Volume 4, Nomor 1, Juni 2023, hlm 27-36
INFOTECH: Jurnal Informatika Teknologi
p ISSN 2722-9378 | e ISSN 2722-9386
http://jurnal.sttmcileungsi.ac.id/index.php/infotech

Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan *oversampling smote*

Using stochastic oversampling, decision tree algorithm analysis is used to predict diabetes

Dikan Ismafillah*, Tatang Rohana, Yana Cahyana

*Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Buana Perjuangan Karawang, Karawang, Indonesia. Jl. Hs. Ronggo Waluyo, Teluk Jambe Karawang, Jawa Barat, Indonesia

Informasi Artikel

Article History:
Submission: 19-01-2023
Revised: 28-01-2023
Accepted: 02-02-2023

Kata Kunci:
Penyakit diabetes; algoritma *machine learnin; k-fold cross validation; confusion matrix;* prediksi.

Keywords:
Diabetes, algorithm machine learning, k-fold cross validation, confusion matrix, prediction.

*** Korespondensi:**
Dikan Ismafillah
if19.dikanismafillah@mhs.ubpkarawang.ac.id

Abstrak

Kumpulan data ini didapat dari situs data dunia Kaggle yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* sebanyak 768 data yang terdiri dari 8 *variable* dan 1 *class* target. Penelitian ini menggunakan model *Random Forest (RF)* + SMOTE dan *Decision Tree (DC)* + SMOTE dengan matriks konfusi serta perhitungan *K-fold cross validation* yang bertujuan untuk memprediksi pengukuran diagnostik apakah seorang pasien menderita diabetes. Untuk mencapai tingkat akurasi terbaik, pada penelitian ini melakukan proses prediksi tingkat diabetes menggunakan dua algoritma, yaitu *Decision Tree* dan *Random Forest*. Pada data penyakit diabetes yang ditemukan terdiri dari *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age,* dan *Outcome(output).* Berdasarkan hasil dari penelitian yang telah dilakukan, pengujian model RF + SMOTE menggunakan *confusion matrix* dan metode K-*Fold Cross Validation* memberikan akurasi yang jauh lebih baik dalam distribusi data diabetes. Hasil pengujian menunjukkan akurasi data sebesar 88,9%. Dengan hasil perbandingan Kurva ROC nilai *Area Under the Curve (AUC) Random Forest* + SMOTE 89,0%.

Abstract

This data set was obtained from the world data site Kaggle which came from the National Institute of Diabetes and Digestive and Kidney Diseases as many as 768 data consisting of 8 variables and 1 target class. This study uses the Random Forest (RF) + SMOTE and Decision Tree (DC)+ SMOTE models with confusion matrices and K-fold cross validation calculations that aim to predict diagnostic measurements whether a patient has diabetes. To achieve the best level of accuracy, in this study the process of predicting the level of diabetes was carried out using two algorithms, namely the Decision Tree and Random Forest. The diabetes data found consisted of Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome(output). Based on the results of the research that has been done, testing the RF + SMOTE model using the confusion matrix and the K-Fold Cross Validation method provides much better accuracy in the distribution of diabetes data. The test results show a data accuracy of 88,9%. With the results of the comparison of the ROC Curve, the Area Under the Curve (AUC) Random Forest + SMOTE value is 89.0%.

1. PENDAHULUAN.

Diabetes atau yang sering disebut kencing manis adalah kondisi metabolisme kronis yang ditandai dengan tingginya kadar glukosa darah secara tidak normal [1]. Menurut *International Diabetes Federation* (IDF), jumlah penderita diabetes mencapai 463 juta orang. Para peneliti memperkirakan bahwa jumlah penderita diabetes akan meningkat menjadi 642 juta pada tahun 2040 [2]. Berdasarkan statistik dari Riset Kesehatan Dasar 2018, angka prevalensi diabetes pada orang dewasa pada tahun 2013 sebesar 6,9%, dan meningkat menjadi 8,5% pada tahun 2018 [3]. *Federasi Diabetes Internasional* memperkirakan bahwa 6,7 juta orang di seluruh dunia akan meninggal akibat diabetes pada tahun 2021, yang memperkirakan adanya kematian setiap 5 detik. China merupakan negara dengan angka kematian akibat diabetes tertinggi di dunia. Pada tahun 2021, 1,39 juta orang akan meninggal karena diabetes di Tiongkok. Pada urutan kedua adalah Amerika Serikat dengan 669.000 kematian [4]. Kemudian India di posisi ketiga dengan total 647.000. Indonesia termasuk pada urutan keenam dalam daftar ini. Jumlah kematian akibat diabetes di Indonesia akan meningkat menjadi 236.000 pada tahun 2021. Menurut *Federasi Diabetes Internasional* (IDF), akan ada 537 juta orang di seluruh dunia (usia 20 hingga 79) yang hidup dengan diabetes pada tahun 2021, ini mewakili satu dari 10 orang. Empat dari lima penderita diabetes tinggal di negara berpenghasilan rendah dan menengah [4].

