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. 2 (2021) 393 - 399 ISSN Media Elektronik: 2580-0760

Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes

Mahendra Dwifebri Purbolaksono[1], Muhammad Irvan Tantowi[2], Adnan Imam Hidayat[3], Adiwijaya[4]
[1,2,3,4]Informatika, Fakultas Informatika, Universitas Telkom
[1]mahendradp@telkomuniversity.ac.id, [2]irvantantowi@students.telkomuniversity.ac.id,
[3]hidayatadnanimam@students.telkomuniversity.ac.id, [4]adiwijaya@telkomuniversity.ac.id

Abstract

Diabetes (diabetes) was a metabolic disorder caused by high levels of sugar in the blood caused by disorders of the pancreas and insulin. According to data from the Ministry of Health of the Republic of Indonesia, Diabetes was the third-largest cause of death in Indonesia with a percentage of 6.7%. The high rate of death from diabetes encouraged this study, with the aim of early detection. This research used a Machine Learning approach to classify the data. In this paper, a comparison of Support Vector Machine (SVM) and Modified Balanced Random Forest (MBRF) was discussed for classifying diabetes patient data. Both methods were chosen because it was proven in previous studies to get high accuracy, so that the two methods are compared to find the best classification model. Several preprocessing methods were used to prepare the data for the classification process. The entire combination of preprocessing steps will be carried out on the two classification methods to produce the same dataset. The evaluation was carried out using the Confusion Matrix method. Based on the experimental results in the process of testing the system being built, the maximum performance results were 87.94% using SVM and 97.8% using MBRF.

Keywords: Diabetes, Machine Learning, Supervised Learning, Support Vector Machine, Modified Balanced Random Forest

Abstrak

Diabetes (kencing manis) merupakan suatu kelainan metabolik yang disebabkan oleh tingginya tingkat kandungan gula dalam darah yang diakibatkan oleh gangguan pada pankreas dan insulin. Menurut data dari Kementrian Kesehatan Republik Indonesia, Diabetes merupakan penyebab kematian terbesar nomor 3 di Indonesia dengan persentase sebesar 6,7%. Tingginya tingkat kematian akibat diabetes mendorong dilakukan penelitian ini, dengan tujuan untuk deteksi dini. Pada penilitian ini akan menggunakan pendekatan Machine Learning untuk melakukan klasifikasi datanya. Dalam makalah ini, dibahas perbandingan Suport Vector Machine (SVM) dan Modified Balanced Random Forest (MBRF) untuk melakukan klasifikasi data pasien diabetes. Kedua metode dipilih karena terbukti pada penelitian sebelumnya mendapatkan akurasi yang tinggi, sehingga kedua metode tersebut dibandingkan untuk mencari model klasifikasi yang terbaik. Beberapa metode preprocessing dilakukan untuk mempersiapkan data agar dapat di lakukan proses klasifikasi. Seluruh kombinasi tahapan dari preprocessing akan dilakukan terhadap kedua metode klasifikasi untuk menghasilkan dataset yang sama juga. Evaluasi dilakukan menggunakan metode Confusion Matrix Berdasarkan hasil eksperimen dalam proses pengujian sistem yang dibangun, diperoleh hasil performasi maksimum 87,94% dengan menggunakan SVM dan 97,8% dengan menggunakan MBRF.

Kata kunci: Diabetes, Machine Learning, Supervised Learning, Support Vector Machine, Modified Balanced Random Forest

1. Pendahuluan

Dewasa ini banyak orang yang bekerja keras tanpa memikirkan diri sendiri. Hal tersebut menyebabkan pola hidup yang tidak sehat tanpa adanya olahraga bahkan mengkonsumsi makanan atau minuman instant ataupun cepat saji. Pola hidup yang buruk tersebut dapat menyebabkan kesehatan tubuh semakin menurun dan juga dapat mengakibatkan penyakit diabetes. Diabetes adalah sebuah penyakit di mana kandungan kadar gula dalam darah menjadi tinggi sehingga tubuh tidak mampu mengolah kadar gula tersebut [1]. Diabetes yang disebabkan oleh gangguan metabolik, terjadi karena pankreas tidak menghasilkan cukup insulin (hormone yang mengatur gula darah) atau tubuh tidak menggunakan insulin yang diproduksi secara efektif, sehingga menyebabkan tingkat glukosa diatas normal

Diterima Redaksi: 10-03-2021 | Selesai Revisi: 27-03-2021 | Diterbitkan Online: 30-04-2021

