Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 1 (2021) 114 - 122 ISSN Media Elektronik: 2580-0760

Peningkatan Hasil Klasifikasi pada Algoritma *Random Forest* untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi

Gde Agung Brahmana Suryanegara[1], Adiwijaya[2], Mahendra Dwifebri Purbolaksono[3]
[1,2,3]Informatika, Fakultas Informatika, Universitas Telkom
[1]brahmasurya@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id, [3]mahendradp@telkomuniversity.ac.id

Abstract

Diabetes is a disease caused by high blood sugar in the body or beyond normal limits. Diabetics in Indonesia have experienced a significant increase, Basic Health Research states that diabetics in Indonesia were 6.9% to 8.5% increased from 2013 to 2018 with an estimated number of sufferers more than 16 million people. Therefore, it is necessary to have a technology that can detect diabetes with good performance, accurate level of analysis, so that diabetes can be treated early to reduce the number of sufferers, disabilities, and deaths. The different scale values for each attribute in Gula Karya Medika's data can complicate the classification process, for this reason the researcher uses two data normalization methods, namely min-max normalization, z-score normalization, and a method without data normalization with Random Forest (RF) as a classification method. Random Forest (RF) as a classification method has been tested in several previous studies. Moreover, this method is able to produce good performance with high accuracy. Based on the research results, the best accuracy is model 1 (Min-max normalization-RF) of 95.45%, followed by model 2 (Z-score normalization-RF) of 95%, and model 3 (without data normalization-RF) of 92%. From these results, it can be concluded that model 1 (Min-max normalization-RF) is better than the other two data normalization models and is able to increase the performance of classification Random Forest by 95.45%.

Keywords: diabetes, classification, min-max normalization, z-score normalization, random forest.

Abstrak

Diabetes merupakan salah satu penyakit yang disebabkan karena gula darah di dalam tubuh yang tinggi atau melampaui batas normal. Penderita diabetes di Indonesia mengalami peningkatan yang cukup signifikan, Riset Kesehatan Dasar menyebutkan penderita diabetes di Indonesia yang semula dari tahun 2013 sebesar 6,9% menjadi 8,5% di tahun 2018 dengan perkiraan jumlah penderita lebih dari 16 juta orang. Oleh karena itu, sangat diperlukan suatu teknologi yang dapat mendeteksi penyakit diabetes dengan kinerja yang baik, tingkat analisis akurat, sehingga penyakit diabetes dapat ditangani lebih awal untuk mengurangi jumlah penderita, kecacatan, dan kematian. Nilai skala yang berbeda tiap atribut pada data Gula Karya Medika dapat mempersulit proses klasifikasi, untuk itu peneliti menggunakan dua metode normalisasi data yaitu *Min-max normalization*, *Z-score normalization*, dan satu tanpa metode normalisasi data dengan *Random Forest* (RF) sebagai metode klasifikasi. *Random Forest* (RF) sebagai metode klasifikasi telah teruji di beberapa penelitian sebelumnya, metode ini mampu menghasilkan kinerja yang baik dengan akurasi yang tinggi. Berdasarkan hasil penelitian, akurasi terbaik dihasilkan model 1 (*Min-max normalization*-RF) sebesar 95.45%, model 2 (*Z-score normalization*-RF) sebesar 95%, dan model 3 (Tanpa normalisasi data-RF) sebesar 92%. Dari hasil tersebut disimpulkan bahwa model 1 (*Min-max normalization*-RF) lebih baik dibandingkan dua model normalisasi data lainya dan mampu meningkatkan performansi klasifikasi *Random Forest* sebesar 95.45%.

Kata kunci: *diabetes*, klasifikasi, *min-max normalization, z-score normalization, random forest.*

Diterima Redaksi: 19-01-2021 | Selesai Revisi: 05-02-2021 | Diterbitkan Online : 20-02-2021

1. Pendahuluan

Diabetes merupakan salah satu penyakit yang disebabkan karena gula darah di dalam tubuh yang tinggi atau melampaui batas normal. Berdasarkan data laporan Kementerian Kesehatan RI tahun 2018 melalui conference Suara Dunia Perangi Diabetes, Indonesia merupakan negara peringkat keenam sebagai penderita diabetes terbanyak di dunia dengan jumlah penderita diabetes usia 20-79 tahun sekitar 10,3 juta [1]. Riset Kesehatan Dasar juga menyebutkan penderita diabetes di Indonesia mengalami peningkatan, yaitu semula dari tahun 2013 sebesar 6,9% menjadi 8,5% di tahun 2018 dengan perkiraan jumlah penderita lebih dari 16 juta orang [1]. Penyakit diabetes dapat mengakibatkan komplikasi penyakit, yang tentunya sangat berbahaya terhadap penderita diabetes. Oleh karena itu, sangat diperlukanya suatu teknologi yang dapat mendeteksi penyakit diabetes dengan tingkat analisis yang akurat, sehingga penyakit diabetes dapat ditangani lebih awal untuk mengurangi jumlah penderita, kecacatan, dan kematian.