Diabetes adalah suatu kondisi yang disebabkan oleh kadar gula darah yang sangat tinggi dalam tubuh manusia. Pada diabetes tipe 1 seseorang akan mengalami kondisi kadar gula yang berlebih ketika pankreas gagal menghasilkan insulin dalam jumlah yang cukup. Di sisi lain, diabetes tipe 2 menggambarkan suatu kondisi di mana tubuh tidak mampu memanfaatkan insulin secara efektif [2][5]. Diabetes yang tidak dikelola dengan baik dapat menimbulkan berbagai akibat jangka panjang, antara lain kerusakan pembuluh darah dan saraf, serta organ penting dalam tubuh, bahkan kematian [6]. Deteksi dini diabetes mungkin merupakan metode pencegahan diabetes yang efektif, karena meningkatkan kemungkinan bahwa individu akan memilih untuk menjalani gaya hidup yang lebih sehat [7]. Data mining dan pembelajaran mesin membuat kemajuan yang stabil untuk menjadi alat yang dapat diandalkan dalam pembuatan model komputer yang memungkinkan tingkat akurasi yang tinggi dalam prediksi diabetes [8][5].

Dari hasil riset menunjukkan bahwa model klasifikasi *Decision tree* berkinerja buruk hanya pada 0,68, sedangkan algoritma *Random Forest* berkinerja cukup baik pada 0,72 [7]. Pada penelitian perbandingan dengan beberapa algoritma machine learning klasifikasi yaitu *Decision Tree, Naïve Bayes, k-Nearest Neighbour*, *Random Forest*, dan *Decison Stump*. Hasil penelitian tersebut mendapatkan nilai *accuracy* sebesar 80.38% dari perbandingan terbaik algoritma Random Forest [9]. Demikian juga, pada penelitian hasil *accuracy* yang didapat menggunakan algoritma *Decision Tree* sebesar 85.28% pada proses identifikasi penyakit diabetes [10]. Selanjutnya, penelitian dengan mencoba menerapkan suatu metode klasifikasi untuk memprediksi apakah seseorang terkena penyakit diabetes menggunakan algoritma *Decision Tree* ID3 dengan nilai *accuracy* sebesar 84,77% [1] .

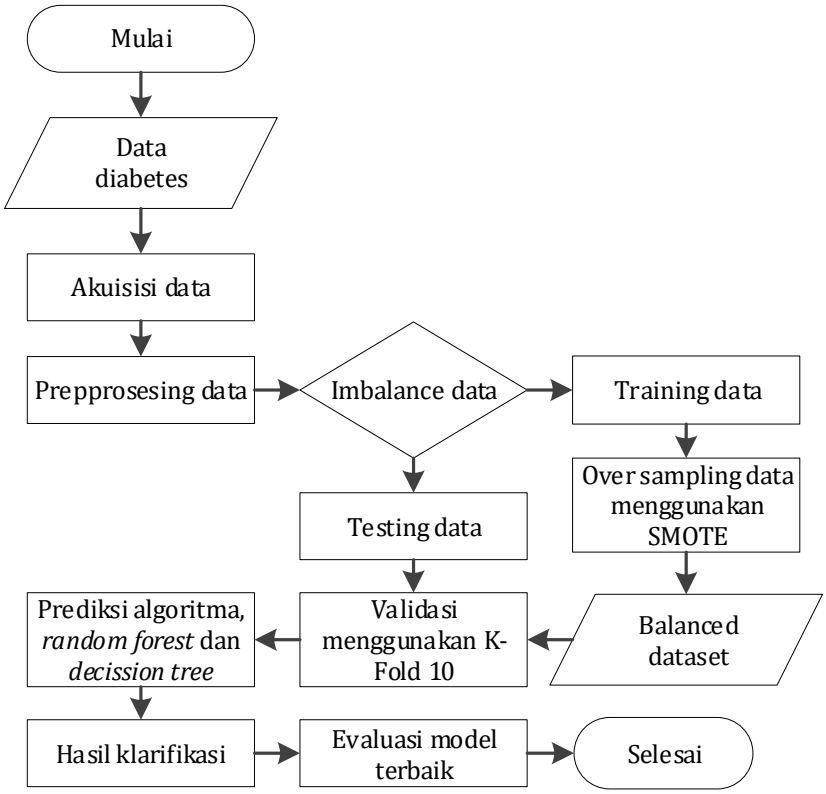
Oleh karena itu, penelitian ini menerapkan algoritma pembelajaran mesin yang bertujuan untuk memprediksi penyakit diabetes. Algoritma ini diharapkan mampu membantu dalam melakukan prediksi awal penyakit diabetes menggunakan dua algoritma yaitu algoritma *Decission Tree (DC)* dan *Random Forest (RF)*. Informasi dianalisis dan dataset diolah untuk menghasilkan produksi nilai prediksi model terbaik dari kedua algoritma. Setelah nilai antisipasi dari algoritma *Decission Tree (DC)* dan *Random Forest (RF)* ditentukan, dilakukan perbandingan evaluasi kinerja algoritma *Decission Tree (DC)* dan *Random Forest (RF)*. untuk memutuskan algoritma mana, dari keduanya, yang lebih efektif. Temuan perbandingan algoritma ini kedepannya akan menjadi standar diagnosis dini dan prediksi penyakit diabetes.