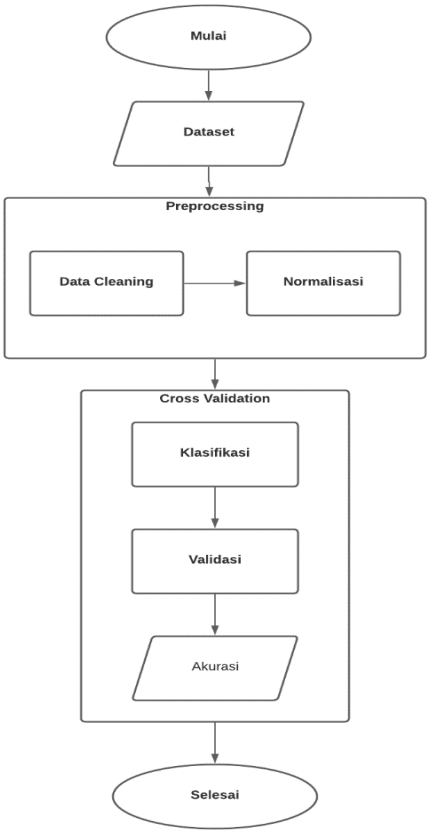
atau disebut hiperglikemia [1]. Diabetes dapat menyebabkan komplikasi seperti penyakit jantung koroner, stroke dan penyakit vaskular perifer, penyakit ginjal tahap akhir (ESRD), retinopati dan neuropati [2]. Secara global, diperkirakan 422 juta orang dewasa hidup dengan diabetes dan 1,5 juta diantaranya meninggal dunia pada tahun 2014[3]. Di Indonesia sendiri diabetes merupakan penyebab kematian terbesar nomor 3 dengan persentase sebesar 6,7 %, setelah Stroke 21,1 % dan penyakit Jantung Koroner 12,9 % pada tahun 2016 saja dan meningkat setiap tahunnya [1].

Berdasarkan data tersebut dapat dikatakan bahwa penyakit diabetes merupakan penyakit yang memiliki penderita yang tinggi di Indonesia bahkan di seluruh dunia. Seiring dengan berkembangnya teknologi saat ini metode machine learning telah banyak digunakan dalam segala bidang tidak terkecuali dalam bidang medis. Penelitian sebelumnya dilakukan klasifikasi dengan data diabetes menggunakan Neural Network menghasilkan akurasi 86.26%[4]. Selain itu juga terdapat penelitian lainnya yang menggunakan dataset diabetes dari mldata.org dengan metode J48 Decision Tree yang di mana menghasilkan akurasi sebesar 73,82% [5]. Adapun penilitan yang dilakukan pada tahun 2021 ini sudah dilakukan penelitian dengan topik deteksi diabetes menggunakan metode pengembangan Decision Tree yaitu Random Forest (RF) yang menghasilkan akurasi sebesar 95.45%[6]. Penelitian yang dilakukan oleh Diniyal Amru pada tahun 2019 dengan melakukan klasifikasi pada data pasien diabetes menggunakan metode Support Vector Machine (SVM) dalam sistem klasifikasi ini menghasilkan tingkat akurasi sebesar 77,92 % [7]. Penelitian ini juga menggunakan metode K-Fold Validation untuk mengantisipasi kemungkinan perolehan model dengan akurasi yang lebih baik. Pada bidang medis yang lain, terdapat penlitian dalam mendeteksi penyakit kanker menggunakan SVM dengan rata-rata akurasi 91,26% [8]. Sedangkan dengan topik yang sama Random Forest berhasil mendapatkan rata-rata akurasi sebesar 96.07% [9]. Kedua pendekatan tersebut diyakini dapat meningkatkan akurasi pada penlitian ini.

Berdasarkan penelitan-penelitan diatas, penelitian ini bertujuan untuk mencari performansi klasifikasi yang terbaik pada dataset diabetes dari Gula Karya Medika. Beberapa kombinasi tahapan akan dilakukan dari preprocessing sampai dengan proses klasifikasi. Preprocessing atau tahap persiapan data harus dilakukan untuk mengatasi data yang tidak lengkap atau missing value (nilai yang hilang). Masalah missing value akan pembersihan data (data cleaning) dengan dua metode yaitu mengganti nilai dengan rata-rata pada kelas yang sama (replace missing value) atau menghapus data yang terdapat missing value tersebut (drop missing value). Selain itu normalisasi data dilakukan untuk merubah nilai pada setiap atribut pada rentang yang sama. Normalisasi dilakukan menggunakan dua model yaitu Min-Max Normalization atau Z-Score Normalization. Pada tahap klasifikasi akan terdapat dua model juga. Model pertama yang akan menggunakan metode klasifikasi SVM, dikarenakan SVM memiliki akurasi yang cukup bagus dibeberapa penelitian dengan kasus yang sama. Pada metode berikutnya klasifikasi akan menggunakan MBRF. MBRF merupakan pengembangan dari algoritma RF. MBRF tidak hanya memiliki kemampuan untuk meningkatkan akurasi namun juga mengurangi kopleksitas waktu proses [10]. Hal ini akan menangani kekurangan yang dilakukan pada penelitian sebelumnya [6] dimana menggunakan dataset yang sama dan menggunakan Random Forest saja.