Beberapa tahun terakhir, penelitian terhadap penyakit diabetes sudah dilakukan dengan menggunakan berbagai macam metode klasifikasi untuk mendeteksi diabetes. Berikut beberapa penelitian yang terkait dengan pengujian data diabetes. Manimaran dan Vanita [2] mengusulkan metode *Decision Tree* dalam melakukan klasifikasi penyakit diabetes. Peneliti juga melakukan *preprocessing* data dan transformasi data seperti mengganti nilai yang hilang dan menormalisasikan data untuk meningkatkan hasil dan efisiensi dalam penambangan data. Pada penelitianya, peneliti menggunakan *cross validation* untuk membagi data ke dalam dua bagian data latih dan data uji dengan perbandingan 70:30. Untuk mengevaluasi model yang sudah dibangun peneliti menggunakan *confusion matrix* dengan hasil *accuracy* yang diperoleh dari data asli tanpa *cross validation* sebesar 83,5937% dan setelah menggunakan *cross validation* hasil *accuracy* klasifikasi didapatkan sebesar 85,0163%. Selanjutnya pada tahun 2020 Diniyal Amru Agatsa [3], membangun model klasifikasi pasien pengidap diabetes menggunakan metode *Support Vector Machine* pada data diabetes dan validasi model menggunakan *K-Fold Cross Validation* untuk membagi data menjadi k bagian dengan hasil akurasi yang diperoleh sebesar 77,92%. Selanjutnya Indrayanti tahun 2017 [4], peneliti menggunakan KNN sebagai metode klasifikasi untuk mengklasifikasi penyakit diabetes melitus, dengan hasil akurasi yang diperoleh sebesar 75,14% dengan nilai k=13 merupakan nilai k yang paling optimal. Selanjutnya Januar Adi Putra tahun 2016 [5], peneliti melakukan klasifikasi penyakit diabetes menggunakan metode penggabungan SVM dengan KNN, dengan hasil akurasi yang diperoleh sebesar 92%. Selanjutnya pada tahun 2019 Safial [6] mengusulkan membangun sistem klasifikasi dari *dataset* Pima Indian Diabetes dengan menggunakan pendekatan *Deep Learning*. Pada penelitianya, peneliti melakukan *data preparation* untuk memeriksa dataset apakah terdapat data yang hilang atau tidak dan membagi dataset menjadi dua bagian diantaranya data latih dan data uji menggunakan *k-fold cross-validation*. Untuk menguji dan melakukan analisis terhadap model *Deep Neural Network* yang dibangun peneliti menggunakan *confusion matrix*. Hasil akurasi klasifikasi yang diperoleh peneliti dengan menggunakan 5-*fold cross validation* sebesar 98,35%. Selanjutnya pada tahun 2017 Amit pandey [7] mengusulkan penelitian Analisis Perbandingan Algoritma KNN dengan Berbagai Teknik Normalisasi. Peneliti menggunakan algoritma KNN sebagai metode klasifikasi dan dua teknik normalisasi data yaitu normalisasi min-max dan normalisasi z-score untuk mengklasifikasikan dataset iris dan mengukur akurasi klasifikasi menggunakan metode *cross validation* menggunakan *R-Programing*. Algoritma KNN digunakan oleh peneliti karena telah banyak digunakan dalam *data mining* dan *machine learning* dengan hasil performa yang sangat baik. Akurasi rata-rata yang dihasilkan dengan normalisasi min-max lebih besar dibandingan normalisasi z-score dengan perbandingan 88.09%:78.56%, membuktikan bahwa min-max dapat meningkatkan hasil akurasi dibandingkan dengan metode lainya. Selanjutnya pada tahun 2017 Fadly Rahman [8], mengusulkan membangun sistem klasifikasi penyakit diabetes menggunakan metode *Bayesian Regularization Neural Network*. Peneliti menggunakan jumlah perbandingan 90% sebagai data latih dan 10% sebagai data uji dari *dataset* semula. Peneliti juga menyebutkan jumlah penggunaan neuron dalam *hidden layer* dapat mempengaruhi hasil akurasi dari proses klasifikasi. Hasil *accuracy* klasifikasi yang didapatkan peneliti sebesar 96,1%.

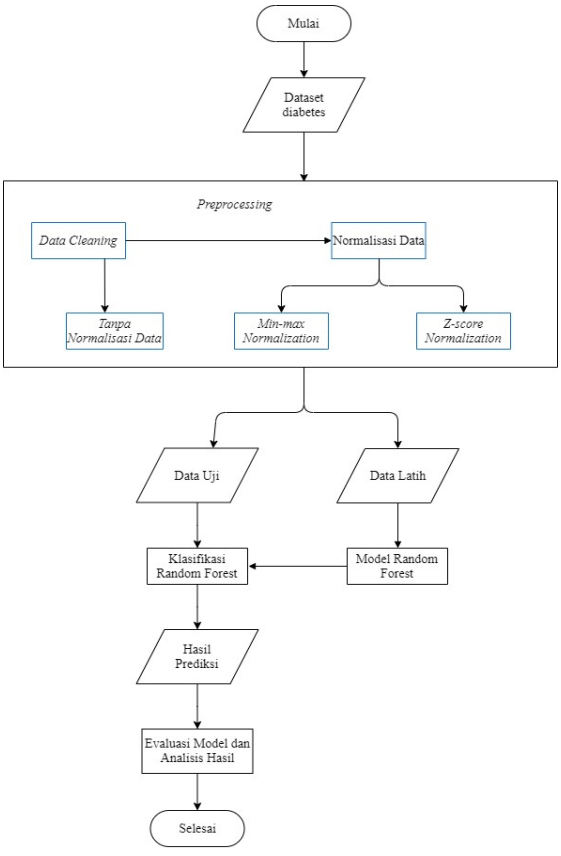
Dari beberapa penelitian yang sudah dilakukan, penelitian kali ini membangun suatu sistem klasifikasi untuk menganalisis dan memprediksi apakah seseorang sebagai penderita diabetes atau tidak berdasarkan *dataset* Gula Karya Medika. Pada tahap *preprocessing* dilakukan transformasi data dengan menghapus nilai yang hilang dan normalisasi data dengan tujuan membuat nilai setiap atribut berada pada rentang yang sama, dengan harapan dapat meningkatkan hasil dan efisiensi dari klasifikasi [2]. Selain itu, peneliti juga menggunakan *cross validation* untuk memisahkan *data training* dengan *data testing* yang diharapkan dapat meningkatkan hasil akurasi dari model. Penelitian ini mengusulkan tiga buah model, dua model dengan normalisasi data yaitu: *Min-max normalization*, *Z-score normalization*, dan satu model tanpa normalisasi data. Pada tahap klasifikasi dibandingkan hasil akurasi yang diperoleh *Random Forest* dari *Min-max normalization*, *Z-score normalization*, dan tanpa normalisasi data untuk mengetahui metode normalisasi data mana yang lebih optimal dan akurat dalam meningkatkan performansi klasifikasi penyakit diabetes. Dalam penelitian ini

DOI: https://doi.org/10.29207/resti.v5i1.2880

menggunakan beberapa skenario pengujian seperti jumlah *tree* pada *Random Forest*, jumlah K pada *cross validation*, dan metode normalisasi data yang digunakan.

