Penerapan Algoritma Hybrid Sampling SMOTE-TomekLink dan Random Forest untuk Klasifikasi Penyakit Diabetes

Farda Milanda Amin, Qalbi Ala Dinika

Universitas Bumigora, Mataram, Indonesia

Correspondence : *e-mail*: fardamilanda1405@gmail.com

Abstrak

Ketidakseimbangan data pada dataset sering kali menjadi kendala dalam meningkatkan akurasi klasifikasi pada data medis, termasuk penyakit diabetes. Penelitian ini bertujuan untuk mengatasi permasalahan tersebut dengan menerapkan algoritma hybrid sampling, yaitu kombinasi metode SMOTE (Synthetic Minority Over-sampling Technique) dan TomekLink, serta memanfaatkan algoritma Random Forest sebagai model klasifikasi. Dataset yang digunakan berasal dari Kaggle, berisi 768 data pasien dengan ketidakseimbangan antara kelas negatif dan positif. Metode SMOTE digunakan untuk menyeimbangkan kelas minoritas, sedangkan TomekLink membantu mengurangi data noise dari kelas mayoritas. Hasil evaluasi menunjukkan bahwa kinerja model Random Forest meningkat secara signifikan setelah diterapkan metode Smote-TomekLink, dengan akurasi mencapai 86,4%, sensitivitas 88,2%, dan spesifisitas 81%. Peningkatan ini membuktikan bahwa kombinasi teknik sampling tersebut efektif dalam menangani masalah data tidak seimbang dan meningkatkan performa klasifikasi pada diagnosis penyakit diabetes.

Kata kunci: *Data Tidak Seimbang, Oversampling, Undersampling, Random Forest*

Abstract

Class imbalance in datasets is often an obstacle in improving classification accuracy in medical data, including diabetes. This research aims to overcome this problem by applying a hybrid sampling algorithm, namely a combination of the SMOTE (Synthetic Minority Over-sampling Technique) and TomekLink methods, and utilizing the Random Forest algorithm as a classification model. The dataset used comes from Kaggle, containing 768 patient data with an imbalance between negative and positive classes. The SMOTE method is used to balance the minority class, while TomekLink helps reduce data noise from the majority class. The evaluation results show that the performance of the Random Forest model increased significantly after applying the Smote-TomekLink method, with accuracy reaching 86.4%, sensitivity 88.2% and specificity 81%. This improvement proves that the combination of sampling techniques is effective in dealing with the problem of imbalanced data and improving classification performance in the diagnosis of diabetes.

Keywords*: Unbalanced Data, Oversampling, Undersampling, Random Forest*

1. Pendahuluan

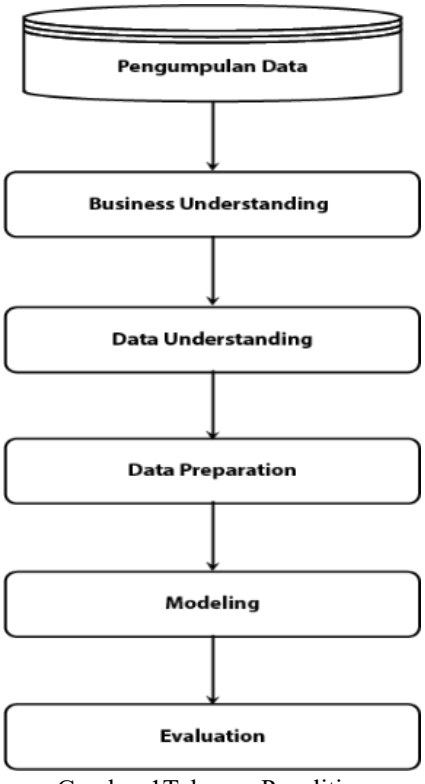
Diabetes atau kencing manis merupakan penyakit metabolik yang diakibatkan oleh tingginya kadar glukosa darah di tubuh dalam waktu yang lama [1]. Bahaya Diabetes yang tidak terkontrol dapat mengakibatkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal dan saraf [2]. Indonesia berada di posisi ke-5 dengan jumlah pengidap diabetes sebanyak 19,47 juta. Jumlah penduduk di Indonesia sebesar 179,72 juta, yang artinya prevalensi diabetes di Indonesia sebesar 10,6% [3]. Permasalahannya adalah kurangnya wawasan pasien tentang ilmu kesehatan, kurangnya perhatian menjaga kesehatan diri, dan kurang memperhatikan pemilihan makanan yang baik di konsumsi. Pengembangan metode klasifikasi yang akurat dan efisien dalam diagnosis penyakit diabetes menjadi sangat penting [4]. Oleh karena itu, penelitian ini menggunakan teknik data *mining* untuk melakukan diagnosis penyakit diabetes. Metode pembelajaran Machine Learning pada data kesehatan khususnya untuk klasifikasi penyakit, telah banyak dipraktikkan [5]. Dataset yang digunakan pada penelitian ini adalah data penyakit diabetes yang diperoleh dari *kaggle* sebanyak 768 data. Namun permasalahannya terdapat ketidakseimbangan kelas pada dataset tersebut yaitu kelas negatif sebanyak 500 data (kelas mayoritas), sedangkan kelas positif sebanyak 268 data (kelas minoritas). Ketidakseimbangan data adalah