2. Metode Penelitian

Sistem yang dibangun pada penelitian ini terdiri dari beberapa proses. Adapun bagan metode penelitian yang dibangun ditunjukkan oleh Gambar 1.



Gambar 1. Desain Metode Penelitian

Penelitian ini mengusulkan rancangan sistem yang dapat melakukan klasifikasi terhadap data diabetes. Dataset yang digunakan telah dijelaskan pada bagian sebelumnya, pada Gambar 1 menunjukkan alur penelitian yang diusulkan yang terdiri dari beberapa

DOI: https://doi.org/10.29207/resti.v5i2.3008

modul yaitu preprocessing, data split dengan menggunkan K-Fold Cross Validation dan klasifikasi menggunakan perbandingan algoritma MBRF dan SVM serta evaluasi yang menggunkan Confusion Matrix. Pada tahapan data cleaning, normalisasi, k-fold dan klasifikasi akan dilakukan beberapa kombinasi satu sama lain yang akan dijelaskan pada skenario pengujian.

2.1. Dataset

Dalam penelitian ini akan menggunakan dataset diabetes yang didapatkan dari Gula Karya Medika. Atribut-atribut yang ada didalamnya adalah atribut yang digunakan pada proses yang biasa dilakukan di laboratorium untuk mendeteksi pasien diabetes atau tidak. Pada dataset ini terdapat 470 record (pasien) yang didalamnya terdapat 290 record yang dinyatakan pasien positif diabetes dan 180 record yang dinyatakan pasien negatif diabetes. Rincian atribut data dapat dilihat pada tabel 1.

Tabel 1. Tipe Data Dataset

Atribut	Tipe Data	Karakteristik
Glucose	Numerik	Atribut
Gender	Nominal	Atribut
Blood Plessure	Numerik	Atribut
BMI	Numerik	Atribut
Usia	Numerik	Atribut
Diabetes	Biner (Ya/Tidak)	Kelas

Dataset ini memiliki 5 atribut dan 1 atribut kelas. Data tersebut diambil oleh Klinik Karya Medika selama satu tahun padatahun 2019-2020. Beberapa data masih mengandung nilai-nilai yang hilang (*missing value*) pada setiap atribut, untuk itu tahap preprocessing sangat penting untuk mengatasi permasalahan data tersebut.

2.2. Preprocessing

Preprocessing adalah sebuah langkah penting dalam proses penambangan data. Data yang akan digunakan dalam proses penambangan data tidak selalu dalam kondisi terbaik untuk diproses. Ada kalanya dalam data tersebut terdapat beberapa masalah yang nantinya dapat mempengaruhi hasil yang diberikan dari proses penambangan itu sendiri seperti terdapat nilai yang hilang, data yang berlebihan, outliner, atau format data yang tidak sesuai dengan sistem. Oleh karena itu untuk mengatasi masalah tersebut perlu dilakukan tahap preprocessing. Preprocessing adalah salah satu langkah dalam menghilangkan masalah yang dapat mengganggu hasil dari pada proses klasifikasi data.

2.2.1 *Cleaning Data*

Tahap pertama yang akan dilakukan dalam preprocessing sistem ini adalah melakukan *Data Cleaning* yaitu apabila terdapat data yang kosong seperti pada atribut BMI dan diberi tanda tanya (?) pada Tabel 2.

Cara yang pertama untuk mengatasi masalah tersebut data kosong dapat diganti dengan mengisi nilai dari data yang kosong tersebut dengan nilai rata-rata dari kelas data yang sama. Tabel 3 adalah contoh dari data yang sudah diganti (*replace missing value*).

Tabel 2. Contoh Data dengan Data Kosong

Glucose	Gender	Blood Preasure	BMI	Usia	Kelas
157	1	80	21,6	49	1
130	1	88	**?**	50	1
115	0	76	36,8	61	1
99	0	86	29.4	74	0

Tabel 3. Contoh Data dengan *Replace Missing Value*

Glucose	Gender	Blood Preasure	BMI	Usia	Kelas
157	1	80	21,6	49	1
130	1	88	(21,6+36,8) / 2 = 29,2	50	1
115	0	76	36,8	61	1
99	0	86	29.4	74	0

Cara yang kedua data yang kosong bisa langsung dihapus. Tabel 4 adalah contoh dari data yang sudah dihapus (*drop missing value*).

