JSITIK, Vol.2 No.2 Juni 2024
ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang

Diterima:
21 April 2024
Revisi:
4 Mei 2024
Terbit:
1 Juni 2024

[1*]Arie Nugroho, [2]Dwi Harini
[1-2]Universitas Nusantara PGRI Kediri

Abstrak— Data tidak seimbang terjadi karena jumlah data pada tiap kelas berbeda jauh dimana akan mempengaruhi hasil prediksi. Dalam penelitian ini menggunakan dataset prediksi diabetes, yang mengandung data yang tidak seimbang. Hasil prediksi ditunjukkan dengan nilai akurasi dan presisi. Tujuan penelitian ini adalah meningkatkan nilai akurasi dan presisi pada data yang tidak seimbang. Metode yang digunakan dalam penelitian ini adalah penentuan sampling dan pembelajaran ensemble. Penentuan sampling yang digunakan adalah dengan cara mengalikan data pada kelas minoritas atau oversampling. Teknik Oversampling yang digunakan adalah Synthetic Minority Oversampling Technique (SMOTE). Pembelajaran ensemble yang digunakan adalah algoritma random forest. Kombinasi algoritma SMOTE dan random forest dapat meningkatkan akurasi dan menyeimbangkan nilai presisi pada setiap kelas. Hasil penelitian ini adalah Kombinasi tersebut menghasilkan nilai akurasi sebesar 97,5% dan nilai presisi pada kelas non pasien sebesar 97% sedangkan nilai presisi pada kelas pasien sebesar 98%.
Kata Kunci— Random Forest; Imbalanced Data; SMOTE

Abstract— *Imbalanced data occurs because the amount of data in each class is very different which will affect the prediction results. This study uses a diabetes prediction dataset, which contains imbalanced data. Prediction results are shown with accuracy and precision values. The aim of this research is to increase accuracy and precision values in unbalanced data. The method used in this research is sampling and ensemble learning. Determining the sampling used is by multiplying the data in the minority class or oversampling. The oversampling technique used is Synthetic Minority Oversampling Technique (SMOTE). The ensemble learning used is the random forest algorithm. The combination of the SMOTE algorithm and random forest can increase accuracy and balance the precision values in each class. The results of this research are that this combination produces an accuracy value of 97.5% and a precision value in the non-patient class of 97%, while the precision value in the patient class is 98%.*
Keywords— *Random Forest; Imbalanced Data; SMOTE*

Penulis Korespondensi:

Arie Nugroho,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: arienugroho@unpkediri.ac.id
ID Orcid: https://orcid.org/0000-0001-9080-5723

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

I. PENDAHULUAN

Data yang tidak seimbang atau *Imbalanced data* merupakan masalah umum yang terjadi pada kasus *machine learning*. *Imbalanced data* menyebabkan prediksi model *machine learning* terlalu mengikuti kelas mayoritas[1]. Kelas minoritas kurang terwakili karena menang datanya sedikit, sehingga hasil prediksi akan cenderung pada kelas mayoritas. Imbalanced data dapat terjadi karena beberapa hal. Penyebab pertama adalah data memang sudah tidak seimbang dari awal. Data pada suatu kejadian atau tempat sudah tidak seimbang karena kondisi tertentu. Penyebab kedua adalah pengambilan jumlah data latih atau data training yang kurang seimbang antar kelas[2], [3]. Penggunaan data *training* yang seimbang akan membuat model *machine learning* memberikan hasil prediksi yang lebih baik[4]. *Oversampling* merupakan salah satu teknik dalam *sampling*. *Oversampling* merupakan duplikasi data pada kelas tertentu untuk menyamakan data sehingga sama dengan kelas yang lain[5]. Teknik *oversampling* ada dua jenis, yaitu *random* dan sintetis. *Random oversampling* merupakan duplikasi secara acak dari data *training*[6]. *Random oversampling* memungkinkan akan ada data yang sama pada data *training*. Sintetis *oversampling* atau *synthetic minority oversampling technique* (SMOTE) merupakan teknik *oversampling* yang dapat menghasilkan data sintetis baru dalam ruang fitur yang berada di antara data-data yang ada[7]. SMOTE dapat meningkatkan kinerja algoritma *machine learning* di kelas minoritas dan mengurangi resiko *overfitting* di kelas mayoritas[8]. *Overfitting* adalah kondisi dimana hasil akurasi tinggi pada masa *training*, tetapi rendah pada saat *testing*[9], [10].