jumlah data pada suatu kelas lebih banyak dibandingkan dengan kelas lainnya. Masalah ketidakseimbangan data secara signifikan dapat mempengaruhi performa model klasifikasi [6]. Ketidakseimbangan data menyebabkan metode klasifikasi lebih dominan mengklasifikasikan kelas mayoritas dibandingkan kelas minoritas. Permasalahan ketidakseimbangan data dapat ditangani dengan pendekatan level data.

Beberapa metode *sampling* bisa digunakan untuk menyelesaikan permasalahan ketidakseimbangan data yaitu *oversampling*, *undersampling,* dan *Hybird Sampling*. Dalam Machine Learning, data Imbalanced merupakan masalah penting untuk dipecahkan[7]. *Oversampling* bekerja untuk menyeimbangkan kelas minoritas, sedangkan *Undersampling* bekerja untuk menyeimbangkan data dengan menghapus kelas mayoritas sehingga menghasilkan data yang seimbang. Namun kedua metode tersebut mempunyai kelemahannya tersendiri. Metode *oversampling* yang berlebihan dapat menyebabkan *overfitting*, sedangkan *undersampling* yang berlebihan berpengaruh pada informasi penting yang ada di dataset akan hilang. Untuk meningkatkan kinerja metode o*versampling*, metode *Smote* dikembangkan untuk mengatasi kelemahan pada metode *oversampling* [8]. *Smote* adalah metode *oversampling* yang paling umum digunakan untuk mengatasi masalah ketidakseimbangan data [9]. Namun metode *Smote* memiliki kelemahannya tersendiri yaitu secara acak lebih dominan menginstance kelas *minoritas* pada *oversampling* sehingga rentan menghasilkan data *noise* karena tidak membedakan antar kelas [10]. Oleh karena itu, metode *undersampling* digunakan untuk meningkatkan kinerja metode *oversampling* dengan membersihkan data *noise* pada kelas mayoritas. Penelitian ini menggunakan salah satu metode *undersampling* yaitu *Tomek Link*. *Tomek link* adalah salah satu penghapusan data *noise* untuk menangani data *imbalanced*. Smote-TomekLink adalah sebuah metode yang menggabungkan teknik Smote dan TomekLink untuk menangani ketidakseimbangan data pada masalah klasifikasi [11]. Smote-TomekLink merupakan metode pengambilan sampel untuk menyeimbangkan data pada data diabetes sehingga meningkatkan kinerja klasifikasi metode Random Forest [12].

2. Metode Penelitian

Tahapan penelitian ini menggunakan metodologi Crisp-dm yang memiliki beberapa tahapan seperti yang ditunjukkan pada Gambar 1.



Gambar 1Tahapan Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan pada penelitin ini adalah dataset penyakit diabetes yang diperoleh dari *kaggle.com*.yang terdiri dari 768 *instance* dan 9 kelas dengan 8 atribut dan 1 hasil. Adapun detail atribut yang digunakan ditunjukkan pada Tabel 1.

