Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

DETEKSI DINI GEJALA AWAL PENYAKIT DIABETES MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Ajeng Citra Mawani[1], Rusdah[2*], Law Li Hin[3], Dian Anubhakti[4]
[1,3,4] Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
[2]Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
Email: [1]ajengmawarni33@gmail.com, [2*]rusdah@budiluhur.ac.id, [3]lihin@budiluhur.ac.id, [4]dian.anubhakti@budiluhur.ac.id
(* : coressponding author)

Abstrak-Diabetes merupakan penyakit kronis yang disebabkan karena pancreas tidak dapat memproduksi insulin sesuai dengan kebutuhan tubuh atau kondisi ketika tubuh tidak dapat menggunakan insulin secara efektif. Pada tahun 2021 Indonesia memperoleh urutan ke-5 didunia dengan populasi penderita penyakit diabetes terbanyak dan terdapat lebih dari 1 orang diantara 10 orang dewasa yang menderita diabetes. Semakin meningkatnya penderita diabetes di Indonesia bahkan di dunia yang sebenarnya sudah positif diderita tetapi tidak menimbulkan komplikasi lebih lanjut hingga kematian. Hal ini disebabkan karena belum adanya model klasifikasi deteksi dini gejala awal diabetes. Maka pada penelitian ini perlu dilakukannya pembuatan model klasifikasi deteksi dini gejala awal penyakit diabetes dengan metode penelitian *Cross Industry Standard Process for Data Mining* (CRISP-DM) yaitu dengan melaksanakan riset jurnal. Penelitian ini menggunakan algoritma *Random Forest*. Data yang akan digunakan bersifat *public* yang didapatkan melalui website www.kaggle.com dengan total 520 *record* dataset yang terdiri dari 17 attribut, terdapat 320 dataset dengan positif diabetes dan 200 dataset dengan *negative* diabetes. Klasifikasi dilakukan dengan dengan komposisi data *training* dan data *testing* 90:10 menggunakan teknik *stratified random sampling* dengan *number of trees* 5, *maximal depth* 5, dan dilakukannya *apply pruning*. Diperoleh akurasi 90.38%, *precision* 100%, *recall* 84.38% dan niai AUC 1.00. Sehingga dapat disimpulkan bahwa model klasifikasi dengan algoritma *Random Forest* dapat bekerja sangat baik terhadap data deteksi dini gejala awal penyakit diabetes.

Kata Kunci: Diabetes, Deteksi Dini, Gejala Awal, Klasifikasi, *Random Forest*

Abstract-*Diabetes is a chronic disease caused by the pancreas not being able to produce insulin according to the body's needs or when the body is unable to use insulin effectively. In 2021 Indonesia ranks 5th in the world with the largest population of people with diabetes and more than 1 person in 10 adults has diabetes. There is an increasing number of diabetics in Indonesia and even in the world who are already positive but do not cause further complications or even death. This is due to the absence of a classification model for early detection of early symptoms of diabetes. So in this research it is necessary to make a classification model for early detection of early symptoms of diabetes with the Cross Industry Standard Process for Data Mining (CRISP-DM) research method, namely by carrying out journal research. This study uses the Random Forest algorithm. The data to be used is public in nature which is obtained through the website www.kaggle.com with a total of 520 dataset records consisting of 17 attributes, there are 320 datasets with positive diabetes and 200 datasets with negative diabetes. Classification was carried out using the composition of training data and testing data 90:10 using a stratified random sampling technique with a total of 5 trees, a maximum depth of 5, and pruning. Obtained 90.38% accuracy, 100% precision, 84.38% recall and 1.00 AUC value. So it can be interpreted that the classification model with the Random Forest algorithm can work very well for the detection of early data on symptoms of diabetes.*

***Keywords**: Diabetes, Early Detection, Early Symptoms, Classification, Random Forest*

1. PENDAHULUAN

Diabetes merupakan suatu penyakit dengan tingkatan kronis yang disebabkan ketidakmampuan pankreas dalam memproduksi *hormone insulin* sesuai dengan kebutuhan tubuh atau kondisi dimana tubuh tidak efektif dalam mengontrol *hormone insulin* yang telah diproduksi oleh pankreas [1].