2. METODE

2.1 Alur penelitian

Pada alur penelitian ini salah satu metode klasifikasi dapat dinilai menggunakan berbagai faktor, akurasi, kecepatan, kehandalan, skalabilitas, dan interpretasi. Pada tahap ini, prediksi diabetes dilakukan dengan menggunakan kedua algoritma machine learning, yaitu *Decission Tree (DC)* dan *Random Forest (RF).* Dataset diabetes yang dimiliki dibagi menjadi data uji dan data latih,

kemudian dilakukan prosesing data dan perbandingan keempat algoritma yang terlampir pada **Gambar 1** untuk mengetahui proses alur penelitian.



Gambar 1. Alur penelitian

Data ini didapat dari situs data dunia Kaggle yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases*, berisi lebih dari 700 data perempuan berusia minimal 21 tahun keturunan India Pima. Data yang dikumpulkan berdasarkan survei data pemeriksaan rutin penduduknya yang tinggal di dekat Phoenix, Arizona, AS yang telah diperbaharui pada tahun 2020 [11]. Jenis data ini merupakan jenis data yang dapat langsung diukur atau dihitung sebagai nilai kategorikal atau numerik. Sebanyak 768 data yang terdiri dari 8 *variable* dan 1 *class* target. Berikut adalah contoh dataset yang digunakan dalam penelitian ini untuk prediksi model algoritma yang akan dilakukan yang disajikan dalam **Tabel 1.**

Tabel 1. *Dataset diabetes*

	Pregnancies	***Glucose***	***BloodPressure***	***SkinThickness***	***Insulin***	***BMI***	***DiabetesPedigreeFunction***	***Age***	***Outcome***
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Pada **Tabel 1** menunjukan data dari 0 sampai 4 merupakan salah satu sebagian data paling awal dari 768 dataset yang terdiri dari: *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age,* dan *Outcome*.

2.2 SMOTE

Metode *Synthetic Minority Oversampling Technique (SMOTE)* adalah pendekatan untuk menyeimbangkan data sampel yang tidak seimbang (mayoritas) dengan berfokus pada kelas minoritas untuk meningkatkan efisiensi metode prediksi [12]. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan kumpulan data dengan melakukan resampling sampel kelas minoritas. Berdasarkan penelitian yang telah dilakukan kurangnya keseimbangan

data pada data set penyakit diabetes. Oleh karena itu, diperlukan teknik pengambilan sampel data untuk menyeimbangkan data yang diperiksa dengan teknik SMOTE. Pada kasus sebelumnya SMOTE dapat menghilangkan noise dan menyelesaikan masalah ketidak seimbangan.

2.3 Decision Tree

Decision tree adalah model prediktif yang hasil akhirnya diklasifikasikan berdasarkan struktur atau pohon hierarkis [13] dapat menganalisis *variabel* data nominal dan numerik secara bersamaan. *Decision tree* akan mencari solusi permasalahan dengan menjadikan kriteria sebagai node yang saling berhubungan membentuk seperti struktur pohon. Setiap pohon memiliki cabang dan cabang mewakili setiap atribut yang dipenuhi untuk menuju cabang selanjutnya hingga berakhir didaun (tidak ada cabang lagi). *Decision Tree* C4.5 merupakan versi perbaikan dari algoritma sebelumnya yang disebut algoritma *iterative dichotomizer* (ID3). Berikut ini adalah daftar tahapan yang dilakukan algoritma C4.5 saat membuat pohon keputusan [14].

- Manfaatkan atribut sebagai root.
- Buat cabang terpisah untuk setiap nilai yang mungkin.
- Membahas setiap contoh individu di dalam cabang.
- Lakukan prosedur untuk setiap proses percabangan berulang kali hingga semua instance di cabang memiliki kelas yang sama.