Tabel 1 Detail Atribut Dataset

No	Atribute	Keterangan	Label
1	*Pregnancies*	Angka kehamilan	X1
2	*Glucose*	Kadar glukosa 2 jam setelah makan. Menurut WHO salah satu yang menjadi kriteria penyakit diabetes, yaitu kadar glukosa minimal 200 mg/dl.	X2
3	*Blood Pressure*	Tekanan Darah	X3
4	*Skin Thickness*	Ketebalan Kulit	X4
5	*Insulin*	Insulin	X5
6	BMI	Berat Badan	X6
7	*Diabetes Pedigree Function*	Riwayat diabetes dalam keluarga	X7
8	*Age*	Umur	X8
9	*Outcome*	Status diabetes (1 = positif diabetes, 0 = negatifdiabetes) .	Y

2.2 Business Understanding

Tahapan pertama adalah *business understanding*. Tahapan ini digunakan untuk kepemahaman kebutuhan berdasarkan penilaian bisnis. Pemahaman tersebut diubah menjadi sebuah rencana awal data mining yang dirancang untuk mencapai tujuan. Pemahaman bisnis mengacu pada jumlah penderita penyakit diabetes yang ditentukan oleh data yang ada. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut. Menerjemahkan tujuan dan batasan dari data yang diambil menjadi formula dari permasalahan data mining mulai dari menyiapkan strategi awal hingga metode yang dibutuhkan untuk mencapai tujuan.

2.3 Data Understanding

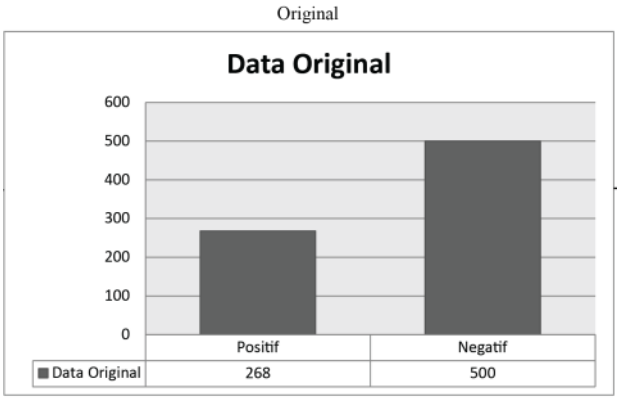
Pada tahap ini, dilakukan pengumpulan *dataset* yang sesuai dengan permasalahan pada skripsi ini. *Dataset* yang digunakan dalam skripsi ini merupakan data publik yang diperoleh dari situs *kaggle.com*. Pada tahapan ini juga dilakukan visualisasi data untuk memudahkan dalam memahami karakteristik dataset yang digunakan.

2.4 Data Preparation

Pada tahap ini digunakan untuk menangani kelemahan data yang ditemukan pada tahap data *understanding*. Pada dataset yang digunakan terdapat ketidakseimbangan *class* dimana *class* negatif lebih banyak daripada *class* positif, jika ini dibiarkan maka metode klasifikasi cenderung mengklasifikasikan *class* yang lebih dominan, oleh karena itu pada penelitian ini menggunakan metode *Smote* untuk menangani ketidakseimbangakan *class* ini menjadi seimbang.