International Diabetes Federation (IDF) Atlas edisi 10 memperkirakan terdapat 536.6 juta jiwa dengan penderita penyakit diabetes diantaranya orang dewasa pada usia 20-79 tahun di 215 wilayah negara. Pervalensi diabetes global yang terjadi pada orang dewasa dengan usia 20-79 tahun dengan perkiraan 10.8% terjadi pada pria dan 10.2% terjadi pada wanita [2]. Memungkinkan ada sekitar 239.7 juta jiwa yang tidak sadar dengan diabetes mereka, dengan variasi yang besar dalam proporsi diabetes yang tidak terdiagnosa disuluruh dunia [3].

Indonesia merupakan negara dengan populasi penderita penyakit diabetes pada tahun 2021 memperoleh urutan ke-5 didunia untuk orang dewasa berusia 20 tahun hingga 79 tahun sebanyak 19,5 juta jiwa. Hal ini menunjukan bahwa lebih dari 1 dari 10 orang dewasa akan menderita diabetes pada tahun 2021, dan jumlahnya akan terus bertambah dimasa yang akan datang [2].

Pada tahun 2021 pervalensi global diabetes menyatakan bahwa masih terdapat penderita diabetes yang tidak terdiagosis dengan angka yang sangat tinggi [3]. Dalam setiap tahunnya, terdapat 3,8 juta jiwa yang meninggal

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

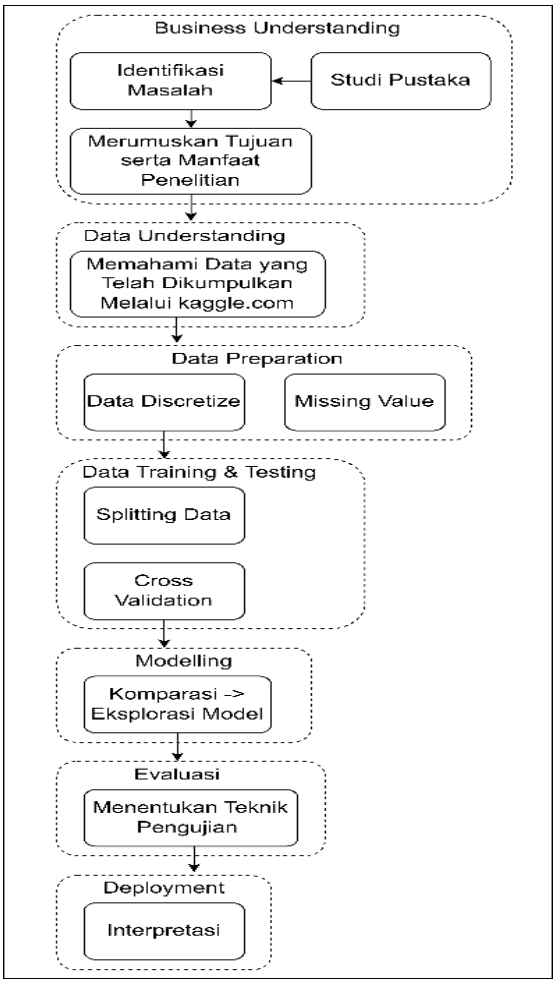
dunia akibat komplikasi diabetes yang disebabkan karena belum adanya model klasifikasi terkait gejala awal penyakit diabetes[4] Dengan adanya model klasifikasi deteksi dini pada gejala awal penyakit diabetes sangat perlu dilakukan karena terdapat fase asimtomatik, yaitu kondisi ketika suatu penyakit sudah pasti diderita tetapi tidak ada gejala klinis yang timbul pada penderitanya [5].

Dalam melakukan deteksi dini gejala awal penyakit diabetes maka perlu dilakukannya membuat model dan klasifikasi dengan data mining terkait gejala awal penyakit diabetes menggunakan data mining dengan tahapan penelitian *Cross-Industry Standard Process for Data Mining (CRISP-DM)* dan algoritma yang akan digunakan yaitu *Random Forest* yang diharapkan sebagai alat bantu masyarakat dalam melakukan deteksi dini penyakit diabetes secara mandiri.

Data mining dapat diartikan yaitu tahapan mencari inti dari suatu informasi dengan proses yang akan memanfaatkan teknik statistik, teknik matematika, dan kecerdasan buatan untuk mendapatkan inti dari suatu informasi dan dilakukannya identifikasi berdasarkan pola pola yang telah terbentuk dari kumpulan data yang jumlahnya sangat besar[6].