Salah satu faktor yang membedakan *Decision Tree* dan *iterative dichotomizer* (ID3) yaitu melalui pruning yang mengacu pada proses menghilangkan *outlier* dan sumber *noise* pada suatu dataset untuk meningkatkan akurasi klasifikasi dan prediksi. *Rasio gain* untuk setiap atribut dihitung menggunakan Algoritma C4.5, dan atribut yang telah ditentukan memiliki nilai terbesar dipilih menjadi node [14]. Proses mining dengan algoritma *Decision Tree* C4.5 dimulai dengan menghitung nilai *Entropy* dan *Gain* dari masing-masing atribut data training yang ada sehingga menghasilkan *Gain Ratio*.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (1)$$

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah partisi atribut A
|Si| = Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

Sementara itu, perhitungan nilai entropi dapat pada persamaan 2:

$$Entropy(S) = \sum_{i=1}^{e} -p_i * log_2 p_i \quad (2)$$

Keterangan:
S = Himpunan kasus
n = Jumlah partisi atribut A
pi = Proporsi dari S_i terhadap S

Atribut yang memiliki *Gain Ratio* terbesar dipilih untuk membuat simpul akar. Selanjutnya menghitung nilai *Gain* dan *Entropy* dari masing–masing atribut dengan menghilangkan atribut yang telah dipilih sebelumnya. Atribut yang memiliki *Gain Ratio* terbesar dipilih untuk membuat simpul internal. Ulangi perhitungan tersebut hingga semua atribut memiliki kelas. Jika semua atribut atau pohon sudah memiliki kelas, maka tampilkan pohon keputusan awal dan generate aturan keputusan awal.

2.4 Random forest

Algoritma *Random Forest* adalah salah satu yang menggunakan pembelajaran *supervised*. Ini menghasilkan kumpulan pohon keputusan, yang sering dilatih menggunakan pendekatan

"*bagging*". Inilah artinya jika mengacu pada "*forest*". Premis dasar di balik pendekatan *bagging* adalah bahwa menggabungkan beberapa model pembelajaran harus mengarah pada peningkatan produk akhir. *Random Forest classifier* merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi [15]. *Random Forest* juga dapat diartikan sebagai terbentuk dari serangkaian pohon keputusan atau *decision tree*. Dapat digunakan untuk memprediksi kategori dengan beberapa nilai yang mungkin dan probabilitas keluaran yang dapat disesuaikan. Satu hal yang harus diperhatikan adalah *overfitting*. *Random Forest* adalah sejenis pembelajaran mesin yang membangun banyak pohon keputusan dan kemudian menggabungkan pohon-pohon itu untuk mendapatkan prediksi yang lebih akurat dan konsisten. Dimungkinkan untuk menggunakan *Random Forest* untuk tugas klasifikasi dan regresi, yang mencakup sebagian besar sistem pembelajaran mesin modern. Ini adalah salah satu keunggulan utama *Random Forest* [16]. Selama proses perancangan model algoritma *Random Forest*, model algoritma *Random Forest* dibangun menggunakan n pohon *Decision Tree* dengan terlebih dahulu menentukan nilai terbaik untuk n, kemudian mendapatkan hasil optimal dari algoritma *Random Forest*, dan terakhir memanfaatkan splitter terbaik untuk mendapatkan hasil akurasi terbaik dalam kasus penyakit diabetes. Semua langkah ini dilakukan untuk memastikan model algoritma *Random Forest* seakurat mungkin [17]. Keuntungan menggunakan algoritma *random forest* adalah dapat mengklasifikasikan data dengan atribut yang tidak lengkap. Baik untuk klasifikasi data dan pengolahan data sampel besar. Pada penelitian sebelumnya, ekstraksi ciri warna dilakukan dengan algoritma random forest, dengan hasil 85% menghasilkan nilai akurasi tertinggi dari proses ekstraksi [18].

3. HASIL DAN PEMBAHASAN

3.1 Pre-processing

Langkah pra-pemrosesan data penelitian dilakukan dengan cara yang membuat data lebih sesuai untuk digunakan pada langkah selanjutnya, termasuk menghilangkan komponen yang tidak lengkap untuk menghindari lebih dari 768 manipulasi. Data dibersihkan untuk menghindari data duplikat atau nilai nol. Terlampir lampiran **Tabel 2** untuk *preprocessing* data.