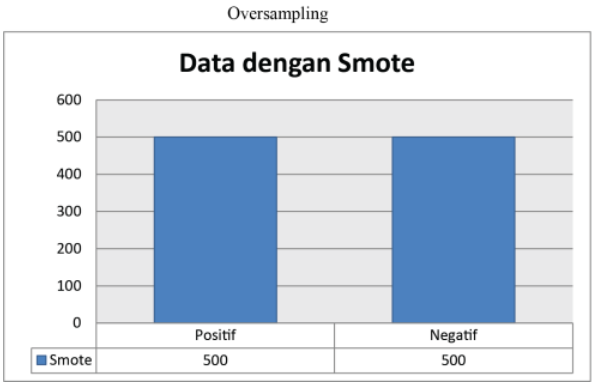
2.4.1 Visualisasi Data

Pada bagian visualisasi data digunakan untuk menggambarkan karakteristik data yang digunakan dalam bentuk diagram. Adapun pengecekan dataset pada tahap ini untuk melihat deskripsi data, korelasi data, *missing value*, dan ketidakseimbangan kelasnya. Pada dataset yang digunakan terdapat ketidakseimbangan *class* dimana *class* negatif lebih banyak daripada *class* positif yang ditunjukkan dengan diagram batang pada Gambar 2.



Gambar 2 Data Awal

Pada bagian visualisasi data telah ditunjukkan digaram batang pada data awal sebelum diolah, maka Gambar 3 merupakan contoh diagram batang pada data yang telah diolah dengan metode Smote.



Gambar 3 Data Setelah Smoote

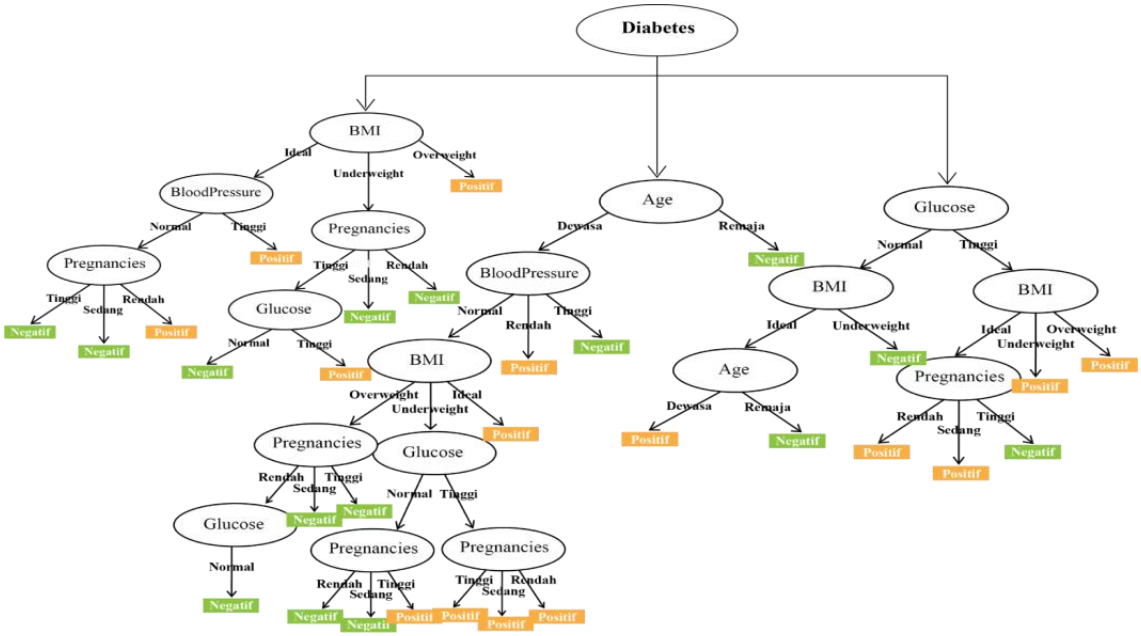
2.5 Modelling

Pada tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatan nilai yang optimal. Data yang digunakan dalam penelitian ini adalah sumber data primer. Data yang dikumpulkan yaitu data pasien penderita penyakit diabetes sebanyak 60 data sebagai contoh perhitungannya. Data yang digunakan dalam penelitian ini akan diklasifikasikan dengan membentuk pohon keputusan menggunakan algoritma *random forest,* dari 60 data tersebut dibentuk menjadi 3 pohon yang dimasing-masing pohonnya terdapat 20 data. Data yang digunakan untuk membentuk pohon keputusan dihitung dari hasil nilai *Entropy* pada rumus persamaan (1) dan *Gain* pada rumus persamaan (2). Proses pembentukan pohon diawali dengan memilih nilai atribut *Gain* tertinggi sebagai *Root*nya dan diikuti dengan cabang pohon berikutnya.

$$Entropy\ (S) = \sum_{i=1}^{n} - Pi * \log_2 Pi$$

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

Hasil dari perhitungan menggunakan algoritma *random forest* dengan rumus persamaan (1) dan (2) dengan contoh data sebanyak 60 data yang telah dikonversikan menghasilkan 3 pohon keputusan pada Gambar 4, yang dimana hasil votting terbanyak adalah kelas positif.