2. Metode Penelitian

Penelitian ini dilakukan dengan membangun suatu sistem yang dapat mendeteksi penyakit diabetes ke dalam dua kelas yaitu *positive diabetes* dan *negative diabetes* menggunakan tiga buah model, dua model dengan normalisasi data yaitu: *Min-max normalization*, *Z-score normalization*, dan satu model tanpa normalisasi data dengan satu algoritma klasifikasi *Random Forest*. Tujuan menggunakan ketiga model tersebut untuk mengetahui metode normalisasi data yang mampu meningkatkan performansi kinerja dari algoritma *Random Forest* dalam mendeteksi penyakit diabetes. Adapun alur kerja sistem yang dibangun secara umum dalam melakukan klasifikasi diabetes, digambarkan pada gambar 1.



Gambar 1. Rancangan umum sistem klasifikasi diabetes.

2.1 *Dataset* Diabetes

Dataset diabetes yang digunakan pada penelitian ini dari Gula Karya Medika. *Dataset* ini memiliki 5 atribut dan 1 atribut kelas dengan jumlah data sebanyak 470 *record* yang terdiri dari 278 pria dan 192 wanita. Dalam dataset ini terdapat 290 orang sebagai penderita *positive diabetes* dan 180 orang *negative diabetes*. Adapun spesifikasi *dataset* diabetes yang digunakan pada penelitian dapat dilihat pada tabel 1.

Tabel 1. Spesifikasi dataset diabetes.

No	Atribut	Deskripsi	Tipe Data
1	Glucose	Kadar gula darah	Numerik
2	Gender	Jenis kelamin	Nominal
3	Blood Plessure	Tekanan darah	Numerik
4	BMI	Berat tubuh	Numerik
5	Usia	Umur	Numerik
6	Class	Positive diabetes (1) dan negative diabetes (0)	Nominal

2.2 *Preprocessing*

Terdapat berbagai permasalahan yang ditimbulkan dari pengolahan data antara lain terlalu banyak atribut, nilai data berada di *range* yang sangat jauh, *missing value*, ataupun format data yang tidak sesuai [9]. Hal tersebut tentunya dapat mengganggu dan menyebabkan hasil dari proses *data mining* yang kurang baik. Oleh karena itu, perlu tahap *preprocessing* data untuk mengatasi permasalahan. *Preprocessing* adalah suatu teknik untuk membuat data menjadi lebih mudah diproses atau digunakan dalam *data mining*. Tujuan dari *preprocessing* ini untuk membuat kualitas data yang baik, termasuk kelengkapan, konsistensi, ketepatan waktu dan meningkatkan hasil akurasi [10]. Adapun beberapa tahapan *preprocessing* yang dilakukan pada *dataset* diabetes Gula Karya Medika.

2.2.1 *Data Cleaning*

Pada *dataset* terdapat atribut yang memiliki *missing value*, sehingga perlu dilakukanya proses *data cleaning*. *Data cleaning* adalah proses menyiapkan data dengan menghapus atau mengisi nilai yang kosong untuk seluruh *dataset* dengan menggunakan rata-rata dari tiap kolom pada nilai yang kosong. Berikut Tabel 2 merupakan contoh keadaan awal *dataset* dari Gula Karya Medika yang terdapat nilai yang hilang.

Tabel 2. Keadaan awal *dataset*.

Glucose	Gender	Blood Plessure	BMI	Usia	Class
197	0	80	29.3	50	1
167	0		24.2	73	1
103	1	80	25	56	1
113	0	90	30.2	51	1
…	…	…	…	…	…

Berikut Tabel 3 merupakan keadaan data setelah dilakukanya *data cleaning* dengan menghapus baris masing-masing nilai yang kosong.

Tabel 3. Keadaan setelah data cleaning.

Glucose	Gender	Blood Plessure	BMI	Usia	Class
197	0	80	29.3	50	1
103	1	80	25	56	1
113	0	90	30.2	51	1
…	…	…	…	…	…

DOI: https://doi.org/10.29207/resti.v5i1.2880

2.2.2 Normalisasi Data

Data yang ada pada *dataset* terkadang memiliki suatu nilai dengan rentang yang tidak sama. Tentunya hal ini dapat mempengaruhi hasil pengukuran analisis data, sehingga perlunya suatu metode normalisasi data. Normalisasi data adalah proses membuat skala nilai atribut ke dalam rentang yang lebih kecil dengan bobot yang sama [10]. Skala nilai atribut data yang baru dapat membantu kinerja klasifikasi karena dapat menghapus fitur dengan noise yang tinggi dan relevansi yang rendah [11]. Terdapat banyak metode normalisasi data seperti *Min-max Normalization*, Z-score Normalization, dan *Decimal scaling*. Penelitian ini menggunakan dua metode normalisasi data, sebagai berikut.

a. *Min-max Normalization*

Min-max normalization adalah suatu metode yang melakukan transformasi linear dengan menggunakan nilai minimum dan maksimum yang menghasilkan keseimbangan antara data satu dengan yang lain pada rentang yang sama [12]. Metode ini untuk mencapai *konvergensi* membutuhkan waktu yang paling singkat dibandingkan metode lainya [7]. *Min-max normalization* dapat dihitung menggunakan rumus berikut:

$$n_i^1 = \text{new\_min}_A + \frac{n_i - \text{min}_A}{\text{maks}_A - \text{min}_A} (\text{new\_maks}_A - \text{new\_min}_A) \quad (1)$$

Hasil *min-max normalization* adalah n_i^1, data yang akan dinormalisasi n_i, nilai minimum pada atribut kolom min_A, nilai maksimum pada atribut kolom maks_A, nilai rentang maksimum 1 new_maks_A, dan nilai rentang minimum 0 adalah new_min_A. Berikut merupakan hasil data *Min-max normalization* dapat dilihat pada tabel 4.