Klasifikasi dapat disebut sebagai proses pengelompokan suatu data dengan kegiatan menempatkan sebuah objek yang akan digunakan pada penelitian ke dalam kelas yang memiliki fungsi serupa. Dataset dengan jenis tipe data nominal atau *binner* dapat terdeskripsi penggunaanya dengan klasifikasi[7].

2. METODE PENELITIAN



Gambar 1. **Tahapan CRISP-DM**

Penelitian ini akan menggunakan metode CRISP-DM dalam melakukan langkah-penelitian agar penelitian dapat berjalan secara efektif dan terstruktur. Penelitian ini memiliki metodologi yang akan digunakan yaitu CRISP-DM. *framework* untuk menginterprestasikan berbagai permasalahan bisnis ke dalam teknik *data mining* dan melaksanakan tugas sesuai *data mining* secara *independent* pada aplikasi serta teknologi yang akan digunakan disebut sebagai *Cross-Industry Strandard Process For Data Mining* (CRISP DM) atau dapat diartikan sebagai proses standar lintas *industry* dalam *data mining*. CRISP-DM merupakan implementasi terfokus pada indutri yang

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

prosesnya diadopsi secara luas bedasarkan *Knowledge Discovery* (KD)[8]. Berikut merupakan penjelas terkait metode CRISP-DM yang akan pada penelitian ini:

2.1. Pemahaman Bisnis *(Business Understanding)*

Tahap yang pertama yaitu *Business Understanding* yang akan dilakukan dengan cara memahami permasalahan penyakit diabetes yang terus bertambah semakin banyak di Indonesia mapun dunia hingga tidak dapat dikendalikan. Peneliti melakukan riset gejala awal penyakit penyakit diabetes yang diharapkan dapat membantu masyarakat dalam melakukan deteksi dini penyakit diabetes secara mandiri.

2.2. Pemahaman Data *(Data Understanding)*

Data Understanding yaitu melakukan pengumpulan data. Data dengan tipe *public* yang telah berhasil dikumpulkan yaitu melalui website www.kaggle.com dengan judul *"Early Stage Diabetes Risk Prediction"*. Dataset ini disetujui oleh rumah sakit di Sylhet, Bangladesh melalui quisioner dari pasien diabetes. Dataset ini memiliki total 520 *record* dataset yang terdiri dari 17 *attributes* dengan total 320 data *positive* diabetes dan 200 data *negative* diabetes.

2.3. Persiapan Data *(Data Preparation)*

Penelitian ini menggunakan data yang bersifat public yang terdapat dari website kaggle.com memiliki data seperti *missing value* sehingga perlu dilakukan dilakukan tahap *pre-processing missing value*. Pada tahap data *pre-processing*, data akan dilakukan *data discretize* dengan tujuan trasformasi data berupa numerical menjadi categorical untuk mendapatkan model terbaik

2.4. *Modelling*

Tahap modelling akan dilakukan eksplorasi modelling dengan tujuan untuk mendapatkan model terbaik terhadap dataset yang telah dilakukan data *preparation.* Akan dilakukannya beberapa metode seperti *discretization by binning, splitting data, dan cross validation.* Pada tahap *discretization*, akan dilakukan transformasi data dengan tipe data *numeric* menjadi *categorical*. Pada tahap splitting data, dataset akan dipisah menjadi 2 sebgai data *training* dan data *testing*. Pada tahap *cross validation* akan dipisah ke dalam 2 bagian sebagai proses validasi atau evaluasi. Pada tahap modelling juga akan dilakukan komparasi menggunakan algoritma *Descision Tree, Naïve Bayes, KNN, Random Forest, dan Support Vector Machine.*). Namun dalam pengambilan sample teknik aapun, penting diingat bahwa kesuluruhan *sample* pengumpulan mewakili seluruh kumpulan data. Jika tidak, generalisasi dari pengklasifikasi yang diawasi tidak diketahui[9].