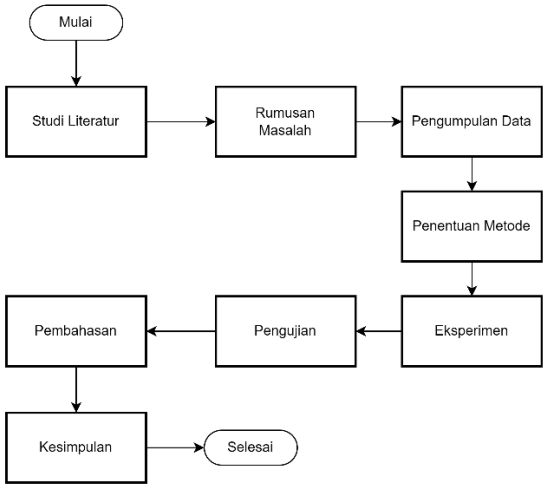
Beberapa penelitian sebelumnya yang dijadikan acuan dalam penelitan ini adalah riset yang membahas penerapan metode klasifikasi pada penyakit diabetes. Penelitian pertama adalah membahas tentang penerapan metode K-Nearest Neighbor (KNN). Pada penelitian ini menguji beberapa nilai K. Nilai akurasi tertinggi adalah 39 % dan nilai presisi sebesar 65 %[11]. Berikutnya adalah penelitian yang membandingkan algoritma *Naive Bayes* dan *Support Vector Machine* (SVM) . Hasil akurasi yang didapat dari naive bayes sebesar 92 % dan SVM sebesar sebesar 96 %[12]. Penelitian berikutnya adalah Implementasi algoritma random forest menggunakan metode normalisasi. Nilai akurasi yang didapat dari kombinasi random forest dan normalisasi sebesar 95.45 %[13]. Dari beberapa penelitian tersebut algoritma naive bayes, SVM dan random forest menghasilkan nilai akurasi yang tinggi, di atas 90 %. Namun tidak dibahas nilai presisinya serta nilai akurasi tersebut dapat ditingkatkan lagi. Perbedaan penelitian ini dengan beberapa penelitian sebelumnya adalah adanya *pre-processing* berupa *balancing data* untuk menyeimbangkan data yang tidak seimbang. Tujuan dari penelitian ini adalah untuk meningkatkan nilai akurasi dan presisi untuk data yang tidak seimbang. Kombinasi teknik oversampling (SMOTE) dan Algoritma Random Forest diharapkan dapat mengatasi masalah

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

imbalanced data[14] dengan meningkatnya nilai akurasi dan presisi. Dalam penelitian ini akan dilakukan dua skenario untuk membandingkan hasil dari nilai akurasinya. Skenario pertama klasifikasi data dengan Random Forest tanpa SMOTE. Skenario kedua klasifikasi data dengan Random Forest dengan SMOTE.

II. METODE

Pada bagian ini akan dijelaskan metode yang akan digunakan dalam penelitian ini. Langkah-langkah yang akan dilakukan dalam penelitian ini akan disajikan dalam diagram penelitian akan ditunjukkan pada gambar 1.