Gambar 4 Pohon Keputusan Random Forest

3. Hasil dan Pembahasan

Pada bagian ini akan membahas mengenai hasil penelitian mulai dari persiapan data, kolerasi antar atribut, distribusi data tidak seimbang, penyeimbangan data dengan metode *smote* dan *tomek*, hasil implementasi pemodelan, performa model klasifikasi dalam memprediksi, dan hasil evaluasi kinerja algoritma *random forest.* Setelah proses persiapan data selesai, proses berikutnya adalah implementasi model algoritma data mining. Ada beberapa tahapan dalam implementasi model algoritma data mining salah satunya adalah klasifikasi.

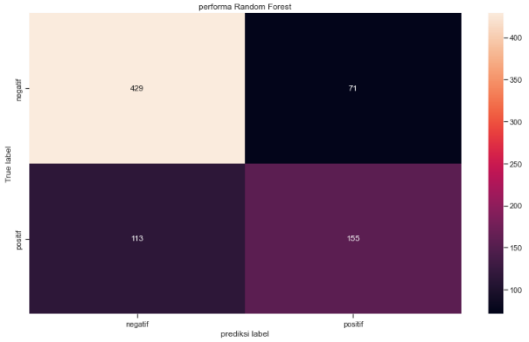
3.1 Kinerja Metode Random Forest pada Data Original

Untuk melihat kinerja dari penggunaan metode *smote*, maka pertama membuat model *random forest* menggunakan data original tanpa smote untuk mendapatkan hasil akurasi dan gambaran tentang kinerja model awal. Pada Gambar 5 model *random forest* dengan *Smote-Tomeklink* memperoleh nilai *True* Positif 419, *True* Negatif 385, *False* Negatif 90, dan *False* Positif 56. Selanjutnya adalah meghitung hasil klasifikasi untuk melihat nilai Akurasi, Sensitifitas, Spesifisitas.

$$\textbf{Akurasi} = \frac{155+429}{155+429+71+113} * 100\% = 0.760$$

$$\textbf{Sensitifitas} = \frac{155}{155+113} * 100\% = 0.578$$

$$\textbf{Spesifisitas} = \frac{429}{429+71} * 100\% = 0.858$$



Gambar 5 Performa Random Forest Data Awal

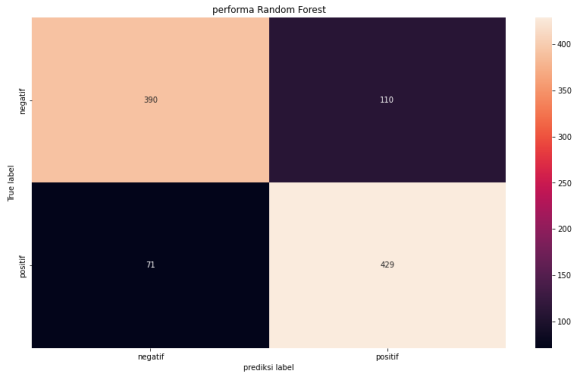
3.2 Kinerja Metode Random Forest pada Data Hasil Smote

Data yang tidak seimbang menyebabkan metode klasifikasi lebih dominan mengklasifikasikan kelas mayoritas dibandingkan kelas minoritas dan akan berpengaruh pada hasil akhir, dengan demikian data akan melalui proses tahap penyeimbangan kelas menggunakan metode *smote,* yang dimana metode *smote* adalah metode *oversampling* yang fungsinya untuk menyeimbangkan data pada kelas minoritas. Pada Gambar 6 model *random forest* dengan *Smote* memperoleh nilai *True* Positif 429, *True* Negatif 390, *False* Negatif 110, dan *False* Positif 71. Selanjutnya adalah meghitung hasil klasifikasi untuk melihat nilai Akurasi, Sensitifitas, Spesifisitas.