Tabel 4. Data hasil *min-max normalization.*

Glucose	Gender	BP	BMI	Usia	Class
0.294	0	0.545	0.566	0.529	1
0.081	1	0.545	0.376	0.6	1
0.104	0	0.772	0.606	0.541	1
…	…	…	…	…	…

b. *Z-score normalization*

Z-score normalization adalah suatu metode normalisasi yang hasilnya didapatkan dari nilai rata-rata dan standar deviasi dari data [7]. Metode ini mempunyai nilai yang stabil terhadap *outlier* maupun adanya nilai yang lebih besar dari maks_A atau lebih kecil dari min_A [12]. *Z-score normalization* dapat dihitung menggunakan rumus berikut:

$$n_i^1 = \frac{n_i - \bar{A}}{\sigma_A} \quad (2)$$

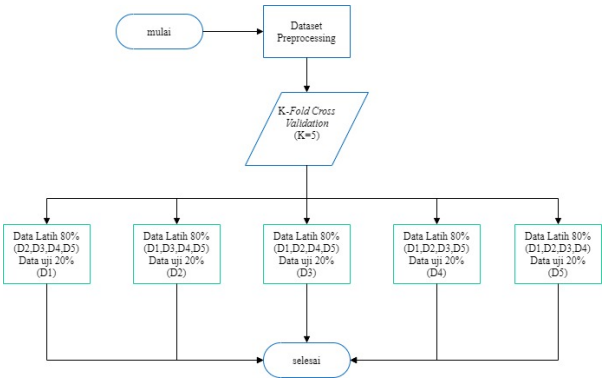
Hasil *Z-score normalization* adalah n_i^1, data yang akan dinormalisasi n_i, nilai rata-rata $\bar{A}$, dan standar deviasi adalah σ_A. Berikut merupakan hasil data *Z-score normalization* dapat dilihat pada tabel 5.

Tabel 5. Data hasil *Z-score normalization.*

Glucose	Gender	BP	BMI	Usia	Class
0.871	0	0.016	1.098	-0.427	1
-0.568	1	0.016	0.045	0.114	1
-0.414	0	1.181	1.318	-0.337	1
…	…	…	…	…	…

2.3 *Cross validation*

Cross validation adalah suatu metode dengan membagi himpunan dataset ke dalam dua bagian seperti data latih dan uji [4]. Dalam membagi himpunan data, istilah yang sering digunakan k-*Fold* yang terdiri dari beberapa bagian. Apabila menggunakan k=5 maka akan didapatkan 5 bagian himpunan data: D1, D2, D3, D4, dan D5 yang setiap himpunan data terdiri dari 4 bagian sebagai data latih dari rumus (k-1) dan 1 bagian sebagai data uji [12]. Himpunan data cross validation dengan k=5 dapat digambarkan sebagai berikut, Gambar 2.



Gambar 2. Skema cross validation.

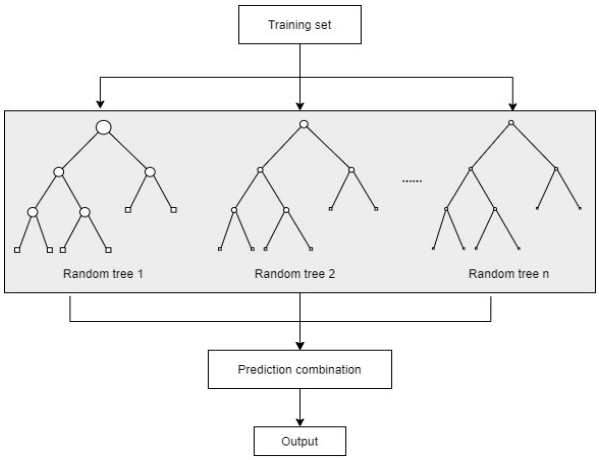
Dengan menggunakan *cross validation* dapat menghasilkan tingkat performa yang lebih stabil, dapat mengukur semua kualitas dari model klasifikasi yang dibangun, dan dapat meningkatkan hasil akurasi yang lebih akurat [3] [4] [12].

2.4 Klasifikasi Menggunakan *Random Forest*

Random Forest adalah salah satu jenis algoritma klasifikasi yang terdiri dari lebih satu pohon keputusan yang setiap pohon keputusan dibentuk bergantung pada nilai-nilai vector acak sampel secara independen dan identik didistribusikan yang sama untuk semua pohon [13]. *Random Forest* masuk ke dalam kelompok *Supervised Learning* yang dikembangkan oleh Leo Bremen. Metode ini merupakan salah satu metode klasifikasi yang sangat akurat digunakan dalam melakukan prediksi, bisa menangani inputan variabel yang sangat besar jumlahnya tanpa overfitting, dan membantu menghilangkan korelasi antara pohon keputusan seperti karakteristik *ensemble methods* [14]. Selain itu, *Random Forest* memiliki tingkat *error rate* yang lebih kecil pada data diabetes dibandingkan algoritma klasifikasi lainya dan memiliki kinerja yang

DOI: https://doi.org/10.29207/resti.v5i1.2880

baik dalam klasifikasi diabetes [15]. Berikut merupakan metodologi cara kerja *Random Forest* seperti gambar 3.



Gambar 3. Cara kerja *random forest*.

Dalam pembuatan pohon keputusan *random forest* menggunakan *random vector*. Adapun tahapan pseudocode dalam pembuatan *Random Forest* [16].

a. Pilih secara acak fitur "R" dari total fitur "m" dimana R<<m.
b. Di antara fitur "R", hitung simpul menggunakan titik perpecahan terbaik.
c. Membagi node menjadi simpul anak menggunakan split terbaik.
d. Ulangi langkah a hingga c hingga "1" jumlah node telah tercapai.
e. Bangun *forest* dengan mengulangi langkah a hingga d untuk jumlah "n" kali untuk membuat "n" jumlah pohon.