Pada penelitian ini akan dipilih algoritma *Random Forest*. *Random forest* merupakan salah satu metodologi pembelajaran untuk klasifikasi, regresi, dll. *Random Forest* dapat bekerja dengan membuat *cluster* dari *decision tree* pada saat *training* dan dapat menghasilkan *class* yang merupakan prediksi rata-rata dari individual *trees*. *Random forest* memiliki hubungan proporsional langsung antara jumlah *trees* yang terjadi dan akurasi yang dihasilkan[10]. Persamaan (1) dan (2) merupakan pehitungan algoritma *Random Forest* .

$$Entropy(Y) = -\sum_{i} p\,(c|Y) log_2 p(c \vee Y) \quad (1)$$

Information Gain, Y merupakan kumpulan kasus dan dan p(c|Y) meupakan rasio nilai Y terhadap kelas c.

$$InformationGain(y, a) = Entropy(Y) - \sum_{veValues} \frac{|Yv|}{|Ya|} Entropy(Yv) \quad (2)$$

Sedangkan *Values* (a) yaitu seluruh nilai yang memungkinan pada suatu kumpulan kasus:

a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a.

Atribut yang dipilih sebagai simpul, baik akar *(root)* maupun simpul internal akan disesuaikan dengan information gain yang diperoleh melalui hasil tertinggi pada atribut – atribut yang tersedia. Untuk penentuannya akan menggunakan rumus *gain ratio*, sedangkan gain ratio diperoleh dengan cara menentukan nilai *split information* dan dijelaskan pada persamaan (3) dan (4).

$$SplitInformation(S, A) = \sum_{i=1}^{c} \left(\frac{|Si|}{|S|}\right) log_2 \left(\frac{|Si|}{|S|}\right) \quad (3)$$

Split Information (S, A) merupakan nilai untuk estimasi *entropy* pada *variable input* S dengan kelas c dan $|Si|/|S|$ yaitu probabilitas kelas i pada atribut.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

$$GainRation(S,A) = \frac{InformationGain(S,A)}{SplitInformation(S,A)} \quad (4)$$

2.5. Evaluasi

Tahap evaluasi akan dilakukannya pengujian dengan menggunakan *confussion matrix* untuk mengukur perfoma algoritma terbaik yang digunakan yaitu *Random Forest* pada *machine learning classification*.

3. HASIL DAN PEMBAHASAN

3.1. Pemahaman Bisnis *(Business Understanding)*

Sesuai permasalahan yang terjadi yaitu semakin meningkat penderita diabetes di Indonesia maupun di dunia yang tidak terkendali, peneliti perlu melakukan riset melalui online yang terdapat pada website *science direct.google scholar, elsivier, serta ieee*. Jurnal yang dapat dilakukan riset hanya jurnal-jurnal dengan dataset yang sama pada penelitian ini dengan fokus penelitian yang sama yaitu terkait deteksi dini gejala awal.

Setelah didapatkan beberapa jurnal dengan kriteria yang sesuai, maka jurnal tersebut dapat dipahami secara teliti mulai dari rumusan masalah, tujuan dibuatnya penelitian, pemodelan dan algoritma yang digunakan pada penelitin tersebut hingga hasil akurasi yang didapatkan.

3.2. Pemahaman Data *(Data Understanding)*

Data yang telah berhasil peneliti kumpulkan bersifat data sekunder yang diperoleh melalui laman web www.kaggle.com dengan judul *Early Stage Diabetes Risk Prediction*. Dataset yang berhassil dikumpulkan terdiri dari 17 *attributes* dengan total 520 *record* dataset yang sudah bersih tanpa adanya *missing value* dan dapat dilihat pada Tabel 1. Pada penelitian ini akan menggunakan seluruh *attributes* yang ada pada dataset bertujuan untuk mengetahui secara luas gejala awal penyakit diabetes. Ke-16 *attributes* akan digunakan sebagai *regular attribute* dan 1 *attribute* lainnya akan digunakan sebagai *special attribute.*

Tabel 1. Attribut Data

No.	Nama *Attribut*	Tipe *Attribut*
1	*Age*	Integer
2	*Gender*	Binomial
3	*Polyuria*	Binomial
4	*Polydipsia*	Binomial
5	*Sudden weight loss*	Binomial
6	*Weakness*	Binomial
7	*Polyphagia*	Binomial
8	*Genital Trush*	Binomial
9	*Visual Blurring*	Binomial
10	*Itching*	Binomial
11	*Irritability*	Binomial
12	*Delayes Healling*	Binomial
13	*Partial Paresis*	Binomial
14	*Muscle Stiffness*	Binomial
15	*Alopecia*	Binomial
16	*Obisity*	Binomial
17	*Class*	Binomial