Tabel 2. Sebelum *processing* data

	Pregnancies	***Glucose***	***BloodPressure***	***SkinThickness***	***Insulin***	***BMI***	***DiabetesPedigreeFunction***	***Age***	***Outcome***	***New BMI***	***NewInsulinScore***	***NewGlucose***
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1	*Obesity 1*	*Abnormal*	*Secret*
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0	*Overweight*	*Normal*	*Normal*
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1	*Normal*	*Abnormal*	*Secret*
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	*Overweight*	*Normal*	*Normal*
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	Obesity 3	*Abnormal*	*Secret*

Tabel 2 menunjukan terjadinya preprocessing data pada variabel kategori dalam kumpulan data yang harus diubah menjadi nilai numerik. Untuk itu proses transformasi ini dilakukan dengan metode Label *Encoding dan One Hot Encoding*. Berikut lampiran **Tabel 3** dan **Tabel 4** yang berisi hasil prosesing data.

Tabel 3. Hasil *processing data*

	Pregnancies	***Glucose***	***BloodPressure***	***SkinThickness***	***Insulin***	***BMI***	***DiabetesPedigreeFunction***	***Age***	***Outcome***	***NewBMI_Obesity 1***	***NewBMI_Obesity 2***	***NewBMI_Obesity 3***
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1	1	0	0
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0	0	0	0
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1	0	0	0
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	0	0	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	0	0	1

Tabel 3 menunjukan hasil dari proses pembersihan data dimana telah dihilangkan nilai nul dan merubah nilai kategori menjadi nilai numerik. Terjadi pada kolom data yang ditransformasikan ditahapan preprocessing *NewBMI, NewInsulinScore, NewGlucose.*

Tabel 4. Hasil *processing* data lanjutan

	NewBMI_Overweight	*NewBMI_Underweight*	*NewInsulinScore_Normal*	*NewGlucose_Low*	*NewGlucose_Normal*	*NewGlucose_Overweight*	*NewGlucose_Secret*
0	0	0	0	0	0	0	1
1	1	0	1	0	1	0	0
2	0	0	0	0	0	0	1
3	1	0	1	0	1	0	0
4	0	0	0	0	0	0	1

Tabel 4 menunjukan hasil *preprocessing* data lanjutan pada tabel 3 data kategorikal dikonversi menjadi numerik dimana 1 = ya dan 0 = tidak. Data *preprocessing* dilakukan untuk mengubah data karakter yang tidak konsisten menjadi numerik. Pengubahan data tersebut dilakukan untuk mempermudah proses modeling menggunakan algoritma *Random Forest* dan *Decision Tree.*

3.2 Evaluasi

Uji kinerja dijalankan menggunakan model algoritma pembelajaran mesin untuk memvalidasi kinerja model algoritma berdasarkan data uji. Perhitungan *confusion matrix* untuk menentukan baik atau tidaknya suatu algoritma pada data diabetes tergantung dari akurasi dan presisi. Akurasi klasifikasi, sensitivitas, dan spesifisitas adalah tiga aspek evaluasi yang dapat digunakan untuk meningkatkan hasil prediksi. Akuisisi klasifikasi Adalah suatu proses untuk mengambil, mengumpulkan dan menyiapkan data, hingga memprosesnya untuk menghasilkan data terbaik, inilah yang dimaksud dengan "akurasi dalam klasifikasi". Sementara itu, sensitivitas adalah ukuran seberapa baik kejadian yang diinginkan dapat diprediksi. Selain itu, spesifisitas adalah kriteria yang menentukan proporsi kejadian yang tidak menguntungkan [19]. Dengan menggunakan angka-angka yang termasuk dalam confusion matriks, seseorang dapat menilai akurasi, sensitivitas, dan spesifisitas klasifikasi. Dengan menganalisis hasil dari beberapa model prediksi, seseorang dapat memutuskan apakah pernyataan tertentu benar atau tidak benar dengan menggunakan perangkat yang dikenal sebagai confusion matriks. Oleh karena itu, untuk memastikan hasil dari confusion matriks, yaitu dengan membandingkannya dengan kategori input utama [20]. Proses evaluasi yang dilakukan menggunakan data tes yang dipisahkan pada proses sebelumnya, dan hasil evaluasi ini menggunakan perhitungan matriks konfusi yang disajikan pada rumusan 3, 4, 5 dan 6.

$$Accuracy = \frac{True\ Positive + True\ Negative}{SUM\ The\ Number\ of\ Data}\ x\ 100\% \qquad (3)$$

$$Precision = \frac{True\ Posistive}{True\ Positive + False\ Positive}\ x\ 100\% \quad (4)$$

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}\ x\ 100\% \quad (5)$$

$$F1 = 2\ x\ \frac{precision\ x\ recall}{precision\ + recall} \quad (6)$$

Pada proses pengujian tahapan ini menggunakan teknik 10 *K-Fold Cross Validation*. Teknik yang bekerja dengan cara memisahkan K data menjadi data uji dan data latih. Kemudian melakukan prediksi menggunakan algoritma *Decision Tree* dan *Random Forest*. Nilai akurasi dan presisi didapatkan dari setiap metode deteksi yang digunakan.