Keuntungan dengan menggunakan algoritma *Random Forest* sebagai metode dalam klasifikasi yaitu dalam menggunakan algoritma *Random Forest* masalah *overfitting* tidak akan pernah muncul dalam masalah klasifikasi, algoritma *Random Forest* dapat digunakan untuk regresi dan klasifikasi, dan *Random Forest* dapat digunakan untuk mengidentifikasi fitur yang paling penting untuk digunakan dari dataset pelatihan [17]. Klasifikasi dilakukan ketika semua data sudah siap untuk digunakan dalam *data mining*. Pada tahap ini menggunakan tiga buah model, dua model dengan normalisasi data yaitu: *Min-max normalization*, *Z-score normalization*, dan satu model tanpa normalisasi data dengan satu algoritma klasifikasi *Random Forest* untuk membangun dan menguji sebuah model. Adapun tahap-tahapan yang dilakukan dalam klasifikasi. Pertama data latih digunakan untuk membangun suatu model dengan proses mesin akan mempelajari terhadap dataset yang digunakan. Kedua model yang sudah dibangun akan diuji menggunakan data uji untuk mengklasifikasikan kelas yang belum diketahui kelasnya.

2.5 Evaluasi Model dan Analisis Hasil

Model klasifikasi yang sudah dibangun perlu dilakukan evaluasi untuk mengetahui hasil dan tingkat performa dari model klasifikasi tersebut dalam melakukan klasifikasi terhadap data uji yang ada. Untuk melakukan evaluasi terhadap model yang sudah dibangun dapat menggunakan *confusion matrix*. *Confusion matrix* adalah metode perhitungan dalam menganalisis kualitas model klasifikasi dalam mengenali tuple-tuple dari kelas yang ada [12]. Dalam perhitungan *confusion matrix* terdapat istilah-istilah TP, TN, FP, dan FN, antara lain *True Positive* (TP) adalah nilai yang benar positive diprediksi oleh model klasifikasi sesuai dengan kelas aktual yang sesungguhnya. *True Negative* (TN) adalah nilai yang benar negative diprediksi oleh model klasifikasi sesuai dengan kelas aktual yang sesungguhnya. *False Positive* (FP) adalah kelas aktual yang berlabelkan negative salah diberi label oleh model klasifikasi dengan hasil prediksi positive. *False Negative* (FN) adalah kelas aktual yang berlabelkan positive salah diberi label oleh model klasifikasi dengan hasil prediksi negative. Berdasarkan istilah tersebut, dapat digambarkan ke dalam *confusion matrix*.

Tabel 6. *Confusion matrix.*

Klasifikasi		Kelas hasil prediksi	
		Ya	Tidak
Kelas aktual	Ya	TP	FN
	Tidak	FP	TN

Dari tabel 6 *confusion matrix* tersebut dapat diukur dan dievaluasi tingkat performa model klasifikasi dengan menghitung akurasi. *Accuracy* adalah persentase data uji dapat diklasifikasikan dengan benar oleh model klasifikasi yang dibangun. Persamaan ini dapat dihitung dengan rumus berikut [12].

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (3)$$

Setelah hasil evaluasi model didapatkan, dilakukan analisis hasil terhadap akurasi yang diperoleh dari tiga buah model, dua model dengan normalisasi data yaitu: *Min-max normalization*, *Z-score normalization*, dan satu model tanpa normalisasi data dengan satu algoritma klasifikasi *Random Forest.* Analisis dilakukan dengan melihat hasil akurasi yang diperoleh *Random Forest* dengan *Min-max normalization*, *Z-score normalization*, dan tanpa normalisasi data untuk menentukan metode normalisasi data mana yang lebih optimal dan akurat dalam meningkatkan performansi akurasi klasifikasi penyakit diabetes.

3. Hasil dan Pembahasan

Penelitian ini menggunakan *dataset* diabetes dari Gula Karya Medika, untuk spesifikasinya dapat dilihat pada tabel 1. Atribut-atribut pada *dataset* diabetes dilakukan tahap *preprocessing data* yang dapat dilihat pada sub bab 2.2 untuk membuat data menjadi lebih mudah diproses atau digunakan dalam *data mining*. Selanjutnya

DOI: https://doi.org/10.29207/resti.v5i1.2880

data yang sudah dilakukan *preprocessing* dibagi menjadi dua bagian ke dalam data latih dan data uji menggunakan *cross validation*. Data latih digunakan untuk membangun suatu model dengan proses mesin mempelajari terhadap dataset yang digunakan. Selanjutnya data uji digunakan untuk mengklasifikasikan kelas yang belum diketahui kelasnya. Pada pengujian, peneliti menggunakan beberapa skenario pengujian seperti jumlah tree pada *Random Forest* (RF) dengan nilai *N tree* = [5, 10, 15] dan jumlah K pada *cross validation* dengan nilai K = [2, 3, 4, 5, 6, 7, 8, 9, 10]. Dalam menentukan akurasi sistem dari setiap iterasi K pada *cross validation*, menggunakan akurasi yang terbesar. Skenario pengujian ini dilakukan untuk memaksimalkan data yang digunakan, menghasilkan suatu sistem dengan kinerja yang baik dan tingkat analisis yang akurat. Skenario tersebut dikombinasikan pada tiga model yang digunakan dengan menghitung akurasi yang nantinya hasil dari akurasi tersebut dibandingkan terhadap masing-masing model, untuk mengetahui metode normalisasi data yang mampu meningkatkan performansi kinerja dari algoritma *Random Forest* dalam mendeteksi penyakit diabetes.

3.1 Hasil Pengujian Model 1 (*Min-max Normalization*-RF)

Hasil akurasi yang diperoleh menggunakan *confusion matrix* dari beberapa skenario pengujian terhadap data diabetes Gula Karya Medika sebagai berikut.

Tabel 7. Model 1 (*Min-max Normalization*-RF).