3.3. Persiapan Data *(Data Preparation)*

Dikarenakan data yang telah berhasil peneliti kumpulkan sudah bersih tanpa adanya kekurangan, maka tahap selanjutnya yaitu data *preparation* akan dilakukan *preprocessing* pada data dan dilakukannya pembagian data latih dan data uji yang akan dijelaskan sebagai berikut:

3.3.1. *Data Pre-processing*

Data precrocessing yang akan dilakukan pada tahap ini yaitu dengan menggunakan *discretize by binning* yaitu untuk membagi data numerik dengan cara melakukan mapping ke dalam *range interval* yang telah ditentukan ke dalam attribute nominal. *Discretization by binning* yang akan dilakukan yaitu dengan *number of bins 2*.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

3.3.2. Data Latih dan Data Uji *(Data Training and Data Testing)*

Data latih dan data uji akan dilakukan dengan menggunakan metode *splitting data*. Pada metode splitting data yang digunakan akan dilakukan *apply pruning* serta *number off trees* 10 dan *maximal depth* 5. Komposisi data latih dan data uji yang akan dilakukan pada penelitian ini yaitu 90:10, 80 : 20, 70 :30 dan 60 :40.

3.4. *Modelling*

3.4.1. *Discretization By Binning*

Setelah berhasil dilakukannya *discretizaion by binning* dengan *number of bins 2* menggunakan algoritma *Decision Tree, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine,* dan *Random Forest* didapatkan hasil pada Tabel 2.

Tabel 2. *Discretizaion by Binning*

Dataset	Akurasi				
	Decision Tree	Naïve Bayes	K-Nearest Neighbor	Random Forest	Support Vector Mechine
Asli	95.19%	87.31%	89.26%	**97.31%**	88.27%
Discretize	94.23%	87.69%	94.04%	95.34%	89.85%

Hasil terbaik setelah dilakukannya *discretization* dengan perbandingan mengunakan dataset asli yaitu pada algoritma *Random Forest* dengan akurasi 97.31% yang artinya dataset asli lebih baik jika dibandingkat dengan dataset yang telah dilakukan *discretization*. Maka kesimpulannya yaitu, *discretization* tidak perlu dilakukan pada penelitian klasifikasi deteksi dini gejala awal penyakit diabetes ini.

3.4.2. Data Latih dan Uji

Splitting data akan dilakukan meggunakan algoritma *Decision Tree, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine*, dan *Random forest* dengan pembagian data latih uji 90:10, 80:20, 70:30, dan 60:40 menggunakan dataset asli didapatkan hasil pada Tabel 3.

Tabel 3. Data Latih dan Uji

Dataset	Akurasi				
	Decision Tree	Naïve Bayes	K-Nearest Neighbor	Random Forest	Support Vector Mechine
90 – 10	96.15%	82.69%	88.46%	**98.08%**	84.62%
80 – 20	97.12%	87.50%	90.38%	96.15%	86.54%
70 – 30	96.79%	87.82%	90.38%	95.51%	85.90%
60 – 40	96.15%	84.62%	89.42%	96.15%	86.54%

Hasil akurasi tertinggi menggunaka teknik *splitting data* yaitu pada algoritma *Random Forest* dengan komposisi 90:10 dengan akurasi 98.08%.

3.4.3. *Cross Validation*

Pemodelan dengan *cross validation* akan dilakukan dengan menggunakan *number of fold* 10 yang artinya akan membagi data latih dan data uji kedalam 10 partisi yang berbeda. *Cross validation* akan dilakukan dengan algoritma yang sama pada *splitting data* yaitu *Decision Tree, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine,* dan *Random Forest* dengan dataset asli. Didapatkan hasil sebagai pada Tabel 4.

Tabel 4. *Cross Validation*

Dataset	Akurasi				
	Decision Tree	Naïve Bayes	K-Nearest Neighbor	Random Forest	Support Vector Mechine
90 – 10	98.29%	88.25%	94.44%	99.79%	89.10%
80 – 20	98.08%	86.30%	94.23%	99.76%	90.62%
70 – 30	98.35%	85.99%	93.68%	99.18%	89.29%
60 – 40	98.04%	84.94%	93.91%	99.36%	89.74%
Cross Validation	95.19%	87.31%	89.26%	97.31%	88.27%

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Hasil akurasi terbaik didapatkan oleh algoritma *Random forest* tanpa dilakukannya *cross validation*. Maka pada penelitian ini *cross validation* tidak perlu dilakukan dalam penelitian ini.