Tabel 4.Contoh Data dengan *Drop Missing Value*

Glucose	Gender	Blood Preasure	BMI	Usia	Kelas
157	1	80	21,6	49	1
115	0	76	36,8	61	1
99	0	86	29.4	74	0

Setelah proses data cleaning selesai maka akan dilakukan normalisasi data agar datanya lebih mudah digunakan ketika proses klasifikasi.

2.2.1 *Normalization Data*

Normalisasi Data merupakan proses dilakukan transformasi sebuah atribut numerik yang di skalakan ke dalam sebuah bentuk lebih sederhana seperti 0 sampai 1. Ada beberapa metode yang dapat digunakan dalam melakukan normalisasi data, diantaranya:

- Min-Max Normalization
  Dalam metode ini data akan di transformasikan secara linear dari suatu nilai menjadi nilai baru lainnya [11], rumusnya sebagai berikut
  $$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (1)$$
- Z-Score Normalization
  Metode normalisasi ini menggunakan rata-rata dan standar deviasi untuk melakukan normalisasi setiap input[12], rumusnya sebagai berikut:
  $$v' = \frac{v - \bar{A}}{\sigma_A} \quad (2)$$
  Di mana ¯A merupakan rata-rata dan σ_A adalah standar deviasi.

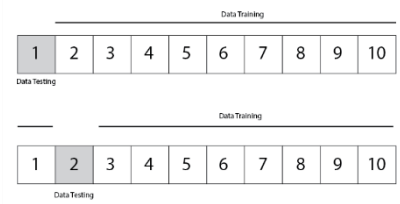
DOI: https://doi.org/10.29207/resti.v5i2.3008

Data yang telah dilakukan proses normalisasi akan berkisar antara 0 sampai dengan 1. Dengan begitu, metode klasifikasi dapat membacanya dengan range yang sama pada setiap atributnya.

2.3. *Cross-validation*

Cross-validasi (*Cross Validation*) atau validasi silang adalah teknik yang digunakan untuk prediksi akurasi sebuah model machine learning. Tujuan dari metode ini adalah memberikan validasi akurasi maksimal dari perputaran data uji dan data latih. K-Fold Cross Validation adalah salah satu metode validasi silang yang berfungsi untuk mengetahui rata-rata tingkat keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak [13].

Metode ini memecah data menjadi K bagian, di mana masing bagiannya memiliki jumlah data yang seimbang, data dibagi 2 bagian di mana data pertama disimpan sebagai data testing untuk validasi dan satu bagiannya digunakan untuk data training. Satu K data yang digunakan sebagai data testing digantikan dengan bagian lain dan terus dilakukan hal yang sama sampai semua bagian data sudah diberlakukan sebagai data testing. Berikut Ilustrasi dari *Cross Validation* dengan menggunakan K-Fold.



Gambar 2. Simulasi *Cross Validation* menggunkan K-Fold Gambar

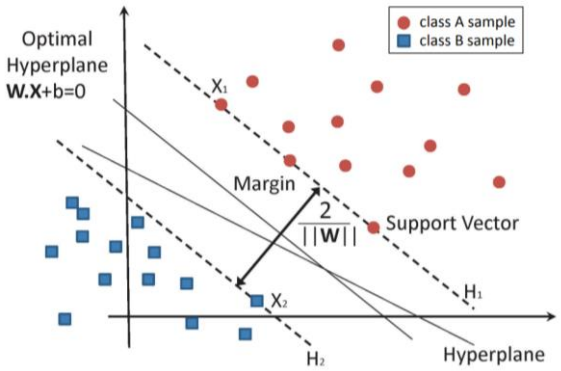
Pada contoh simulasi di gambar 2 menunjukan jumlah K adalah 10. Data testing akan bergerak dari 1 hingga 10. Maka dari itu, proses klasifikasi akan dilakukan sebanyak nilai K dan hasil evaluasi pun juga akan sama dengan jumlah nilai K.

2.4. *Classification*

Tahapan kalsifikasi adalah tahapan dilakukannya prediksi terhadap dataset untuk menemukan suatu pola yang sebelumnya telah melewati tahap preprocessing dan tahap *Data Split*. Pada tahap ini metode klasifikasi yang digunakan adalah algoritma Support Vector Machine (SVM) dan Modified Balanced Random Forest (MBRF).

2.4.1. Support Vector Machine

Prinsip yang mendasar dari SVM adalah bagaimana mencari fungsi hyperplane (garis pemisah) yang dapat memisahkan antara kedua kelas secara maksimal. Maksimal yang dimaksud yaitu hyperplane dapat memisahkan data kedua kelas dengan margin yang paling baik. Margin merupakan jarak garis hyperplane dengan anggota-anggota terdekat dari kedua kelas. Margin yang mampu memisahkan kelas secara maksimal disebut sebagai Optimal Hyperplane.