K	Akurasi		
	T=5	T=10	T=15
2	72.36%	72.36%	75.37%
3	81.06%	81.20%	82.70%
4	88.00%	83.83%	87.00%
5	85.00%	91.13%	88.75%
6	86.36%	87.87%	88.05%
7	89.28%	89.47%	**92.85%**
8	87.75%	90.00%	92.00%
9	**93.18%**	**95.45%**	90.90%
10	87.50%	92.50%	92.50%

Berdasarkan hasil pengujian model 1 (min max normalization) dengan algoritma *Random Forest*, seiring bertambahnya nilai K pada *cross validation* menghasilkan peningkatan akurasi yang signifikan, terlihat pada tabel 7 dari nilai K=2 dengan K=9 akurasi yang dihasilkan mengalami peningkatan sebesar 23%. Hal ini menunjukan bahwa nilai iterasi K pada *cross validation* dalam kasus diabetes mampu membuat sistem lebih banyak belajar terhadap datasetnya sehingga menghasilkan model yang memiliki kinerja dan performa yang baik. Pada tabel 7 akurasi dengan jumlah *tree*=5 tidak ada perbedaan yang signifikan dengan jumlah *tree*=10 dan 15. Dimana hasil akurasi jumlah *tree*=10 lebih besar dibandingkan dengan jumlah tree=15, hal ini memiliki artian bahwa semakin besar nilai jumlah *tree* (*N tree*) yang dibangun pada model *Random Forest* tidak menjamin model menghasilkan akurasi yang optimal dan akurat, sebaliknya hanya meningkatkan waktu yang lama terhadap eksekusi program [18].

Pada hasil penelitian ini dari skenario pengujian dengan nilai K = [2, 3, 4, 5, 6, 7, 8, 9, 10] dan nilai *N tree* (T) = [5, 10, 15] akurasi model ditentukan berdasarkan nilai tertinggi pada iterasi K *cross validation*¸ sehingga dapat disimpulkan bahwa nilai K=9 dan *N tree* (T)=10 menghasilkan kinerja yang baik dengan akurasi tertinggi 95.45 dari jumlah data 44 sebagai *data testing* dan 398 sebagai *data training*.

3.2 Hasil Pengujian Model 2 (*Z-score Normalization*-RF)

Hasil akurasi yang diperoleh menggunakan *confusion matrix* dari beberapa skenario pengujian terhadap data diabetes Gula Karya Medika sebagai berikut.

Tabel 8. Model 2 (*Z-score Normalization*-RF).

K	Akurasi		
	T=5	T=10	T=15
2	75.87%	70.85%	76.88%
3	79.69%	76.69%	84.96%
4	83.00%	82.00%	85.00%
5	82.50%	87.34%	88.60%
6	86.36%	89.39%	89.39%
7	85.96%	89.47%	89.28%
8	86.00%	88.00%	89.79%
9	86.66%	90.90%	93.18%
10	**87.50%**	**92.50%**	**95.00%**

Berdasarkan hasil pengujian model 2 (*z-score normalization*) dengan algoritma *Random Forest*, seiring bertambahnya nilai K pada *cross validation* menghasilkan peningkatan akurasi yang signifikan, terlihat pada tabel 8 akurasi terendah sebesar 70.85% dan akurasi tertinggi sebesar 95%, dimana akurasi yang dihasilkan mengalami peningkatan sebesar 24.15%. Hal ini menunjukan bahwa nilai iterasi K pada *cross validation* dalam kasus diabetes mampu membuat sistem lebih banyak belajar terhadap datasetnya sehingga menghasilkan model yang memiliki kinerja dan performa yang baik. Pada tabel 8 berbanding terbalik dengan hasil pengujian tabel 7, dimana nilai jumlah *tree* yang semakin besar menghasilkan peningkatan akurasi yang signifikan. Dapat dilihat hasil akurasi dengan jumlah *tree*=5 dan *tree*=10 lebih kecil dibandingkan dengan jumlah *tree*=15, hal ini memiliki artian bahwa jumlah *tree* yang dibangun pada model *Random Forest* dapat meningkatkan performansi kinerja dalam melakukan klasifikasi penyakit diabetes.

Pada hasil penelitian ini dari skenario pengujian dengan nilai K = [2, 3, 4, 5, 6, 7, 8, 9, 10] dan nilai *N tree* (T) = [5, 10, 15] akurasi model ditentukan berdasarkan nilai tertinggi pada iterasi K *cross validation*¸ sehingga dapat disimpulkan bahwa nilai K=10 dan *N tree* (T) = 15 menghasilkan kinerja yang baik dengan akurasi tertinggi 95.00 dari jumlah data 40 sebagai *data test* dan 342 sebagai *data training*.

DOI: https://doi.org/10.29207/resti.v5i1.2880

3.3 Hasil Pengujian Model 3 (Tanpa normalisasi data-RF)

Hasil akurasi yang diperoleh menggunakan *confusion matrix* dari beberapa skenario pengujian terhadap data diabetes Gula Karya Medika sebagai berikut.

Tabel 9. Model 3 (Tanpa normalisasi data-RF).

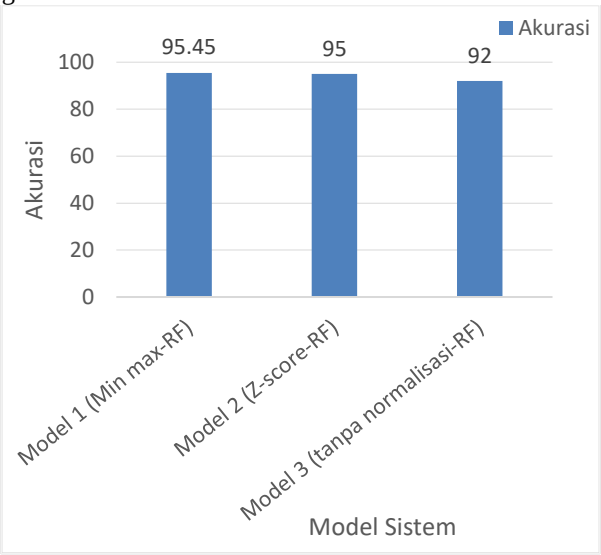
K	Akurasi		
	T=5	T=10	T=15
2	77.38%	70.35%	75.37%
3	81.06%	77.44%	78.94%
4	82.82%	83.83%	87.00%
5	83.54%	87.50%	88.60%
6	86.36%	89.39%	88.05%
7	85.96%	89.47%	89.28%
8	84.00%	87.75%	**92.00%**
9	**90.90%**	88.63%	88.63%
10	90.00%	**90.00%**	89.74%