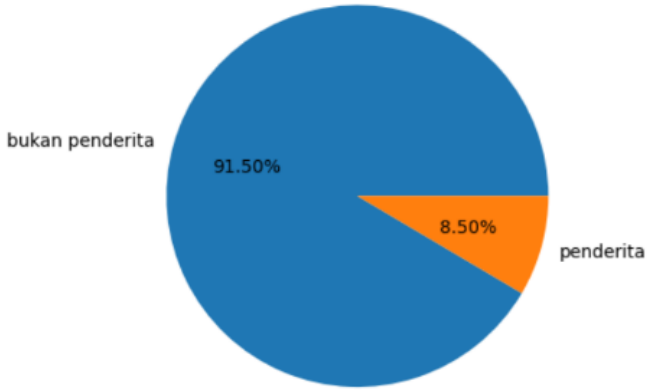


Gambar 1. Diagram Penelitian

Pada Gambar 1 ditunjukkan diagram penelitian yang dimulai dari studi literatur. Studi literatur meliputi semua studi yang digunakan dalam penelitian ini, berupa artikel penelitian selama 10 tahun terakhir. Penentuan rumusan masalah dalam penelitian ini adalah bagaimana meningkatkan nilai akurasi dan presisi pada data yang tidak seimbang. Pengumpulan data dilakukan dengan mengunduh dataset publik dari Kaggle.com. Data yang digunakan adalah klasifikasi dari prediksi penyakit diabetes. Dataset ini mempunyai 100.000 data atau baris dan 8 atribut atau kolom. Dataset ini disimpan dalam file Comma Separated Value (CSV). Dataset berisi tentang data-data yang terkait tentang indikasi atau ciri penyakit diabetes. Dalam dataset tersebut mempunyai beberapa *tipe* data yang berbeda. Atribut *age*, bmi, hbA1c dan *blood glucose level* mempunyai *tipe* data *float*. Atribut *hypertension*, *heart disease* dan diabetes mempunyai *tipe* data integer dengan kemungkinan hanya 1 dan 0 (biner). Atribut *gender* dan *smoking history* mempunyai *tipe* data *object*(*string*). Berdasarkan data dari dataset tersebut label diabetes untuk nilai 1 (penderita)

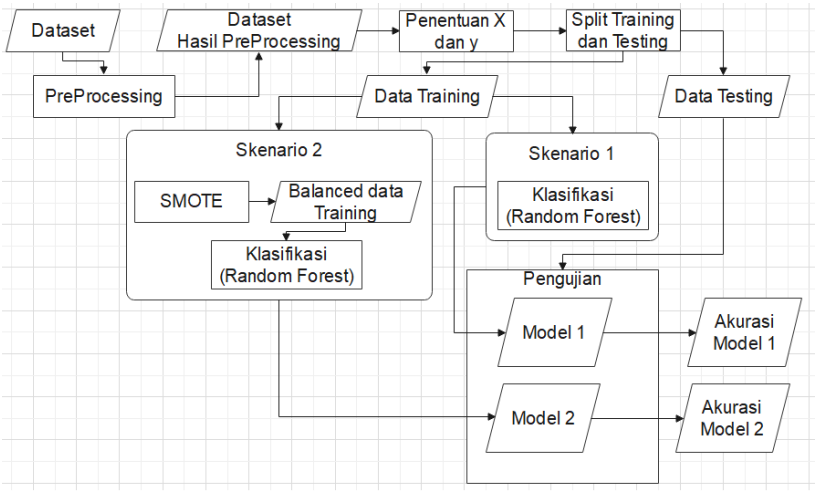
ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

sebanyak 8.500 data, sedangkan untuk nilai 0 (bukan penderita) sebanyak 91.500 data. Grafik perbandingan label penderita diabetes dan yang bukan penderita ditunjukkan pada Gambar 2.



Gambar 2. Perbandingan Label

Pada Gambar 2 ditampilkan perbandingan label dari dataset, yaitu penderita diabetes dan bukan penderita. Label untuk bukan penderita sebanyak 91,5 %, sedangkan penderita hanya 8,5%. Berdasarkan perbandingan tersebut label untuk bukan penderita merupakan kelas mayoritas sedangkan label untuk penderita merupakan kelas minoritas, sehingga dataset tersebut termasuk dalam Imbalanced data[15],[16]. Dalam penelitian ini akan membandingkan 2 skenario. Skenario 1 menggunakan dataset asli atau yang belum seimbang. Skenario 2 menggunakan dataset yang sudah diseimbangkan(balanced). Perbandingan yang akan dilakukan dari kedua skenario adalah akurasi dan precision. Precision dipakai untuk mengukur seberapa presisi dari prediksi[17]. Precision dihitung dari true positif dengan penjumlahan dari true positif dan false negative[18]**.** Alur penelitian ditunjukkan pada Gambar 3.



Gambar 3. Alur Penelitian

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

Pada Gambar 3 ditampilkan Alur penelitian dalam riset ini. Langkah pertama adalah load dataset prediksi penyakit diabetes. Langkah kedua adalah pre-processing. Pre-processing yang dilakukan adalah konversi data dari beberapa atribut[19], [20]. Konversi data yang dilakukan adalah mengubah data dari bentuk teks ke bentuk angka[21]. Pada dataset yang digunakan, atribut dengan data teks adalah gender dan smoking_history. Dataset hasil pre-processing ditunjukkan pada Gambar 4.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0	80.0	0	1	4	25.19	6.6	140	0
1	0	54.0	0	0	0	27.32	6.6	80	0
2	1	28.0	0	0	4	27.32	5.7	158	0
3	0	36.0	0	0	1	23.45	5.0	155	0
4	1	76.0	1	1	1	20.14	4.8	155	0