Gambar 1. Klasifikasi data menggunakan Support Vector Machine (SVM) [14]

Pada gambar 3 menunjukkan bahwa garis H1, H2, dan hyperplane merupakan pemisah kedua kelas. Dimana X adalah dot product dari variabel dan konstanta pada setiap notasi dan W adalah nilai yang tegak lurus dengan X.

$$w.Xi + b \leq -1 \quad (3)$$

Persamaan 3 merupakan *hyperplane* yang bersinggungan terhadap data yang ada pada kelas A (H1).

$$w.Xi + b \geq +1 \quad (4)$$

Persamaan 4 merupakan hyperplane yang bersinggungan terhadap data yang ada pada kelas B (H2).

$$w.X + b = 0 \quad (5)$$

Dan Persamaan 5 merupakan hyperplane yang berada di antara hyperplane kelas A dan kelas B (Garis *Hyperplane*). Sedangkan untuk data yang bersinggungan dengan H1 di kelas A dan H3 di kelas B disebut dengan Support Vector.

Pencarian titik minimal disebut juga dengan *Quadratic Programming* (QP). Penentuan margin diperlukan untuk menentukan titik minimal yaitu dengan $\frac{1}{\|w\|}$. Berikut ini adalah persamaan untuk mencari titik minimal [15]:

$$\min_{w} \tau(\mathrm{W}) = \frac{1}{2}\|w\|^2 \quad (6)$$

Dengan memperhatikan nilai constrain:

$$y_i(Xi.w + b) - 1 \geq 0, \quad \forall_i \quad (7)$$

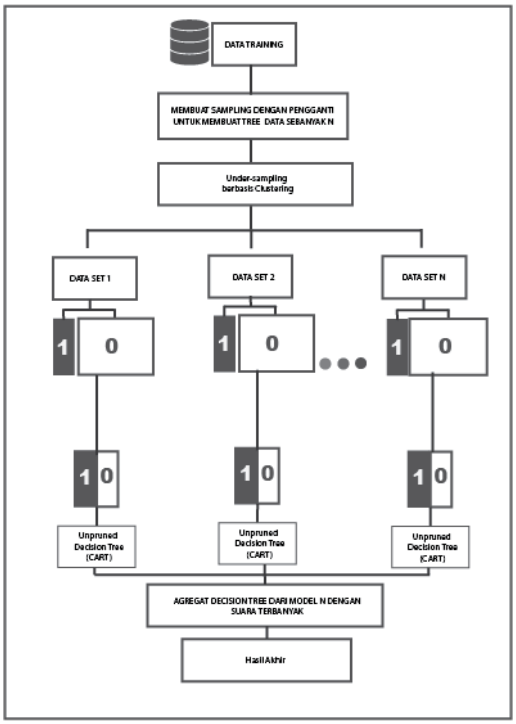
Sampel data yang ada tidak semua memiliki data yang terpisah secara *linear* sehingga tidak bisa menggunakan

DOI: https://doi.org/10.29207/resti.v5i2.3008

SVM *linear*. Apabila dipaksakan nantinya akan memberikan hasil klasifikasi yang buruk dan tidak optimal. Sehingga harus merubah SVM *linear* menjadi SVM *non-linear* agar dapat berjalan dengan optimal, caranya yaitu dengan memanfaatkan metode *kernel*. Pendekatan ini berbeda dengan metode klasifikasi secara umum, yang sebenarnya mengurangi dimensi awal untuk menyederhanakan proses komputasi dan meningkatkan performansi[16].

2.4.2. Modified Balanced Random Forest

Modified Balanced Random Forest adalah metode yang dikembangkan dari Random Forest dan Balanced Random Forest yang bertujuan untuk meningkatkan akurasi prediksi, mengurangi kompleksitas waktu dan penanganan *Imbalanced* yang menjadi masalah utama dalam algoritma klasifikasi *Machine Learning*. Algoritma Modified Balanced Random Forest dikembangkan dengan menggunakan bantuan algoritma lain yaitu algoritma clustering. Metode ini mengubah proses algoritma Balanced Random Forest yang membuang sebagian besar data atau dengan kata lain menggantikan *random undersampling* dalam BRF digantikan dengan teknik clustering. Teknik distribusi data juga disesuaikan dengan jumlah parameter Random Forest yang digunakan dan jumlah *clusters* dalam metode ini disesuaikan dengan jumlah kelas minoritas.