Langkah-langkah untuk k-fold crossing adalah: Pertama bagi data dengan faktor k. Kemudian, pada percobaan pertama, buat bagian pertama sebagai data uji dan bagian lainnya sebagai data pelatihan. Pada percobaan kedua, mengubah satu bagian untuk data uji dan satu bagian lagi untuk data pelatihan. Tes ketiga mengubah bagian ketiga menjadi data uji dan bagian sisanya menjadi data pelatihan. Ketiga, dari hasil tersebut, matriks konfusi digunakan untuk menyimpan skor evaluasi kinerja model. Kemudian tentukan rata-rata untuk setiap percobaan. Keempat, dilakukan eksperimen dengan menggunakan model algoritma terpilih yang dapat dijadikan referensi. Berikut tabel hasil persilangan k-*fold 10* yang terlampir pada **Tabel 5**.

Tabel 5. Hasil persilangan *K-fold* 10

K-Fold	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
DC + SMOTE	90,0%	88,5%	84,2%	90,0%	85,7%	86,9%	91,3%	88,4%	86,9%	84,0%
RF + SMOTE	88,5%	91,4%	90,0%	95,7%	92,8%	91,3%	95,6%	94,2%	84,0%	89,8%

Tabel 5 menunjukan hasil metode persilangan *K-Fold Cross validation* yang digunakan untuk mengevaluasi kinerja model. Nampak pada bagian *Random Forest (RF)* + SMOTE jika dirata – rata kinerja algoritma ini jauh lebih baik dibandingkan *Decision Tree (DC)* + SMOTE.