3.5. Evaluasi

Pada tahap evaluasi akan dilakukan pengujian algoritma terbaik dengan model terbaik menggunakan teknik *confusion matrix*. Algoritma terbaik yang akan dilakukan evaluasi yaitu *Random Forest* dengan menggunakan metode *splitting data* komposisi 90:10 dengan dataset asli tanpa dilakukannya *discretization* dan *cross validation*. Didapatkan hasil terbaik hasil pada Tabel 5.

Tabel 5. Evaluasi

	True Positive	True Negative	Class Precision
Pred.Positive	27	0	100%
Pred.Negative	5	20	80%
Class Recall	84.38%	100%	

Berikut merupakan perhitungan *confusion matrix* dari Tabel 5:

$$Precision = \frac{TP}{TP + FP}$$
$$Precision = \frac{27}{27+0} = 100\%$$

$$Recall = \frac{TP}{TP + FN}$$
$$recall = \frac{27}{27 + 5} = 84.38\%$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$accuracy = \frac{27+20}{27+20+0+5} = 90.38\%$$

Dari hasil *confussion matrix* yang telah dilakukan diatas, *precision* didapatkan sebesar 100%, *Recall* mendapatkan hasil 84.38% dan memiliki nilai akurasi didapatkan sebesar 90.38%.

4. KESIMPULAN

Berdasarkan hasil pengujian serta analisa yang telah berhasil dilakukan serta komprasi untuk mencari model terbaik yang nantinya akan digunakan untuk menguji kinerja performa algoritma terbaik, maka telah didapatkan algoritma dengan hasil akurasi terbaik yaitu *Random Forest* dengan komposisi data *training* 90:10 data *testing* dengan menggunakan teknik *sampling* yaitu *stratified random sampling* dam pada algoritma *Random forest* hanya menggunakan *number of trees* 5 dan *maximal depth* 5 dengan tujuan untuk mempermudah dalam membaca *trees* yang terbentuk, serta dengan dilakukannya *apply pruning* agar *trees* yang terjadi tidak terlalu luas dan menggunakan teknik pengujian *Confusion Matrix* dengan hasil akurasi 90.38% terkait dataset "*Early Stage Diabetes Risk Prediction"* pada model klasifikasi deteksi dini gejala awal penyakit diabetes.

DAFTAR PUSTAKA

[1] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0175-6.

[2] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, p. 109119, 2022, doi: 10.1016/j.diabres.2021.109119.

[3] K. Ogurtsova *et al.*, "IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021," *Diabetes Res. Clin. Pract.*, vol. 183, 2022, doi: 10.1016/j.diabres.2021.109118.

[4] R. Andanika Siallagan, "PREDIKSI PENYAKIT DIABETES MELLITUS MENGGUNAKAN ALGORITMA C4.5," *J. RESPONSIF*, vol. 3, no. 1, pp. 44–52, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti

[5] E. Karyadiputra and A. Setiawan, "Penerapan data mining untuk prediksi awal kemungkinan terindikasi diabetes," pp. 221–232, 2022.

[6] N. M. Putry, "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," *EVOLUSI J. Sains dan Manaj.*, vol. 10, no. 1, 2022, doi: 10.31294/evolusi.v10i1.12514.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 165-171
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

[7] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model," *Procedia CIRP*, vol. 79, pp. 403–408, 2019, doi: 10.1016/j.procir.2019.02.106.

[8] L. Husna, S. Syahputra, and B. S. Ginting, "Penerapan data mining menggunakan metode K-Means cluster untuk pengelopokkan data perizinan Madrasah Diniyah Taklimiyah Awwaliyah (MDTA) studi kasus Kementerian Agama Stabat," *J. Inform. Kaputama*, vol. 6, no. 3, 2022.

[9] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sens.*, vol. 11, no. 2, 2019, doi: 10.3390/rs11020185.

[10] J. S. Komputer, K. Buatan, and A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," 2020.