JURNAL NUANSA INFORMATIKA
Volume 14 Nomor 1, Januari 2020
p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

INTEGRASI NAÏVE BAYES DENGAN TEKNIK SAMPLING SMOTE UNTUK MENANGANI DATA TIDAK SEIMBANG

Nina Sulistiyowati[1], Mohamad Jajuli[2]
Teknik Informatika Universitas Singaperbangsa Karawang
Email: nina.sulistio@unsika.ac.id[1], mohamad.jajuli@unsika.ac.id[2]

Klasifikasi pada data dengan kelas tidak seimbang merupakan masalah utama pada bidang *machine learning* dan *data mining*. Jika bekerja pada data tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas. Penelitian ini akan mengimplementasikan metode *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi data tidak seimbang pada data nasabah kredit di koperasi guru Rawamerta. Metodologi penelitian menggunakan SEMMA dengan tahapan penelitian *Sample*, Explore, *Modify*, *Model*, dan *Asses*. Tahap *Sample* dilakukan pemilihan data nasabah kredit Koperasi Guru Rawamerta tahun 2015-2017 dengan total data sebanyak 878 data dengan atribut yang digunakan yaitu pendapatan, jumlah simpanan, jumlah pinjaman, lama angsuran, jasa, cicilan, dan status kredit. Tahap *Explore* menganalisis kelas lancar yang dikategorikan sebagai kelas mayoritas karena berjumlah 813 data, sedangkan kelas macet dapat dikategorikan sebagai kelas minoritas karena berjumlah 65 data. Data tersebut menunjukkan adanya ketidakseimbangan data diantara kedua kelas. Tahapan *Modify* melakukan proses SMOTE 500%. Tahap *Model* melakukan klasifikasi menggunakan *Naïve Bayes*. Pemodelan *naïve bayes* dengan SMOTE menghasilkan 1131 data berhasil diklasifikasikan dengan benar dan 72 data tidak diklasifikasikan dengan benar sedangkan tanpa SMOTE menghasilkan 818 data berhasil diklasifikasikan dengan benar dan 60 data tidak diklasifikasikan dengan benar.
Kata kunci: Naïve Bayes, SMOTE, data tidak seimbang

Classification of data with unbalanced classes is a major problem in the field of machine learning and data mining. If working on unbalanced data, almost all classification algorithms will produce much higher accuracy for majority classes than minority classes. This research will implement the Synthetic Minority Over-sampling Technique (SMOTE) method to overcome unbalanced data on credit customer data in Rawamerta teacher cooperatives. The research methodology uses SEMMA with the stages of research Sample, Explore, Modify, Model, and Asses. The Sample Phase was conducted to choose the data of the Rawamerta Teachers Cooperative credit customers for 2015-2017 with a total of 878 data with the attributes used namely income, total deposits, loan amount, duration of installments, services, installments, and credit status. The Explore phase analyzes current classes which are categorized as majority classes because there are 813 data, while traffic classes can be categorized as minority classes because there are 65 data. The data shows an imbalance of data between the two classes. The Modify stages perform the 500% SMOTE process. The Model Stage classifies using Naïve Bayes. Naïve Bayes modeling with SMOTE produced 1131 successfully classified data correctly and 72 data were not classified correctly while without SMOTE resulted in 818 data was classified correctly and 60 data were not classified correctly.
Keywords: Naïve Bayes, SMOTE, unbalanced data

1. PENDAHULUAN

Klasifikasi pada data dengan kelas tidak seimbang merupakan masalah utama pada bidang *machine learning* dan *data mining*. Jika bekerja pada data tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas (Q. Gu, X.-M. Wang, Z. Wu, B. Ning, C.-S. Xin). Perbedaan ini merupakan suatu indikator performa klasifikasi yang buruk. Pada beberapa kasus, kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas. Untuk mengatasi permasalahan tersebut dapat digunakan dua pendekatan yaitu, pendekatan sample dan algoritma (Chawla et al. 2002). Pada penelitian ini