Gambar 2. Model Modified Balanced Random Forest [10]

Metode ini dimulai dari mengambil data dari data latih (D), yang akan dipecah sebanyak masukan dari N *tree* pada Di (D1,D2....Dn) *tree*, yang memiliki Xi dan Yi, di mana Xi data vekor dan Yi label sebuah kelas dan algoritma Random Forest dijalankan yang berlabel (–) dan (+) dan membentuk pohon dua baris dan akan dilakukan proses *undersamping*[10].

Data *training* dan data *testing* kemudian di *undersampling* untuk menyeimbangkan penyebaran data dan menghilangkan *imbalanced* data menggunakan algoritma K-Means. Data dipartisi kedalam satu atau lebih *cluster* yang memiliki karasteristik data yang sama satu sama lain. Hal ini memberikan karakteristik antar *cluster* yang bervariasi, sedangkan data dalam satu *cluster* memiliki karakteristik yang sama. Data yang memiliki kemiripan didalam *cluster* diminimalkan dan mempertahankan keterwakilan variasi antar data, sehingga terjadi pengurangan data yang cenderung sama di tiap *cluster*. Data yang dihasilkan oleh proses *undersampling* kemudian masuk kedalam proses klasifikasi dengan menggunakan algoritma Random Forest dengan metode:

- Menentukan jumlah k atau jumlah pohon yang akan digunakan dalam porses klasifikasi Random Forest. Nilai k yang umumnya digunakan adalah k = 50 karena telah memberikan hasil yang baik untuk klasifikasi dan nilai diatas k = 100 rata rata memiliki tingkat misklasifikasi yang rendah [17].
- Melakukan Bootstrap sampling untuk membangun pohon prediksi sejumlah k.
- Menentukan kriteria pemisahan *node* dalam pohon prediksi dengan menggunakan *entropy*. *Entropy* adalah ukuran kuantitatif dari ketidakteraturan dalam suatu sistem, dengan melakukan perhitungan untuk mencari kesamaan pada dataset dan membagi menjadi beberapa kelas, jika kelas yang dihasilkan berisi data yang serupa, maka *entropy* bernilai nol dan jika kelas yang dihasilkan dapat dibagi menjadi 2 maka *entropy* akan menjadi satu. Metode ini juga menghitung *impurity* dataset yang berarti semakin tinggi nilai *entropy* maka nilai yang dihasilkan maka akan lebih banyak informasi konten. *Entropy* digunakan untuk mengukur seberapa informatif sebuah *node*[18].

$$Entropy = \sum_{i=1}^{c} P_l log p_i \quad (8)$$

- Random Forest kemudian melakukan klasifikasi dengan melakukan *majority vote* dari hasil setiap pohon keputusan.
- Menentukan akurasi ketepatan klasifikasi.

2.5. Validasi

Metode evaluasi yang dilakukan pada penilitian menggunakan Confusion Matrix. Confusion Matrix atau biasa juga disebut Error Matrix adalah metode digunakan dalam melakukan perhitungan akurasi untuk proses klasikasi atau *Supervised Learning*. Pada perhitungan akurasi terdapat empat 4 kombinasi nilai prediksi dan nilai aktual. Keempat istilah tersebut adalah Nilai *True Positive* (TP), Nilai *True Negative* (TF), Nilai

DOI: https://doi.org/10.29207/resti.v5i2.3008

False Positive (FP) dan Nilai *False Negative* (FN). Nilai *True Positive* (TP) merupakan data positif yang diprediksi benar. Nilai *True Negative* (TN) merupakan jumlah dari data negatif yang terdeteksi dengan benar. Nilai *False Positive* (FP) atau disebut juga Type-1 Error merupakan data negatif namun terdeteksi sebagai data yang positif. Nilai *False Negative* (FN) atau disebut juga Type-2 *Error* adalah kebalikan dari *True Positive*, di mana data positif tetapi terdeteksi sebagai data negatif [5].

Tabel 5.Confusion Matrix

	Data Aktual	
	Positif	Negatif
Data Prediksi Positif	True Positive (TP)	False Positive (FP)
Data Prediksi Negatif	False Negative (FN)	True Negative (TN)

Dari table 5 tersebut dapat rumuskan *accuracy* untuk evaluasi performansi dari model klasifikasi yang dibangun. Rumus *accuracy* sebagai berikut [13]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Accuracy adalah nilai perbandingan prediksi benar dengan keseluruhan data. Setelah hasil evaluasi model didapatkan, dilakukan analisis hasil terhadap nilai *accuracy* dari masing-masing kombinasi metode.

