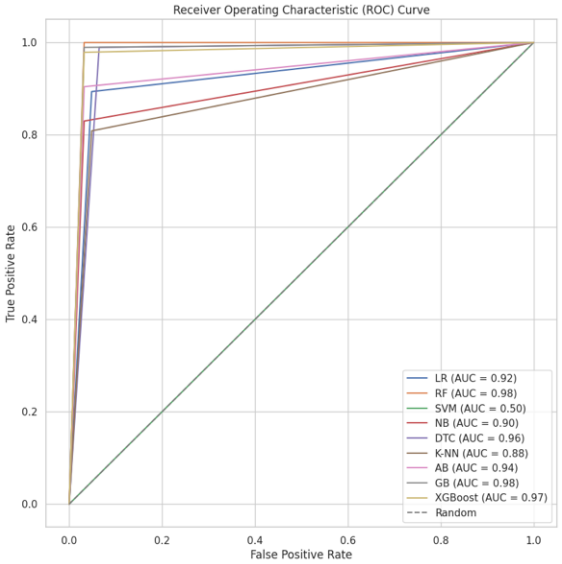
Tabel 2. Hasil Eksperimen

Algo	TP	FP	FN	TN	Akur	Pres	Rcll	F1S
NB	60	2	16	78	0,88	0,97	0,79	0,87
RF	**60**	**2**	**0**	**94**	**0,99**	**0,97**	**1,00**	**0,98**
KNN	59	3	18	76	0,87	0,95	0,77	0,85
DTC	58	4	1	93	0,97	0,94	0,98	0,96
XGB	60	2	2	92	0,97	0,97	0,97	0,97
SVM	0	62	0	94	0,60	0,00	0,00	0,00
LR	59	3	10	84	0,92	0,95	0,86	0,90
AB	60	2	9	85	0,93	0,97	0,87	0,92
GB	60	2	2	92	0,97	0,97	0,97	0,97

Tabel 2 merupakan hasil eksperimen dari hasil penerapan algoritma untuk menemukan algoritma terbaik untuk prediksi Diabetes, dari table diatas dapat disimpulkan bahwa Random Fores merupakan algoritma terbaik dengan hasil Akurasi 0,99, *Recall* 1,00, *Precision* 0,97, *f1-score* 0,98 .

2) *Receiver Operating Characteristic* (ROC) *Curve*.

ROC *(Receiver Operating Characteristic)* adalah salah satu jenis alat pengukuran kinerja untuk masalah klasifikasi dalam menentukan ambang batas suatu model. ROC merupakan alat ukur dengan Melihat Nilai AUC *(Area Under Curve)* semakin tinggi Nilai AUC maka hasil klasifikasi semakin bagus. Gambar 3 menunjukan nilai AUC dari Deteksi dini penyakit diabetes.



Gambar 3. Grafik *Receiver Operating Characteristic* (ROC)

Nilai AUC dari Sembilan algoritma yang digunakan adalah sebagai berikut untuk *Logistic Regression* (LR) Nilai AUC nya adalah 0,92, Random Forest (RF) nilai AUC sebesar 0,98 , Support Vector Machine (SVM) sebesar 0.50, Naïve Bayes (NB) sebesar 0,90, Decision Tree (DT) sebesar 0,96, K-Nearest Neighbors (KNN) sebesar 0,88, Ada Boost (AB) sebesar 0,94, Gradient Boosting (GB) sebesar 0,98 dan XGBoost Classifier (XGB) sebesar 0,97. Dari hasil nilai AUC maka hanya ada satu algoritma yang memiliki tingkat klasifikasi *Failure* yaitu model dengan algoritma SVM, sedangkan untuk RF, NB, DT, XGB, AB,

LR, GB, dan K-NN memiliki tingkat klasifikasi *Excellent*.