Gambar 4. Dataset Hasil Pre-Processing

Pada Gambar 4 ditampilkan Dataset hasil konversi data dari beberapa atribut. Atribut *gender* diubah dari bentuk teks (male/female) menjadi 0 atau 1. Atribut *Gender* dengan data *female* menjadi angka 0 dan *male* menjadi angka 1. Atribut *smoking_history* juga diubah dari bentuk teks (no info, never, current, former) menjadi angka dengan *range* 0 sd 4. Atribut-atribut lainnya tidak dikonversi karena sudah dalam bentuk angka. Langkah ketiga adalah penentuan variabel x dan y. Langkah ini dilakukan untuk memisahkan atribut bebas yaitu x dengan atribut terikat yaitu y. Atribut x adalah mulai dari atribut *gender,age* sampai dengan *blood_glocose_level*. Atribut y adalah label atau kelasnya yaitu diabetes. Langkah keempat adalah *split* data *training* dan *testing*. Data *training* digunakan untuk proses pelatihan model. Data *testing* digunakan untuk menguji model yang dihasilkan dari proses pelatihan. Pada riset ini data *training* dialokasikan 80 % dari banyaknya data keseluruhan dan sisanya 20 % adalah data *testing*. Langkah kelima adalah menjalankan skenario. Skenario 1 adalah proses klasifikasi dengan algoritma random forest dari data asli. Skenario 2 diawali proses penyeimbangan (*balancing*) dataset. Proses *balancing* menghasilkan dataset dengan jumlah data yang sama dari setiap kelasnya[22]. Pada dataset yang telah diseimbangkan banyaknya data pada label bukan penderita dan penderita menjadi sama, yaitu 91.500 data. Setelah dataset diseimbangkan berikutnya adalah klasifikasi dengan algoritma *random forest*. Langkah keenam adalah pengujian model. Pengujian dilakukan untuk kedua skenario, yaitu skenario 1 dan skenario 2. Masing-masing model dari setiap skenario akan diuji dengan data *testing* untuk mengetahui nilai akurasi dan presisinya.

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

III. HASIL DAN PEMBAHASAN

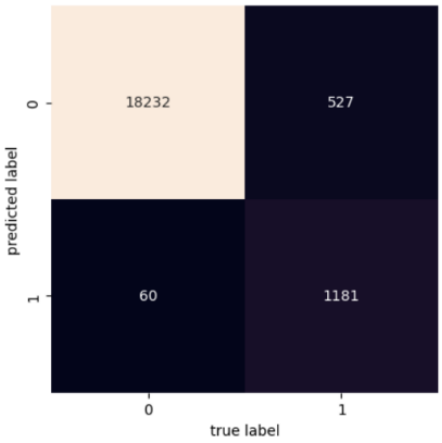
Pada bagian ini akan dijelaskan pembahasan dari alur penelitian yang telah dilakukan menggunakan dataset prediksi penyakit diabetes. Pengaturan parameter yang digunakan dalam algoritma random forest dalam riset ini menggunakan 100 pohon / trees. Pada skenario 1 menggunakan data training sebesar 80.000 data dan 20.000 data testing. Skenario 1 menghasilkan nilai akurasi sebesar 97,065 %. Nilai presisi pada kelas bukan penderita mencapai nilai 1.00. Nilai presisi pada kelas penderita hanya mencapai nilai 0.69. Hasil pengujian pada skenario 1 ditampilkan pada Gambar 5.

```
Accuracy of model:  0.97065
              precision    recall  f1-score   support

           0       1.00      0.97      0.98     18759
           1       0.69      0.95      0.80      1241

    accuracy                           0.97     20000
   macro avg       0.84      0.96      0.89     20000
weighted avg       0.98      0.97      0.97     20000
```