3. Hasil dan Pembahasan

Pada penelitian ini terdapat 3 macam skenario pengujian yang dilakukan. Pengujian pada skenario pertama memiliki tujuan untuk mengetahui pengaruh dari proses penggunaan *data cleaning*. Pengujian kedua dilakukan dengan tujuan untuk mengetahui pengaruh proses penggunaan metode normalisasi. Kemudian untuk pengujian ketiga dilakukan untuk mengetahui bagaimana pengaruh penggunaan klasifikasi. Selanjutnya data yang akan digunakan untuk setiap pengujian sama dan pembagian data *training* dan *testing* data akan diaplikasikan K-Fold sebesar K=3, K=5 dan K=7. K-Fold akan membagi data dalam jumlah yang seimbang antara kelas *True* dan kelas *False*.

3.1. Pengujian Pengaruh *Data Cleaning*

Dalam skenario yang kedua ini pengujian dilakukan untuk mengetahui pengaruh dari penggunaan metode *data cleaning* untuk mengatasi missing value dan menggunakan MBRF untuk metode klasifikasi. Ada 2 metode yaitu *Replace Missing Value* dan *Drop Missing Value*. Hasil pengujian bisa dilihat pada table 6.

Berdasarkan tabel 6 dari hasil pengujian drop missing value memiliki nilai 97,8% pada jumlah k = 7 ini cukup tinggi berbandingkan dengan hasil dari replace missing value yang memiliki nilai maksimum 90,2%. Dari hasil akurasi dapat disimpulkan drop missing value memiliki akurasi yang lebih baik.

Tabel 6. Hasil uji penggunaan metode *data cleaning*

Data Cleaning		**Akurasi Maksimum**	
		Drop Missing Value	Replace Missing Value
K-Fold	**3**	79,9%	78,3%
	5	87,3%	84,7%
	7	97,8%	90,2%

3.2. Pengujian Pengaruh Normalisasi

Dalam skenario yang kedua ini pengujian dilakukan untuk mengetahui pengaruh dari penggunaan metode Normalisasi Min-Max dan Z-Score dalam semua kondisi menggunkan SVM sebagai metode klasifikasi. Hasil pengujian dapat dilihat pada Tabel 7.

Tabel 7. Hasil uji penggunaan metode normalisasi

Normalisasi		**Akurasi Maksimum**	
		Min-Max	Z-Score
K-Fold	**3**	80,00%	80,89%
	5	77,75%	85,10%
	7	85,00%	91,48%

Berdasarkan tabel hasil pengujian pada skenario pertama diketahui bahwa menggunakan normalisasi Min-Max menghasilkan nilai akurasi maksimal sebesar 85,00% sedangkan untuk penggunaan normalisasi Z-Score rata-rata akurasi yang di berikan sebesar 91,48%. Dalam penggunaan metode normalisasi banyak faktor yang menentukan hasil terbaik dalam memberikan akurasi, seperti kondisi yang digunakan, pemilihan metode untuk membagi data, dan lain-lain. Dari pengujian pengaruh penggunaan preprocessing dapat diketahui bahwa penggunaan proprocessing Z-Score terbukti lebih optimal dalam melakukan klasifikasi dalam kondisi yang sama.

3.2. Pengujian Pengaruh Klasifikasi

Dalam skenario yang kedua ini pengujian dilakukan untuk mengetahui pengaruh dari penggunaan metode klasifikasi antara SVM dan MBRF. Keduanya sama-sama menggunakan *drop missing value* untuk mengatasi *missing value*. Kemudian untuk metode normalisasinya menggunakan Z-Score. SVM menggunkan *Kernel Linier*, sedangkan pada MBRF menggunkan K-Means untuk *Undersampling*, menggunakan jumlah *tree* = 90 dan metode *Entropy* untuk pemisahan nodenya. Hasil pengujian dapat dilihat pada Tabel 8.

Tabel 8. Hasil uji penggunaan metode klasifikasi

Normalisasi		**Akurasi Maksimum**	
		SVM	MBRF
K-Fold	**3**	80,89%	80,90%
	5	85,10%	87,30%
	7	91,48%	97,80%

Pada tabel 8 dapat dilihat bahwa nilai akurasi yang ditunjukan oleh MBRF dan SVM. Dalam penggunaan

DOI: https://doi.org/10.29207/resti.v5i2.3008

MBRF nilai akurasi maksimum yang dihasilkan sebesar 97,80%. Sedangkan dalam penggunaan SVM nilai akurasi maksimum yang dihasilkan berada diatas 91,48%. Dalam kasus ini, MBRF lebih unggul dibandingkan SVM. Hal tersebut dikarenakan MBRF melakukan proses klasifikasi lebih *smooth* pada setiap kelasnya karena adanya proses *undersampling* data.