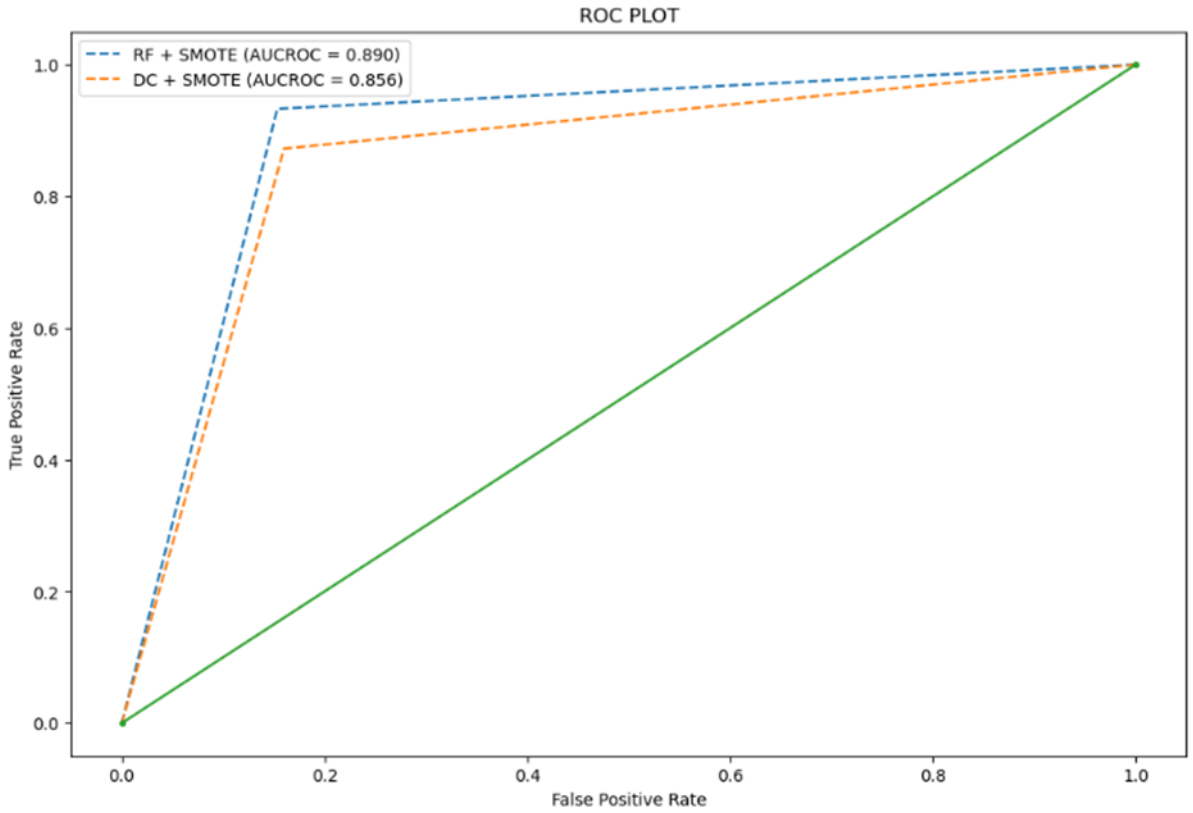
3.3 Perbandingan hasil

Berdasarkan penelitian sebelumnya, SMOTE digunakan untuk menyeimbangkan data. Hasil yang diperoleh untuk tingkat akurasi yang dibandingkan dengan tingkat akurasi prediksi menggunakan kedua algoritma. Penelitian yang telah dilakukan menggunakan Teknik *oversampling* SMOTE, untuk menyeimbangkan data sampel untuk kelas (mayoritas) yang terlalu tidak seimbang dengan berfokus pada kelas minoritas. Dapat dilihat pada **Tabel 6** penggunaan SMOTE pada *Random Forest* sangat berpengaruh bahkan penggunaan SMOTE dapat memperbaiki *accuracy* dengan pembagian data secara seimbang menggunakan K-fold dan oversampling.

Tabel 6. Hasil perbandingan *K-fold* 10

K-FOLD CROSS VALIDATION 10	
Metode	**Akurasi**
DC + SMOTE	87,6%
RF + SMOTE	91,3%

Pada **Tabel 6** menjelaskan hasil perbandingan K-fold menggunakan algoritma *Random Forest* + SMOTE memiliki *accuracy* lebih baik sebesar 91,3% dibandingkan *Decision Tree* dengan *accuracy* 87,6%. Dari hasil prediksi tersebut accuracy algoritma *Random Forest* ini sangat efektif digunakan untuk prediksi penyakit diabetes dan membantu tenaga medis dalam mengambil langkah–langkah yang tepat untuk mencegah penyakit diabetes. Dibutuhkan juga pengukuran kinerja algoritma menggunakan kurva untuk mengetahui tingkat statistiknya yang terlampir pada **Gambar 2.**



Gambar 2. Hasil perbandingan kurva *ROC*

Gambar 2 menunjukan hasil pengujian model yang dilakukan adalah untuk mengukur tingkat akurasi dan kinerja algoritma berdasarkan nilai AUC (*Area Under Curve*) dari prediksi penyakit diabetes dengan metode *cross validation*. Hasil dari pengujian model yang telah dilakukan adalah untuk mengukur kinerja suatu algoritma dan AUC (*Area Under Curve*). Grafik ROC RF + SMOTE dengan nilai AUC (*Area Under Curve*) sebesar 0.890 lalu dikalikan 100% untuk hasil akhir berupa persentase 89,0% yang terlampir pada **Tabel 7**.

Tabel 7. Hasil prediksi penyakit diabetes

HASIL PREDIKSI		
Metode	**Akurasi**	**ROC**
DC + SMOTE	85,6%	85,6%
RF + SMOTE	88,9%	89,0%

Tabel 7 menunjukan evaluasi yang dilakukan dengan confusion matrix dan kurva ROC keduanya menunjukkan bahwa hasil pengujian algoritma Random Forest + SMOTE memiliki nilai akurasi yang lebih tinggi jika dibandingkan dengan hasil pengujian algoritma Decision Tree + SMOTE. Kesimpulan tersebut dapat ditarik dari hasil pengujian yang telah disajikan sebelumnya. Model Random Forest + SMOTE memiliki rating 88,9% untuk nilai akurasinya, sedangkan hasil kinerja algoritma menggunakan kurva ROC dan menghasilkan nilai kinerja model Area Under the Curve (AUC) sebesar 89,0%.

4. SIMPULAN

Berdasarkan penelitian yang dilakukan dapat disimpulkan tujuan penelitian ini adalah untuk mengetahui seberapa akurat model algoritma *machine learning* yang digunakan dalam kasus prediksi penyakit diabetes. Penelitian ini menggunakan dataset yang diperoleh melalui Kaggle yang berasal dari *National Institute of Diabetes* penduduk Arizona, AS tahun 2020. Data Diabetes

yang sebelumnya 768 data menjadi 760 data setelah dilakukan tahapan *pre-processing*. *Accuracy, Recall, Precision, F1 Score*, yang didasarkan pada Confusion Matrix, dan grafik ROC/AUC digunakan untuk mengevaluasi hasil penelitian. Penerapan algoritma *machine learning* dalam prediksi penyakit diabetes berhasil dilakukan dengan baik. Berdasarkan hasil prediksi analisis *accuracy* penyakit diabetes menggunakan algoritma *Decision Tree* dan *Random Forest* didapatkan hasil perbandingan terbaik untuk menentukan metode mana yang lebih efisien dalam memprediksi kasus ini. Pada proses perhitungan K-fold 10 dan SMOTE, penerapan algoritma *Random Forest* menggunakan SMOTE sangat efisien dibandingkan dengan DC + SMOTE yang memiliki tingkat prediksi kurang akurat. Evaluasi model juga berhasil dilakukan dengan nilai prediksi yang sangat baik sebesar 88,9% untuk *Random Forest* dengan nilai akurasi tertinggi pada kasus penyakit Diabetes. Dengan hasil Kurva ROC dengan nilai *Area Under the Curve (AUC)* 89,0%.