Gambar 5. Akurasi model Skenario 1

Pada Gambar 5 ditunjukkan akurasi model pada skenario 1. Nilai presisi pada kelas penderita kurang baik karena hanya 69 %. Hal ini disebabkan data pada kelas penderita adalah kelas minoritas. Nilai recall untuk kedua kelas cenderung seimbang. Kelas bukan penderita mencapai nilai 97 % sedangkan untuk kelas penderita mencapai nilai 95 %. Nilai f1-score dari kedua mempunyai selisih yang cukup banyak yaitu 18 %. Kelas bukan penderita memperoleh nilai 98 %, sedangkan kelas penderita memperoleh nilai 80 %. Pada pengujian dengan confusion matrix hasil prediksi pada setiap label akan dibandingkan dengan label sesungguhnya[23]. Pada confusion matrix menjelaskan perbandingan antara prediksi dan hasil sesungguhnya. Nilai True terbagi menjadi True Positif (TP) dan True Negatif(TN). Nilai False terbagi menjadi False Positif (FP) dan False Negatif(FN). Prediksi sesuai (True) mencapai 19.413 data sedangkan prediksi salah (False) mencapai 587 data. Gambar hasil pengujian dengan confusion matrix ditunjukkan pada Gambar 6.

JSITIK, Vol.2 No.2 Juni 2024
ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX



Gambar 6.Hasil Confusion_matrix Skenario 1

Pada Gambar 6 ditampilkan hasil pengujian skenario 1 dengan confusion matrix. Pada confusion matrix tersebut nilai 0 mewakili kelas bukan penderita, sedangkan nilai 1 mewakili kelas penderita. Pada hasil predicted label dari kelas bukan penderita yang sesuai (TN) dengan true label-nya bukan penderita mencapai 18.232 data. Pada kelas penderita banyaknya data predicted label yang sesuai (TP) dengan true-labelnya penderita mencapai 1.181 data. Hasil predicted label untuk kelas bukan penderita yang tidak sesuai (FN) dengan true label-nya bukan penderita mencapai 527 data, sedangkan untuk kelas penderita yang tidak sesuai (FP) true label-nya penderita mencapai 60 data. Skenario 2 menggunakan data training sebesar 146.400 data dan 36.600 data testing. Skenario 2 menghasilkan akurasi sebesar 97,464 %. Nilai presisi pada kelas bukan penderita mencapai nilai 0.98, sedangkan nilai presisi pada kelas penderita mencapai nilai 0.97. Nilai presisi dan f1-score kelas bukan penderita dan penderita cenderung sama atau seimbang. Hasil pengujian pada skenario 2 ditampilkan pada Gambar 7.

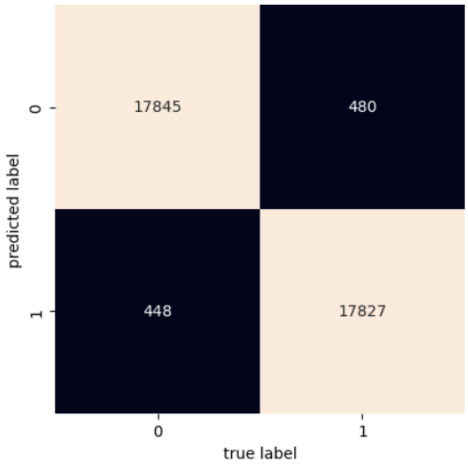
```
Accuracy of model:  0.9746448087431694
              precision    recall  f1-score   support

           0       0.98      0.97      0.97     18325
           1       0.97      0.98      0.97     18275

    accuracy                           0.97     36600
   macro avg       0.97      0.97      0.97     36600
weighted avg       0.97      0.97      0.97     36600
```

Gambar 7. Akurasi model Skenario 2

Pada Gambar 7 ditunjukkan akurasi model pada skenario 2. Nilai presisi pada kelas penderita mencapai 97 %, sedangkan kelas bukan penderita mencapai 98 %. Nilai recall untuk kedua kelas cenderung seimbang. Kelas bukan penderita mencapai nilai 98 % sedangkan untuk kelas penderita