4. Kesimpulan

Penelitian yang sudah dilakukan mengusulkan perbandingan algoritma Support Vector Machine (SVM) dan Modified Balanced Random Forest (MBRF) yang digunakan dalam proses klasifikasi terhadap data pasien diabetes dengan dataset yang berasal dari Gula Karya Medika, maka menghasilkan kesimpulan sebagai berikut: (1). Penggunaan algoritma MBRF terbukti lebih efektif untuk mengatasi kasus ini. Hal tersebut dapat dilihat dari akurasi yang di hasilkan MBRF sampai dengan 97,8%. Sedangkan SVM sendiri hanya sebesar 91,48%. Model yang dihasilkan oleh MBRF mampu menangani dataset yang memiliki atribut yang relative kecil. Seperti yang dijelaskan sebelumnya, MBRF mampu membuat model yang lebih efisien atau ringkas. (2). Pada metode preprocessing menggunakan metode *Drop Missing Value* dan Z-Score *Normalization* terbukti yang paling efektif. Hal tersebut dapat dilihat dari kedua metode klasifikasi mencapai akurasi maksimum dengan penggunaan metode tersebut. (3). Jika dilihat pada penelitian sebelumnya yang menggunakan Random Forest biasa menghasilkan akurasi yang lebih kecil daripada MBRF. Bahkan dengan menggunakan metode yang sama yaitu SVM, dibandingkan dengan penlitian sebelumnya menghasilkan performansi yang lebih rendah. Ini membuktikan bahwa metode *preprocessing* pun terbukti lebih efektif menghasilkan data untuk input pada proses klasifikasi.

Pada penelitian selanjutnya, penulis menyarankan untuk menambah dataset dengan atribut yang lebih besar lagi. Dataset yang lebih banyak akan mempermudah proses pembuatan model.

Daftar Rujukan

[1] Kemenkes RI, "Hari Diabetes Sedunia Tahun 2018," *Pus. Data dan Inf. Kementrian Kesehat. RI*, pp. 1–8, 2018.

[2] J. L. Harding, M. E. Pavkov, D. J. Magliano, J. E. Shaw, and E. W. Gregg, "Global trends in diabetes complications: a review of current evidence," *Diabetologia*, vol. 62, no. 1, pp. 3–16, 2019, doi: 10.1007/s00125-018-4711-2.

[3] WHO, "Global Report on Adult Learning Executive Summary," 2016, [Online]. Available: http://apps.who.int/iris/bitstream/10665/204874/1/WHO_NMH_NVI_16.3_eng.pdf?ua=1.

[4] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 4, pp. 530–535, 2019, doi: 10.1016/j.cegh.2018.12.004.

[5] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci.*, vol. 47, no. C, pp. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182.

[6] G. A. B. Suryanegara, Adiwijaya, and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Resti*, vol. 5, no. 1, pp. 114–122, 2021.

[7] Agatsa, D.A, Rismala, R, and Wisesty, U.N, "Klasifikasi Pasien Pengidap Diabetes menggunakan Metode Support Vector Machine," *J. Telkom Univ.*, vol. 7, no. 1, pp. 1–9, 2020.

[8] H. Aydadenta and Adiwijaya, "On the classification techniques in data mining for microarray data classification," 2018, doi: 10.1088/1742-6596/971/1/012004.

[9] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1521–1530, 2018, doi: 10.3844/jcssp.2018.1521.1530.

[10] Z. P. Agusta and Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction," *Int. J. Adv. Intell. Informatics*, vol. 5, no. 1, pp. 58–65, 2019, doi: 10.26555/ijain.v5i1.255.

[11] R. A. Wijayanti, M. T. Furqon, and S. Adinugroho, "Penerapan Algoritme Support Vector Machine Terhadap Klasifikasi Tingkat Risiko Pasien Gagal Ginjal," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 10, pp. 3500–3507, 2018, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/2647/991/.

[12] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[13] Suyanto, *Data Mining Untuk Klasifikasi Dan Klasterisasi Data*. 2019.

[14] E. Excavations, L. Classifiers, E. García-gonzalo, Z. Fernández-muñiz, P. José, and G. Nieto, "Hard-Rock Stability Analysis for Span Design in," pp. 1–19, 2016, doi: 10.3390/ma9070531.

[15] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine, Teori dan Aplikasinya dalam Bioinformatika," *Proc. Indones. Sci. Meet. Cent. Japan*, 2013, doi: 10.1109/CCDC.2011.5968300.

[16] S. V. . Nugroho, "Paradigma Baru Dalam SoftComputing dan Aplikasinya," 2018.

[17] Y. L. Pavlov, "Random forests," *Random For.*, pp. 1–122, 2019, doi: 10.1201/9780429469275-8.

[18] P. Gulati, A. Sharma, and M. Gupta, "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review," *Int. J. Comput. Appl.*, vol. 141, no. 14, pp. 19–25, 2016, doi: 10.5120/ijca2016909926.

DOI: https://doi.org/10.29207/resti.v5i2.3008