Berdasarkan hasil pengujian model 3 (tanpa normalisasi data) dengan algoritma *Random Forest*, seiring bertambahnya nilai K pada *cross validation* menghasilkan peningkatan akurasi yang signifikan, terlihat pada tabel 9 akurasi terendah sebesar 70.35% dan akurasi tertinggi sebesar 92%, dimana akurasi yang dihasilkan mengalami peningkatan sebesar 21.65%. Hal ini menunjukan bahwa nilai iterasi K pada *cross validation* dalam kasus diabetes mampu membuat sistem lebih banyak belajar terhadap datasetnya sehingga menghasilkan model yang memiliki kinerja dan performa yang baik. Pada tabel 9 berbanding terbalik dengan hasil pengujian tabel 7, dimana nilai jumlah *tree* yang semakin besar menghasilkan peningkatan akurasi yang signifikan. Dapat dilihat hasil akurasi dengan jumlah *tree*=5 dan *tree*=10 lebih kecil dibandingkan dengan jumlah *tree*=15, hal ini memiliki artian bahwa jumlah *tree* yang dibangun pada model *Random Forest* dapat meningkatkan performansi kinerja dalam melakukan klasifikasi penyakit diabetes. Namun model 3 (tanpa normalisasi data) dengan algoritma *Random Forest* masih belum bisa menghasilkan akurasi yang maksimal pada data diabetes Gula Karya Medika. Salah satu penyebabnya karena nilai tiap atribut yang tidak normal dengan memiliki rentang yang tidak sama satu sama lain, noise yang tinggi, dan relevansi yang rendah, sehingga mempersulit dan mempengaruhi hasil pengukuran model *Random Forest* dalam melakukan proses klasifikasi penyakit diabetes [11].

Pada hasil penelitian ini dari skenario pengujian dengan nilai K = [2, 3, 4, 5, 6, 7, 8, 9, 10] dan nilai *N tree* (T) = [5, 10, 15] akurasi model ditentukan berdasarkan nilai tertinggi pada iterasi K *cross validation*¸ sehingga dapat disimpulkan bahwa nilai K=8 dan *N tree* (T)=15 menghasilkan kinerja yang baik dengan akurasi tertinggi 92.00% dari jumlah data 50 sebagai *data test* dan 332 sebagai *data training*.

3.4 Analisis Perbandingan Performa Model dalam Klasifikasi Diabetes

Berdasarkan hasil pengujian yang telah dilakukan seiring bertambahnya nilai K *cross validation*, jumlah *tree Random Forest* (RF), dan metode normalisasi data yang digunakan memiliki hasil akurasi yang berbeda-beda dari masing-masing model yang dibangun dalam melakukan proses klasifikasi penyakit diabetes. Pada beberapa percobaan jumlah *tree* yang semakin besar tidak menjamin model menghasilkan akurasi yang optimal dan akurat, sebaliknya hanya meningkatkan waktu yang lama terhadap eksekusi program [18]. Penggunaan teknik C*ross validation* dalam model membuat model klasifikasi lebih banyak belajar terhadap data latih secara bergantian pada dataset, sehingga dapat menghasilkan model yang memiliki kinerja dan performa yang baik. Penelitian ini melakukan pengujian terhadap tiga model yang dibangun yaitu model 1 (*Min max normalization*-RF), model 2 (*Z-score normalization*-RF), dan model 3 (Tanpa normalisasi data-RF) dengan tujuan untuk mengetahui metode normalisasi data yang mampu meningkatkan performansi kinerja dari algoritma *Random Forest* dalam mendeteksi penyakit diabetes. Berikut merupakan perbandingan hasil akurasi yang diperoleh dari masing-masing model dapat dilihat pada gambar 4.



Gambar 4. Perbandingan performa model.

Gambar 4 menunjukan bahwa algoritma *Random Forest* menghasilkan kinerja yang baik ketika data dilakukan normalisasi data, dibandingkan tidak dilakukan normalisasi data. Hal ini menunjukan bahwa proses dari normalisasi data pada data Gula Karya Medika dapat membuat skala nilai atribut ke dalam rentang yang lebih kecil dengan bobot yang sama, sehingga memudahkan, meningkatkan kualitas data, dan meningkatkan efisiensi sistem dalam proses *learning* dengan kesalahan minimum terhadap *training model.* Menggunakan normalisasi data juga dapat mempercepat waktu

DOI: https://doi.org/10.29207/resti.v5i1.2880

learning dari *data training* untuk setiap fitur dalam skala yang sama, dibandingkan tanpa normalisasi data yang pada umumnya nilai berada pada skala yang berbeda dan ukuran ruang fitur yang tinggi [19].

Berdasarkan hasil akurasi yang diperoleh menggunakan *confusion matrix* terhadap data diabetes Gula Karya Medika, dari ketiga model yang digunakan dalam penelitian, dua model menghasilkan kinerja yang baik dengan menggunakan normalisasi data, dibandingkan model yang tidak menggunakan normalisasi data. Dalam penelitian ini model 1 (*Min-max normalization*-RF) dan model 2 (*Z-score normalization*-RF) menghasilkan akurasi yang lebih tinggi dibandingkan model 3 (Tanpa normalisasi data-RF). Pada pengujian tersebut model 1 (*Min max normalization*-RF) menghasilkan akurasi sebesar 95.45%, model 2 (*Z score normalization*-RF) menghasilkan akurasi sebesar 95%, dan model 3 (Tanpa normalisasi data-RF) menghasilkan akurasi sebesar 92%. Nilai akurasi memiliki artian bahwa model 1 (*min-max normalization*-RF) sebagai model klasifikasi mampu mengklasifikasikan data uji yang belum diketahui kelasnya lebih akurat dibandingkan dua model lainya. Sehingga disimpulkan model 1 (*min-max normalization*-RF) dengan skenario pengujian *cross validation* K=9 dan jumlah tree (T)=10 sebagai model dengan kinerja yang baik, dengan akurasi tertinggi, dan mampu meningkatkan performansi hasil klasifikasi algoritma *random forest* dalam mendeteksi penyakit diabetes.