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

akan dilakukan pendekatan sampel yaitu menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi data tidak seimbang pada data nasabah kredit di koperasi guru Rawamerta. Metode SMOTE merupakan metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan *dataset* dengan cara sampling ulang sampel kelas minoritas. Rumusan masalah pada penelitian ini adalah: (1) Bagaimana klasifikasi penentuan potensi kredit di koperasi guru Rawamerta dengan data tidak seimbang dan data yang seimbang; (2) Bagaimana evaluasi klasifikasi penentuan potensi kredit di koperasi guru Rawamerta dengan data tidak seimbang dan data yang seimbang.

Tujuan penelitian ini adalah: (1) Menerapkan algoritma Naïve Bayes dan SMOTE untuk mengklasifikasikan penentuan potensi kredit di koperasi guru Rawamerta pada data tidak seimbang dan data yang seimbang; (2) Mengetahui evaluasi klasifikasi algoritma Naïve Bayes dan SMOTE menggunakan *accuracy* dan G-Mean.

2. METODOLOGI PENELITIAN

Metodologi penelitian ini adalah SEMMA dengan tahapan penelitiannya adalah *Sample*, *Explore*, *Modify*, *Model*, dan *Assess*.

3. HASIL DAN PEMBAHASAN

3.1 Sample

Pada tahap ini dilakukan pemilihan data yang sesuai dengan tujuan penelitian. Data yang digunakan merupakan data nasabah kredit dari tahun 2015-2017 yang bersumber dari Koperasi Guru Rawamerta dengan total data sebanyak 878 data. Data yang dipilih berjumlah 7 atribut yaitu pendapatan, jumlah simpanan, jumlah pinjaman, lama angsuran, jasa, cicilan, dan status kredit.

3.2 Explore

Nasabah dapat dikategorikan ke dalam dua kategori status kredit yaitu nasabah yang memiliki status kredit lancar dan nasabah yang memiliki status kredit macet. Nasabah dengan status kredit lancar adalah nasabah yang mampu membayar cicilan sesuai dengan lama angsuran yang telah disepakati sedangkan nasabah dengan status kredit macet adalah nasabah yang tidak dapat membayar cicilan sesuai dengan lama angsuran yang telah disepakati, atribut status kredit menjadi kelas target yang mana memiliki dua kelas yaitu lancar dan macet. Kelas lancar dapat dikategorikan sebagai kelas mayoritas karena berjumlah 813 data, sedangkan kelas macet dapat dikategorikan sebagai kelas minoritas karena berjumlah 65 data. Data tersebut menunjukkan adanya ketidakseimbangan data diantara kedua kelas.

Gambar 1. Persentase Data Nasabah Kredit

3.3 Modify

Pada tahapan ini melakukan proses SMOTE sebesar 500% yang berfungsi untuk menambah data buatan ke dalam data minoritas agar jarak jumlah data antar mayoritas ke minoritas tidak begitu jauh. Cara kerja SMOTE adalah dengan menggunakan teknik membangkitkan amatan buatan pada kelas minor berdasarkan konsep *k*-tetangga terdekat. Nilai *k* tetangga terdekat dipilih secara acak tergantung dari diperlukannya pengambilan *over-sampling* (Chawla et. al., 2002). Jumlah *k*-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan dalam melaksanakannya (Fuadin, 2017).