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

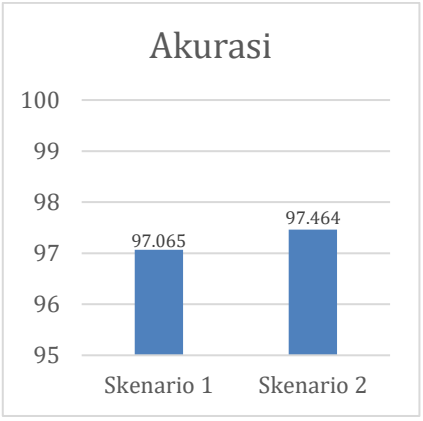
mencapai nilai 97 %. Nilai f1-score dari kedua kelas mempunyai nilai yang sama yaitu 97 %. Pada pengujian dengan confusion matrix hasil prediksi pada setiap label akan dibandingkan dengan label sesungguhnya. Pada confusion matrix menjelaskan perbandingan antara prediksi dan hasil sesungguhnya. Nilai True terbagi menjadi True Positif (TP) dan True Negatif(TN). Nilai False terbagi menjadi False Positif (FP) dan False Negatif(FN)[24]. Prediksi sesuai (True) mencapai 35.672 data sedangkan prediksi salah (False) mencapai 928 data. Gambar hasil pengujian dengan confusion matrix ditunjukkan pada Gambar 8.



Gambar 8.Hasil Confusion_matrix Skenario 2

Pada Gambar 8 ditampilkan hasil pengujian skenario 2 dengan confusion matrix. Pada confusion matrix tersebut nilai 0 mewakili kelas bukan penderita, sedangkan nilai 1 mewakili kelas penderita. Pada hasil predicted label dari kelas bukan penderita yang sesuai (TN) dengan true label-nya bukan penderita mencapai 17.845 data. Pada kelas penderita banyaknya data predicted label yang sesuai (TP) dengan true label-nya penderita mencapai 17.827 data. Hasil predicted label untuk kelas bukan penderita yang tidak sesuai (FN) dengan true label-nya bukan penderita mencapai 480 data, sedangkan untuk kelas penderita yang tidak sesuai (FP) dengan true label-nya penderita mencapai 448 data. Berikutnya adalah perbandingan nilai akurasi dari kedua skenario yang ditunjukkan pada Gambar 9.

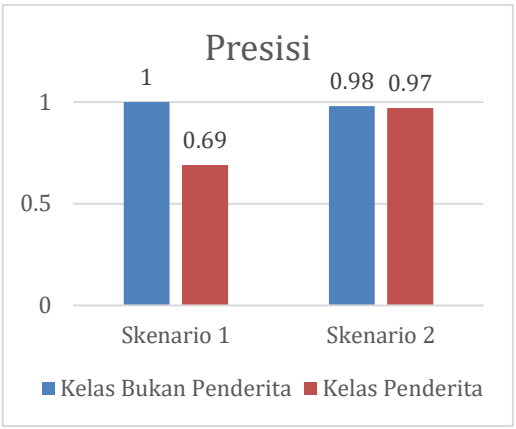
ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX



Gambar 9.Grafik Perbandingan Nilai Akurasi

Pada Gambar 9 ditunjukkan perbnadingan nilai akurasi skenario 1 mencapai 97,065 %, sedangkan skenario 2 mencapai nilai akurasi sebesar 97,464 %. Selisih dari kedua skenario mencapai 0,399 %. Perhitungan nilai akurasi didapat dari pembagian banyaknya prediksi yang sesuai dengan banyaknya semua prediksi. Perbedaan nilai akurasi antara skenario 1 dan 2 dikarenakan perbedaan banyaknya data yang digunakan dan keseimbangan data antar kelas. Skenario 1 menggunakan 100.000 data dengan rincian data training sebanyak 80.000 data dan data testingnya sebanyak 20.000 data. Skenario 2 menggunakan 183.000 data dengan rincian 146.400 data training dan 36.600 data testing. Skenario 1 menggunakan data yang belum seimbang. Kelas penderita mempunyai banyak data yang lebih sedikit daripada kelas bukan penderita. Skenario 2 menggunakan data yang sudah diseimbangkan. Banyaknya data pada kelas penderita sama dengan dengan kelas bukan penderita. Selain nilai akurasi, nilai presisi pada kedua skenario juga mengalami perubahan. Model yang dihasilkan dari kedua skenario masing-masing berupa gabungan beberapa (*ensemble*) pohon keputusan (tree)[25]. Perbedaaannya hanya pada data training yang digunakan pada tahap pre-processing. Perbandingan nilai presisi pada kedua skenario ditampilkan pada gambar 10.