4. Kesimpulan

Berdasarkan penelitian yang sudah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* mampu mengklasifikasikan data diabetes Gula Karya Medika dengan performansi yang baik jika data dilakukan normalisasi data. Hal ini menunjukan bahwa proses dari normalisasi data dapat mengubah nilai atribut yang memiliki rentang terlalu jauh ke dalam rentang tertentu dari setiap atribut, sehingga memudahkan dan meningkatkan efisiensi sistem dalam proses *learning* dengan kesalahan minimum terhadap *training model*. Selain itu nilai K *cross validation* dan jumlah *tree* pada *Random Forest* memiliki pengaruh yang signifikan terhadap performansi suatu model *Random Forest* dalam melakukan proses klasifikasi penyakit diabetes. Dari ketiga model yang digunakan dalam penelitian, dua model menghasilkan kinerja yang baik dengan menggunakan normalisasi data, dibandingkan model yang tidak menggunakan normalisasi data. Model 1 (min-max normalization-RF) memperoleh akurasi tertinggi sebesar 95.45%, mengungguli dua model lainya yaitu model 2 (Z-score normalization-RF) dengan akurasi sebesar 95% dan model 3 (Tanpa normalisasi data-RF) dengan akurasi sebesar 92%. Dengan artian bahwa model 1 (Min-max normalization-RF) sebagai model klasifikasi mampu mengklasifikasikan data uji yang belum diketahui kelasnya lebih akurat dibandingkan dua model lainya dan sebagai model yang mampu meningkatkan performansi hasil klasifikasi algoritma *Random Forest* dalam mendeteksi penyakit diabetes.

Daftar Rujukan

[1] Kementerian Kesehatan Republik Indonesia, 2018. Cegah, Cegah, dan Cegah: Suara Dunia Perangi Diabetes. [Online] (Update 13 Dec 2018). Tersedia di: http://p2ptm.kemkes.go.id/kegiatan-p2ptm/pusat-/cegah-cegah-dan-cegah-suara-dunia-perangi-diabetes [Accessed 6 Juni 2020]

[2] Manimaran, R. and Vanitha, Dr. M, 2017. Novel Approach to Prediction of Diabetes using Classification Mining Algorithm. *International Journal of Innovative Research in Science, Engineering and Technology*, 6 (7), pp. 14481–14487. doi: 10.15680/IJIRSET.2017.0607266.

[3] Agatsa, D. A., Rismala, R., and Wisesty, U.N, 2020. Klasifikasi Pasien Pengidap Diabetes menggunakan Metode Support Vector Machine. *Journal of Telkom University*, pp. 1–9.

[4] Indrayanti, Sugianti, D., and AL Karomi, M. A., 2017. Optimasi Parameter K pada Algoritma K-Nearest Neighbour untuk Klasifikasi Penyakit Diabetes Mellitus. *Jurnal Neliti*, 14 (4), pp. 823–829.

[5] Putra, J. A. and Akbar, A. L., 2016. Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Penggabungan Metode Support Vector Machine dan K-Nearest Neighbour. *Informatics J. UNEJ*, 1 (2), pp. 47–52.

[6] Ayon, S. I. and Islam, M. M., 2019. Diabetes Prediction: A Deep Learning Approach. *International Journal of Information Engineering and Electronic Business*, 2, pp. 21–27.

[7] Pandey, A. and Jain, A., 2017. Comparative Analysis of KNN Algorithm using Various Normalization Techniques. *I.J. Computer Network and Information Security*, 11, pp. 36–42. doi: 10.5815/ijcnis.2017.11.04.

[8] Rahman, M. F., Darmawidjadja, M. I., and Alamsah, D., 2017. Klasifikasi untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). *Journal of Garuda*, 11 (1), pp. 36–45.

[9] Chairunisa, R., Adiwijaya, and Astuti, W., 2020. Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray. *Jurnal RESTI*, 4(5), pp. 805–812. doi: https://doi.org/10.29207/resti.v4i5.2083.

[10] Han, J., Kamber, M., and Pei, J., 2011. *Data Mining Concepts and Techniques*. (3rd ed.). USA: Morgan Kaufmann.

[11] Khoirunnisa, A. and Rohmawati A., A., 2019. Implementing Principal Component Analysis and Multinomial Logit for Cancer Detection based on Microarray Data Classification. *In 2019 7th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–6. doi: 10.1109/ICoICT.2019.8835320.

[12] Suyanto, 2018. *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung.

[13] Breiman, L., 2011. *Random Forests*. Netherlands: Kluwer Academic Publishers.

[14] Nuklianggraita, T. N., Adiwijaya, and Aditsania, A., 2020. On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier. *Jurnal INFOTEL*, 12 (3), pp. 89–96. doi: https://doi.org/10.20895/infotel.v12i3.485.

[15] Benbelkacem, S. and Atmani, B., 2019. Random Forests for Diabetes Diagnosis. *2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–4. doi: 10.1109/ICCISci.2019.8716405.

[16] VijiyaKumar, K., 2019. Random Forest Algorithm for the Prediction of Diabetes. *Proceeding of International Conference on Systems Computation Automation and Networking 2019*, pp. 1–5. doi: 10.1109/ICSCAN.2019.8878802.

[17] Polamuri, S., 2017. How The Random Forest Algorithm Works in Machine Learning. [Online] (Update 22 May 2017). Tersedia di: https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/ .

[18] Agusta, Z. P. and Adiwijaya., 2019. Modified balanced random forest for improving imbalanced data prediction. *International Journal of Advances in Intelligent Informatics*, 5 (1), pp. 58–65.

[19] Singh, D. and Singh, B., 2019. Investigating the impact of data normalization on classification performance. *Applied Soft Computing Journal*, pp. 1568–4946. doi: https://doi.org/10.1016/j.asoc.2019.105524.

DOI: https://doi.org/10.29207/resti.v5i1.2880