$$\textbf{Akurasi} = \frac{429+390}{429+390+110+71} * 100\% = 0.819$$

$$\textbf{Sensitifitas} = \frac{429}{429+71} * 100\% = 0.858$$

$$\textbf{Spesifisitas} = \frac{390}{390+110} * 100\% = 0.780$$



Gambar 6 Performa Random Forest Data Smote

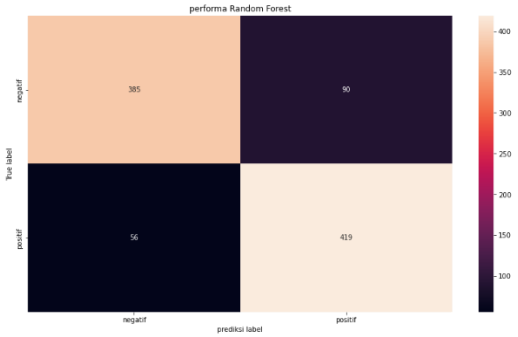
3.3 Kinerja Metode Random Forest pada Data Smote-TomekLink

Jika metode *smote* adalah metode *oversampling*, maka *tomek* adalah metode *undersampling*. Tahap berikutnya yaitu melakukan proses *undersampling* dengan mengurangi hasil *oversampling* dengan metode *tomeklink.* Metode *undersampling* mengurangi sampel dengan cara 2 data yang saling berdekatan antara kelas minoritas dan kelas mayoritas. Pada Gambar 7 model *random forest* dengan *Smote-Tomeklink* memperoleh nilai *True* Positif 419, *True* Negatif 385, *False* Negatif 90, dan *False* Positif 56. Selanjutnya adalah meghitung hasil klasifikasi untuk melihat nilai Akurasi, Sensitifitas, Spesifisitas.

$$\textbf{Akurasi} = \frac{\mathbf{419+385}}{\mathbf{419+385+90+56}} * 100\% = 0.864$$

$$\textbf{Sensitifitas} = \frac{\mathbf{419}}{\mathbf{419+56}} * 100\% = 0.882$$

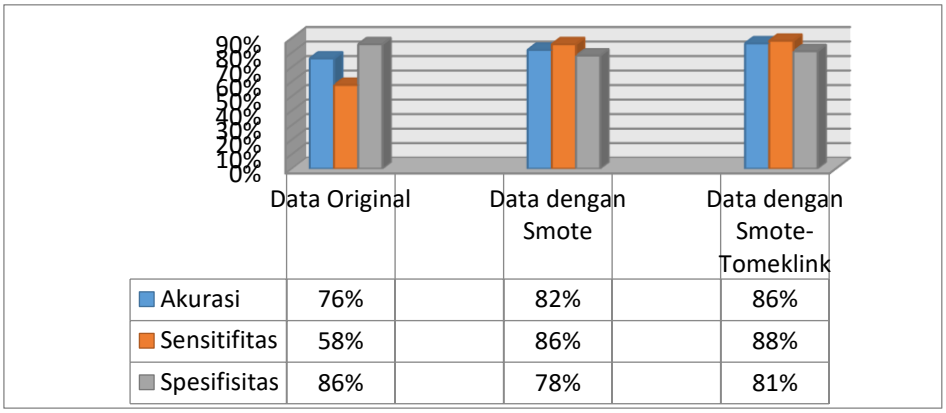
$$\textbf{Spesifisitas} = \frac{\mathbf{385}}{\mathbf{385+90}} * 100\% = 0.810$$



Gambar 7 Performa Random Forest Data Smote-TomekLink

3.4 Perbandingan Model Tanpa Smote, dengan Smote dan Smote-TomekLink

Berdasarkan hasil pengujian pada ketiga model, terbukti bahwa metode *Smote* dapat mengurangi *overfiting* pada model dan dapat meningkatkan kinerja dari model yang dibangun. *Smote* dapat menyeimbangkan kelas dengan menambahkan data sintetis pada kelas minoritas dan *Smote-Tomeklink* mengurangi jumlah data pada kelas mayoritas. Hal tersebut dapat dilihat dari kenaikan nilai Akurasi, Sensitifitas, dan Spesifisitas yang ditunjukkan pada Gambar 8.