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX



Gambar 10. Grafik Perbandingan Nilai Presisi

Pada Gambar 10 ditunjukkan pada Skenario 1 kelas bukan penderita menghasilkan nilai presisi sebesar 1 sedangkan kelas bukan penderita mempunyai nilai presisi sebesar 0,69 atau 69 % Pada Skenario 2 kelas bukan penderita menghasilkan nilai presisi sebesar 0,98 atau 98 % sedangkan kelas bukan penderita mempunyai nilai presisi sebesar 0,97 atau 97 %. Perbandingan antara skenario 1 dan 2 adalah kelas bukan penderita mengalami penurunan sebesar 0,02 atau 2 % sedangkan kelas penderita mengalami kenaikkan sebesar 0,28 atau 28 %. Skenario 1 dan 2 menunjukkan nilai akurasi yang tidak terlalu berbeda (sigifikan). Perbedaan signifikan terlihat pada nilai presisi pada setiap kelasnya.

Temuan dari penelitian ini adalah perbedaan nilai presisi menunjukkan pentingnya keseimbangan data training yang akan digunakan. Hasil model yang telah ditraining dengan data yang seimbang akan mempunyai kemampuan untuk memprediksi data testing lebih baik[26]. Keseimbangan nilai presisi menunjukkan keakuratan nilai hasil prediksi pada setiap kelasnya.

IV. KESIMPULAN

Imbalanced data dalam klasifikasi dapat mempengaruhi hasil prediksi. Kombinasi SMOTE dan Algoritma Random Forest dapat mengatasi masalah imbalanced data pada dataset prediksi penyakit diabetes. Algoritma Random forest dengan imbalanced data mengahsilkan nilai akurasi yang tinggi yaitu 97,1 % akan tetapi mempunyai nilai presisi yang rendah pada kelas minoritas (kelas penderita) yaitu 69 %. Algoritma Random Forest dengan balanced data menggunakan SMOTE menghasilkan akurasi sebesar 97,5 % dan mempunyai nilai presisi yang seimbang pada kedua kelas, yaitu kelas bukan penderita sebesar 97 % dan kelas penderita sebesar 98 %. Hasil penelitian ini membuktikan bahwa optimasi algoritma random forest dengan SMOTE dapat meningkatkan nilai akurasi dan keseimbangan nilai presisi pada setiap kelasnya. Penelitian selanjutnya adalah menggunakan algoritma *ensemble* yang lain, misalnya menggunakan

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

AdaBoost, *Light Gradien Boosting Machine* (LGBM) dan menggunakan dataset yang lain untuk menguji nilai akurasi dan keseimbangan nilai presisi pada kedua kelasnya.

DAFTAR PUSTAKA

[1] R. O'Brien and H. Ishwaran, "A random forests quantile classifier for class imbalanced data," *Pattern Recognition*, vol. 90, pp. 232–249, 2019, doi: 10.1016/j.patcog.2019.01.036.

[2] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. 2018. doi: 10.1007/978-3-319-98074-4_5.

[3] U. Nila, R. Firliana, and S. Sucipto, "Analisis Data Transaksi Penjualan Produk Pertanian Menggunakan Algoritma FP-Growth," *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, vol. 7, no. 1, pp. 175–183, Jul. 2023, doi: 10.29407/INOTEK.V7I1.3426.

[4] J. Brownlee, "Imbalanced Classification with Python Choose Better Metrics, Balance Skewed Classes , and Apply Cost-Sensitive Learning".

[5] Y. Jian, M. Ye, Y. Min, L. Tian, and G. Wang, "FORF-S: A Novel Classification Technique for Class Imbalance Problem," *IEEE Access*, vol. 8, pp. 218720–218728, 2020, doi: 10.1109/ACCESS.2020.3040978.

[6] L. K. Xin and N. binti A. Rashid, "Prediction of depression among women using random oversampling and random forest," *2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021*, 2021, doi: 10.1109/WIDSTAIF52235.2021.9430215.

[7] N. S. S. Pranavi, T. K. S. S. Sruthi, B. J. Naga Sirisha, M. S. Nayak, and V. S. Gupta Thadikemalla, "Credit Card Fraud Detection Using Minority Oversampling and Random Forest Technique," in *2022 3rd International Conference for Emerging Technology (INCET)*, May 2022, pp. 1–6. doi: 10.1109/INCET54531.2022.9824146.

[8] A. Amin *et al.*, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access*, vol. 4, no. c, pp. 7940–7957, 2016, doi: 10.1109/ACCESS.2016.2619719.

[9] M. Molinier and J. Kilpi, "Avoiding overfitting when applying spectral-spatial deep learning methods on hyperspectral images with limited labels," *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2019-Janua, pp. 5049–5052, 2019, doi: 10.1109/IGARSS.2019.8900328.