DAFTAR PUSTAKA

[1] M. S. Efendi and H. A. Wibawa, "<i>Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik (Diabetes Prediction using ID3 Algorithm with Best Attribute Selection) <i/>," *<i>Juita<i/>*, vol. VI, no. 1, pp. 29–35, 2018.

[2] B. M. K. P, S. P. R, N. R K, and A. K, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier," *Int. J. Cogn. Comput. Eng.*, vol. 1, no. July, pp. 55–61, 2020, doi: 10.1016/j.ijcce.2020.10.002.

[3] A. S. Putri, "Tahun 2018 Penderita Diabetes di Indonesia Meningkat," *2018*, 2018. https://www.fimela.com/lifestyle/read/3739252/tahun-2018-penderita-diabetes-di-indonesia-meningkat

[4] R. Pahlevi, "Kasus Kematian Akibat Diabetes di Indonesia Terbesar Keenam di Dunia," *2021*, 2021. https://databoks.katadata.co.id/datapublish/2021/11/26/kasus-kematian-akibat-diabetes-di-indonesia-terbesar-keenam-di-dunia

[5] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[6] D. Care and S. S. Suppl, "1. Improving care and promoting health in populations: Standards of medical care in diabetes-2020," *Diabetes Care*, vol. 43, no. January, pp. S7–S13, 2020, doi: 10.2337/dc20-S001.

[7] W. Nugraha and R. Sabaruddin, "Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4.5, Random Forest, dan SVM Resampling Technique for Handling Class Imbalance in the Classification of Diabetes using C4.5, Random Forest, and SVM," *Agustus*, vol. 20, no. 3, pp. 352–361, 2021, [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[8] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.icte.2018.10.005.

[9] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[10] J. B. Junior, R. R. Saedudin, and V. P. Widharta, "Perbandingan Akurasi Algorima Decision Tree Dan Algoritma Support Vector Machine Pada Penyakit Diabetes," vol. 8, no. 5, pp. 9749–9756, 2021.

[11] Rdrr.io, "Pima Indians Diabetes Database.," *Subgroup Discovery dan Bump Hunting*, 2020.

[12] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," *Isd* , vol. 3, no. 1, pp. 2528–5114, 2018.

[13] N. Azizah, "Komparasi Metode Klasifikasi Decision Tree Algoritma C4.5 Dan Random Forest Untuk Prediksi Penyakit Stroke," 2021.

[14] A. Franseda, W. Kurniawan, S. Anggraeni, and W. Gata, "Integrasi Metode Decision Tree dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 3, p. 282, 2020, doi: 10.26418/justin.v8i3.40982.

[15] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[16] dqlab.id, "Studi Kasus Random Forest Machine Learning untuk Pemula Data."

[17] A. Samosir, M. Hasibuan, W. E. Justino, and T. Hariyono, "Komparasi Algoritma Random Forest , Naïve Bayes dan K- Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung," pp. 214–222, 2021.

[18] N. Khasanah, R. Komarudin, N. Afni, Y. I. Maulana, and A. Salim, "Skin Cancer Classification Using Random Forest Algorithm," *Sisfotenika*, vol. 11, no. 2, p. 137, 2021, doi: 10.30700/jst.v11i2.1122.

[19] J. J. Pangaribuan, C. Tedja, and S. Wibowo, "PERBANDINGAN METODE ALGORITMA C4.5 DAN EXTREME LEARNING MACHINE UNTUK MENDIAGNOSIS PENYAKIT JANTUNG KORONER," 2019.

[20] E. Prasetyo, B. Prasetiyo, and P. Korespondensi, "PENINGKATAN AKURASI KLASIFIKASI ALGORITMA C4.5 MENGGUNAKAN TEKNIK BAGGING PADA DIAGNOSIS PENYAKIT JANTUNG INCREASED CLASSIFICATION ACCURACY C4.5 ALGORITHM USING BAGGING TECHNIQUES IN DIAGNOSING HEART DISEASE," vol. 7, no. 5, pp. 1035–1040, 2020, doi: 10.25126/jtiik.202072379.