Gambar 8 Akurasi, Sensitifitas, Spesifisitas pada Data Awal, Smote, dan Smote-TomekLink

4. Kesimpulan

Berdasarkan penelitian yang dilakukan pada dataset penyakit diabetes yang diperoleh dari *kaggle.com* dan diolah menggunakan algortima *Random Forest* dengan menerapkan metode *Smote* dan *Smote-Tomeklink* dalam klasifikasi penyakit diabetes dapat diambil kesimpulan bahwa hasil penelitian menggunakan algoritma *Random Forest* dengan metode *Smote* dan *Smote-Tomeklink* menghasilkan kinerja yang lebih baik dibandingkan dengan data original yang dimana, Algoritma *Random Forest* tanpa *Smote* memperoleh hasil akurasi dibawah 80%, algoritma *Random Forest* dengan *Smote* memperolah hasil akurasi diatas 80% dan algoritma *Random Forest* dengan *Smote-TomekLink* memperoleh hasil akurasi diatas 85%. Ketidakseimbangan data menjadi penyebab utama hasil akhir yang dilakukan pada data awal tidak efektif dibandingkan dengan data yang telah melakukan proses penyeimbangan data dengan metode *Smote* dan *Smote-Tomeklink.*

Daftar Pustaka

[1] R. P. Kurniadi, R. R. Saedudin, and V. P. Widartha, “Perbandingan Akurasi Algoritma K-Nearest Neighbor Dan Logistic Regression Untk Klasifikasi Penyakit Diabetes,” in *e-Proceeding of Engineering*, 2021, pp. 9757–9764.

[2] D. A. Agatsa, R. Rismala, and U. N. Wisesty, “Klasifikasi Pasien Pengidap Diabetes Metode Support Vector Machine,” *e-aproceeding of Enginering*, vol. 7, no. 1, pp. 2517–2525, 2020.

[3] M. Hassanein *et al.*, *Diabetes and Ramadan: Practical guidelines 2021*, vol. 185. 2021. doi: 10.1016/j.diabres.2021.109185.

[4] M. Salsabil, N. Lutvi, and A. Eviyanti, “Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost,” *J. Ilm. Komputasi*, vol. 23, no. 1, pp. 51–58, 2024, doi: 10.32409/jikstik.23.1.3507.

[5] H. Hairani and D. Priyanto, “A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 585–590, 2023, doi: 10.14569/IJACSA.2023.0140864.

[6] Sutarman, R. Siringoringo, D. Arisandi, E. Kurniawan, and E. B. Nababan, “Model Klasifikasi Dengan Logistic Regression Dan Recursive Feature Elimination Pada Data Tidak Seimbang,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 735–742, 2024, doi: 10.25126/jtiik.1148198.

[7] R. Ridwan, E. H. Hermaliani, and M. Ernawati, “Penerapan: Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian,” *Comput. Sci.*, vol. 4, no. 1, pp.

80–88, 2024, [Online]. Available: https://jurnal.bsi.ac.id/index.php/co-science/article/view/2990

[8] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, 2021, doi: 10.33387/jiko.v4i1.2561.

[9] A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *Int. J. Inf. Educ. Technol.*, vol. 13, no. 2, pp. 289–295, 2023, doi: 10.18178/ijiet.2023.13.2.1806.

[10] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk Menangani Ketidakseimbangan Kelas dalam Kalsifikasi Penyakit Diabetes dengan C4.5, SVM, dan naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[11] L. G. R. Putra, K. Marzuki, and H. Hairani, "Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction," *Eng. Appl. Sci. Res.*, vol. 50, no. 6, pp. 577–583, 2023, doi: 10.14456/easr.2023.59.

[12] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.