4. Conclusions

Penelitian ini menggunakan *Knowledge Discovery in Database Process* (KDD) untuk mengukur tingkat akurasi dalam Deteksi Dini Penyakit Diabetes hasil dari eksperimen yang dilakukan adalah sebagai berikut nilai akurasi dari Logistic Regression (LR) sebesar 91,67% dengan nilai AUC sebesar 0,92 , Random Forest (RF) nilai akurasi sebesar 98,72% dengan AUC sebesar 0,98, Support Vector Machine (SVM) nilai akurasi sebesar 60,26% dengan nilai AUC sebesar 0,50, Naïve Bayes (NB) dengan nilai akurasi 88,46% sedangkan untuk nilai AUC sebesar 0,90, Decision Tree (DT) dengan nilai akurasi sebesar 95,51% nilai AUC sebesar 0,96, K-Nearest Neighbors (KNN) nilai akurasi sebesar 86,54% nilai AUC sebesar 0,88 , Nilai akurasi dari Ada Boost (AB) sebesar 0,94 , Gradient Boosting (GB) memiliki nilai akurasi sebesar 98,08% sedangkan untuk nilai AUC sebesar 0,98 dan nilai akurasi dari XGBoost Classifier (XGB) sebesar 97,44% sedangkan untuk nilai AUC sebesar 0,97. Dari hasil eksperimen yang dilakukan maka model terbaik dalam penelitian ini adalah model dengan algoritma Random Forest (RF) dimana nilai akurasi sebesar 98,72% dan nilai AUC sebesar 0,98.

References

Agustina, V., Irma, M., Fanisa, T., Arum, C., Wulandari, D., Weya, A., & Lampongajo, O. G. C. (2021). Deteksi Dini Penyakit Diabetes Militus. *Jurnal Pengabdian Masyarakat*, *2*(2), 300–309.

algorit. (2022, March 11). *Cara Kerja Algoritma Random Fores*. Algorit. https://algorit.ma/blog/cara-kerja-algoritma-random-forest-2022/

Aprianti, B., Aulia, A., & Purnamasari, I. (2022). Algoritma Linear Regression dalam Memprediksi Pertumbuhan Jumlah Penduduk Menurut Provinsi dan Jenis Kelamin. *Jurnal Ilmiah Wahana Pendidikan*, *8*(5), 89–92.

Arta, J. K. I., Indarawan, G., & Dantes, R. G. (2019). Data Mining Rekomendasi Calon Mahasiswa Berprestasi Di STMIK Denpasar Menggunakan Metode Technique For Others Reference By Similarity To Ideal Solution. *Jurnal Ilmu Komputer Indonesia (JIKI)*, *4*(1).

Binus. (2021, September 3). *Proses Data Mining KDD*. Binus. https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/

DITPUI. (2023, January 16). *Diabetes Penyebab Kematian Tertinggi di Indonesia: Batasi dengan Snack Sehat Rendah Gula*. Direktorat Pengembangan Usaha UGM. https://ditpui.ugm.ac.id/diabetes-penyebab-kematian-tertinggi-di-indonesia-batasi-dengan-snack-sehat-rendah-gula/#

Mardi, Y. (2016). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Jurnal Edik Informatika*, *2*(2), 213–219.

Maulidah, N., Supriyadi, R., Utami, D. Y., Hasan, F. N., Fauzi, A., & Christian, A. (2021). Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes. *Indonesian Journal on Software Engineering (IJSE)*, *7*(1), 63–68. http://ejournal.bsi.ac.id/ejurnal/index.php/ijse63

Noviandi, N. (2018). Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes. *Indonesian of Health Information Management Journal (INOHIM)*, *6*(1), 1–5. https://doi.org/10.47007/INOHIM.V6I1.142

P2PTM. (2022, January). Penyakit Diabetes Melitus. *Direktorat P2PTM Kementerian Kesehatan*. https://p2ptm.kemkes.go.id/informasi-p2ptm/penyakit-diabetes-melitus

Rizki, R., Athallah, R., Cholissodin, I., & Adikara, P. P. (2022). Prediksi Potensi Pengidap Penyakit Diabetes berdasarkan Faktor Risiko Menggunakan Algoritme Kernel K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6*(8), 3777–3785. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11439

Wardana, Y., & Sari, D. P. (2023). Prediksi Penyakit Diabetes Dengan Naive Bayes. *Journal of Mathematics UNP*, *8*(3), 89–97. https://doi.org/10.24036/UNPJOMATH.V8I3.15070