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

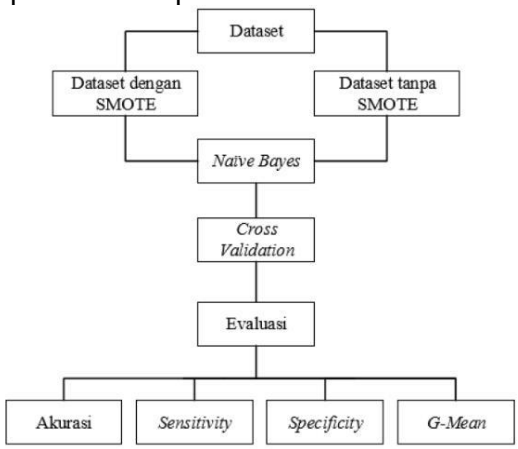
Tabel 1. Perbandingan Data Awal dengan Data SMOTE 500%

Kelas	Data Awal	SMOTE 500%
Mayor	813 (93%)	813 (68%)
Minor	65 (7%)	390 (32%)
Jumlah	878 (100%)	1203 (100%)

Pada Tabel 1 terjadi penambahan data kelas minoritas dari yang awalnya 65 data menjadi 390 data hampir mendekati keadaan *balanced data*.

3.4 Model

Pada tahapan ini dibuat dua skenario seperti terlihat pada Gambar 2.



Gambar 2. Skenario Model Penelitian

Skenario penelitian membandingkan antara dataset yang sudah dilakukan proses sampling SMOTE dan dataset yang tidak dilakukan proses sampling SMOTE untuk diuji dengan algoritma klasifikasi yaitu *Naïve Bayes*, kemudian model yang didapat dari algoritma tersebut divalidasi menggunakan 10-*fold cross validation*, karena metode ini telah menjadi metode standar dalam hal praktis (Putri & Wahono, 2015). Dengan bantuan software WEKA didapat hasil klasifikasi Naïve Bayes seperti pada Gambar 3 dan 4.

```
Correctly Classified Instances        1131               94.015  %
Incorrectly Classified Instances        72                5.985  %
Kappa statistic                          0.8678
Mean absolute error                      0.0595
Root mean squared error                  0.2287
Relative absolute error                 13.5687 %
Root relative squared error             48.867  %
Total Number of Instances             1203
```

Gambar 3. Hasil Naïve Bayes+SMOTE

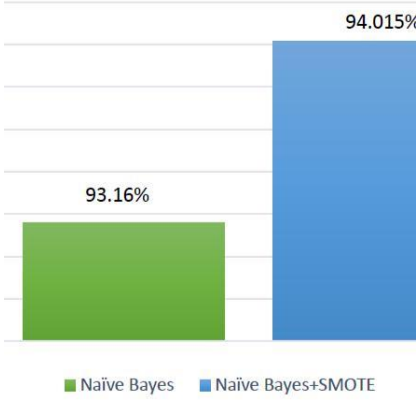
```
Correctly Classified Instances         818               93.1663 %
Incorrectly Classified Instances        60                6.8337 %
Kappa statistic                          0.628
Mean absolute error                      0.0659
Root mean squared error                  0.2395
Relative absolute error                 47.7407 %
Root relative squared error             91.4529 %
Total Number of Instances              878
```

Gambar 4. Hasil Naïve Bayes tanpa SMOTE

Berdasarkan Gambar 3 dan 4 dapat dilihat bahwa hasil yang diperoleh dari pemodelan *Naïve Bayes*+SMOTE menghasilkan persentase *correctly classified instances* sebesar 94.01% sedangkan persentase *incorrectly classified instances* sebesar 5.98% atau dari total 1203 data terdapat 1131 data yang berhasil diklasifikasikan dengan benar dan 72 data tidak diklasifikasikan dengan benar sedangkan pemodelan *Naïve Bayes* tanpa SMOTE menghasilkan persentase *correctly classified instances* sebesar 93.16% sedangkan persentase *incorrectly classified instances* sebesar 6.83% atau dari total 1203 data terdapat 818 data yang berhasil diklasifikasikan dengan benar dan 60 data tidak diklasifikasikan dengan benar

3.5 Evaluation

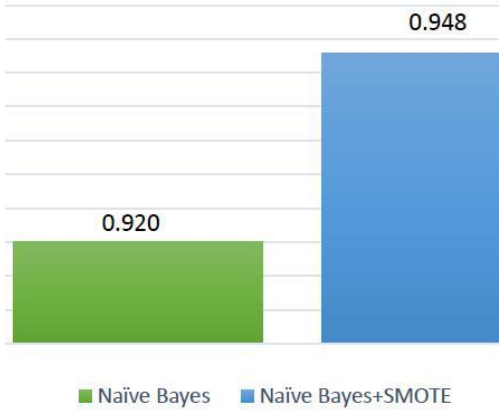
Berikut hasil evaluasi berdasarkan dua skenario penelitian yang telah dilakukan.