[10] H. Hairani and T. Widiyaningtyas, "Augmented Rice Plant Disease Detection with Convolutional Neural Networks," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan*

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v2i2.XX

Teknologi Sistem Informasi, vol. 8, no. 1, pp. 27–39, Feb. 2024, doi: 10.29407/INTENSIF.V8I1.21168.

[11] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[12] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," *Journal of Information Technology Ampera*, vol. 1, no. 3, pp. 133–143, 2020, doi: 10.51519/journalita.volume1.isssue3.year2020.page133-143.

[13] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[14] N. Iriawan *et al.*, "On The Comparison: Random Forest, SMOTEBagging, and Bernoulli Mixture to Classify Bidikmisi Dataset in East Java," *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2018 - Proceeding*, pp. 137–141, 2018, doi: 10.1109/CENIM.2018.8711035.

[15] K. Vijiyakumar, B. Lavanya, I. Nirmala, and S. Sofia Caroline, "Random forest algorithm for the prediction of diabetes," *2019 IEEE International Conference on System, Computation, Automation and Networking, ICSCAN 2019*, pp. 1–5, 2019, doi: 10.1109/ICSCAN.2019.8878802.

[16] S. C. Gupta and N. Goel, "Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method," *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, no. Icssit, pp. 980–986, 2020, doi: 10.1109/ICSSIT48917.2020.9214129.

[17] S. Hosseini, B. Turhan, and D. Gunarathna, "A systematic literature review and meta-analysis on cross project defect prediction," *IEEE Transactions on Software Engineering*, vol. 45, no. 2, pp. 111–147, 2019, doi: 10.1109/TSE.2017.2770124.

[18] D. O. Ratmana, G. Fajar Shidik, A. Z. Fanani, Muljono, and R. A. Pramunendar, "Evaluation of feature selections on movie reviews sentiment," *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, pp. 567–571, 2020, doi: 10.1109/iSemantic50169.2020.9234287.

[19] A. Nugroho and A. Husin, "Analisis Performa Random Forest Menggunakan

Normalisasi Atribut," *SISTEMASI: Jurnal Sistem Informasi*, vol. 11, no. 1, pp. 186–196, 2022, doi: https://doi.org/10.32520/stmsi.v11i1.1681.

[20] S. Sucipto, D. D. Prasetya, and T. Widiyaningtyas, "Educational Data Mining: Multiple Choice Question Classification in Vocational School," *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 23, no. 2, pp. 367–376, 2024, doi: 10.30812/matrik.v23i2.3499.

[21] A. Nugroho, A. Z. Fanani, and G. F. Shidik, "Evaluation of Feature Selection Using Wrapper for Numeric Dataset with Random Forest Algorithm," *Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021*, pp. 179–183, 2021, doi: 10.1109/iSemantic52711.2021.9573249.

[22] A. Nugroho, M. A. Soeleman, R. A. A. Pramunendar, and A. Nurhindarto, "Peningkatan Performa Ensemble Learning Pada Segmentasi Semantik Gambar Dengan Teknik Oversampling Untuk Class Imbalance." Jurnal Teknologi Informasi dan Ilmu Komputer(JTIIK), pp. 899–908, 2023. doi: 10.25126/jtiik.2023106831.

[23] A. R. Ismail, A. Zainul Fanani, G. F. Shidik, and Muljono, "Implementation of naive bayes algorithm with particle swarm optimization in classification of dress recommendation," *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, pp. 174–178, 2020, doi: 10.1109/iSemantic50169.2020.9234293.

[24] Z. Bingzhen, Q. Xiaoming, Y. Hemeng, and Z. Zhubo, "A random forest classification model for transmission line image processing," *15th International Conference on Computer Science and Education, ICCSE 2020*, no. Iccse, pp. 613–617, 2020, doi: 10.1109/ICCSE49874.2020.9201900.

[25] Z. Xu and Z. Wang, "A Risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier," *11th International Conference on Advanced Computational Intelligence, ICACI 2019*, pp. 278–283, 2019, doi: 10.1109/ICACI.2019.8778622.

[26] F. Khanam and Md. R. H. Mondal, "Ensemble Machine Learning Algorithms for the Diagnosis of Cervical Cancer," in *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, Aug. 2021, pp. 1–5. doi: 10.1109/ICSCT53883.2021.9642612.