Gambar 5. Hasil Evaluasi *Accuracy*

Accuracy adalah persentase dari total data yang diprediksi secara benar. Berdasarkan Gambar 5 menunjukkan bahwa accuracy Naïve Bayes+SMOTE lebih tinggi dibandingkan Naïve Bayes tanpa SMOTE dengan selisih 0.855%.

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom



Gambar 6. Hasil Evaluasi *G-Mean*

G-Mean digunakan untuk mengukur performa keseluruhan (*overall classification performance*). Gambar 6 menunjukkan bahwa nilai *G-Mean* Naïve Bayes+SMOTE lebih tinggi dibandingkan Naïve Bayes tanpa SMOTE dengan selisih 0.028.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan maka didapatkan kesimpulan sebagai berikut:

1. Penelitian ini menggunakan metode SMOTE untuk menambah data sintetis pada kelas minoritas agar menjadi seimbang dengan kelas mayoritas yang mana terdapat 813 data nasabah kredit macet dan 65 data nasabah kredit lancar.
2. Terdapat dua skenario penelitian yang dilakukan, skenario pertama melakukan klasifikasi Naïve Bayes pada dataset asli sedangkan skenario kedua melakukan klasifikasi Naïve Bayes pada dataset yang sudah dilakukan metode SMOTE.
3. Secara keseluruhan metode SMOTE mampu menangani permasalahan *imbalanced data* dan SMOTE baik dikombinasikan dengan algoritma klasifikasi *Naïve Bayes* yang mana memiliki tingkat akurasi 94.015% dan *G-mean* 0.948.

5. SARAN

Dari hasil penelitian yang dilakukan, maka saran yang dapat diberikan untuk penelitian selanjutnya, yaitu:

1. Pengujian juga dapat dicoba menggunakan nilai persentase SMOTE lebih besar dari 500% dan menggunakan nilai *k* selain 5.
2. Pengujian pada *data mining* dilakukan dengan beberapa algoritma klasifikasi untuk melihat perbandingan algoritma mana yang lebih baik dikombinasikan dengan SMOTE.
3. Pengujian juga dapat diuji coba dengan berbagai variasi nilai pada *k-fold cross validation* untuk memperoleh pemahaman yang baik

DAFTAR PUSTAKA

[1] Anindya, A., Indahwati, & Suteyo, B. (2018). Application of SMOTE on CART Method to Handle Imbalanced Data (Study Case: Labor Force Classification in Banten Province). *IOP Conference Series: Earth and Environmental Science*, 1-18.

[2] Bramer, M. (2007). *Principles of Data Mining.* Springer Science.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artifical Intelligence Research, 16*, 321-357.

[4] Defiyanti, S., & Jajuli, M. (2015). Integrasi Metode Klasifikasi Data Clustering dalam Data Mining. *Konferensi Nasional Informatika (KNIF)*, 39-44.

[5] Fuadin, D. N. (2017). *Deteksi Botnet Menggunakan Naive Bayes Classifier dengan SMOTE dan Metode BFS.* Surabaya: Institut Teknologi Sepuluh Nopember.

[6] Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques.* Romania: Springer-Verlag Berlin Heidelberg.

[7] Hairani, Setiawan, N. A., & Adji, T. B. (2016). Metode Klasifikasi Data Mining dan Teknik Sampling SMOTE Menangani Class Imbalance Untuk Segmentasi Customer Pada Industri Perbankan. *Prosiding SNST, 7*, 168-172