ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 4, Tahun 2022, Halaman 499 - 505
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PERBANDINGAN SMOTE DAN ADASYN PADA DATA *IMBALANCE* UNTUK KLASIFIKASI RUMAH TANGGA MISKIN KABUPATEN TEMANGGUNG DENGAN ALGORITMA *K-NEAREST NEIGHBOR*

Dinda Virrliana Ramadhanti[1*], Rukun Santoso[2], Tatik Widiharih[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail** : 208.dindavirrliana@gmail.com

DOI: 10.14710/j.gauss.11.4.499-505

Article Info:
Received: 2022-08-23
Accepted: 2022-10-02
Available Online: 2023-02-25

Keywords:
K Nearest Neighbor; Household Poverty; Imbalanced data; SMOTE; ADASYN.

Abstract: Poverty is a global problem that has occurred in various countries with various impacts. Poverty conditions are characterized by the inability of a person or household to meet the basic needs of life. Socio-economic problems, such as poverty, can be handled using machine learning, one of which is classification. The classification of households based on poverty criteria is expected to assist the government in preparing programs that are right on target. *K-Nearest Neighbor* is one of the easy-to-use classification algorithms. this classification is based on the closest neighborliness. The problem that can be experienced when classifying is if the data used is imbalanced. The data imbalance will causing the classification process to focus more on the majority class. SMOTE and ADASYN are used to solve the problem of imbalanced data. This study resulted in the addition of SMOTE and ADASYN to imbalanced data can improve classification performance, especially on the G-mean value. G-mean is a performance measure that is widely used in the case of imbalanced data. The result of this study is that SMOTE can increase the G-mean value to 58.5%, while ADASYN is 57.3%. Therefore, it can be concluded that SMOTE-KNN is the best classification model for household poverty classification.

1. PENDAHULUAN

Kemiskinan (*poverty*) merupakan salah satu masalah dunia yang harus diselesaikan dan menjadi indikator *Sustainable Development Goals* (SDG's) yaitu menghapus kemiskinan. Kemiskinan merupakan ketidakmampuan dari sisi ekonomi, serta memenuhi kebutuhan makanan dan non-makanan yang diukur berdasarkan pengeluaran. (BPS, 2021). Kemiskinan ialah situasi kehidupan yang dialami seseorang atau rumah tangga yang serba kekurangan sehingga tidak mampu memenuhi kebutuhan dasar. Kebutuhan dasar meliputi pangan, Kesehatan, Pendidikan, perumahan, pekerjaan, air bersih, pertanahan, sumber daya alam, lingkungan hidup (Ispriyanti *et al*, 2019). Badan Pusat Statistik membagi kategori rumah tangga menjadi dua berdasarkan rata rata pengeluaran rumah tangga perkapita perbulan dibandingkan dengan garis kemiskinan, yaitu rumah tangga miskin jika pengeluaran perkapita berada di bawah garis kemiskinan, sedangkan rumah tangga tidak miskin jika pengeluaran perkapita berada di bawah garis kemiskinan.

Kabupaten Temanggung merupakan kabupaten yang memiliki persentase kemiskinan 9,96% pada tahun 2020. Kabupaten Temanggung memiliki garis kemiskinan sebesar Rp323.705 kapita/bulan pada tahun 2020. Pemerintah telah merancang beberapa program untuk mengatasi masalah kemiskinan, namun sering kali ditemukan program yang salah sasaran dikarenakan kesulitan untuk menentukan dan memilah kategori rumah tangga.

Machine learning digunakan untuk memudahkan dalam menyelesaikan masalah pengkategorian rumah tangga, yaitu metode klasifikasi. Klasifikasi adalah proses yang

bertujuan untuk mendeskripsikan dan membedakan kelas data untuk masalah tertentu. Salah satu metode klasifikasi yaitu *K-Nearest Neighbor* (KNN) yang menggunakan konsep kedekatan tetangga. Kelas baru akan ditetapkan berdasarkan kedekatan kesamaannya dengan titik set pelatihan. Ukuran performa algoritma dapat dilihat melalui nilai akurasinya. Permasalahan yang berpengaruh dalam proses klasifikasi adalah adanya data *imbalance*. Data *imbalance* merupakan kondisi ketika banyaknya label kelas data latih yang satu lebih sedikit daripada banyaknya label kelas data yang lainnya. Menurut He & Ma (2013) algoritma klasifikasi akan menghasilkan suatu model dengan tingkat kepekaan yang minim terhadap kelas minoritas. Penanganan data *imbalance* dapat dilakukan dengan oversampling yaitu SMOTE (*Synthetic Minority Over-sampling Technique*) dan ADASYN (*Adaptive Syntethic Sampling Approach).*

Penelitian ini bertujuan untuk mengklasifikasi kategori rumah tangga di Kabupaten Temanggung serta mengatasi data *imbalance* dengan SMOTE dan ADASYN. Hasil dari penelitian ini adalah menentukan metode terbaik untuk klasifikasi kategori rumah tangga di Kabupaten Temanggung dengan Algoritma KNN, SMOTE-KNN, dan ADASYN-KNN.

2. TINJAUAN PUSTAKA

Klasifikasi merupakan pemodelan yang membedakan kelas data dan bertujuan agar dapat digunakan untuk prediksi kelas yang label kelasnya belum diketahui (Han & Kamber, 2006). *K-Nearest Neighbor*s adalah algoritma non parametrik yang menggunakan kesamaan fitur untuk memprediksi nilai titik data baru (Patgiri, 2021). Jumlah tetangga terdekat (*nearest neighbor*) dinyatakan dengan nilai k. Label kelas hasil prediksi ditentukan dari anggota tetangga yang memiliki jumlah paling banyak. Klasifikasi *K-Nearest Neighbor*s juga didasarkan pada jarak, perhitungan jarak pada penelitian ini menggunakan jarak Euclidean. Peringkat untuk k tetangga terdekat berdasarkan nilai kesamaan dihitung menggunakan jarak Euclidean yang didefinisikan dengan Persamaan (1) (Sreemathy, 2012).

$$d(\boldsymbol{x_i}, \boldsymbol{y_i}) = \sqrt{\sum_{l=1}^{N} (diff(x_{il}, y_{il}))^2} \quad (1)$$

dengan $diff(x_{il}, y_{il})$ adalah nilai ketidaksamaan antara data uji ke-i pada variabel ke-l (x_{il}) dan data latih ke-i pada variabel ke- l (y_{il}).

Perhitungan nilai ketidaksamaan berdasarkan tipe untuk tiap variabel disajikan pada Tabel 1 (Prasetyo, 2012).

Tabel 1 Ketidaksamaan Dua Data dengan Satu Atribut

Tipe Atribut	Formula Jarak
Nominal	$diff(x_{il}, y_{il}) = \begin{cases} 0, Jika\ x_{il} = y_{il} \\ 1, Jika\ x_{il} \neq y_{il} \end{cases}$
Ordinal	$diff(x_{il}, y_{il}) = \lvert x_{il} - y_{il} \rvert / (n-1)$ n adalah banyaknya pengkategorian dalam x
Interval atau Rasio	$diff(x_{il}, y_{il}) = \lvert x_{il} - y_{il} \rvert$

Randomized Search merupakan metode yang dapat digunakan untuk mencari parameter terbaik, dengan mencobakan beberapa kombinasi parameter yang mungkin menghasilkan model terbaik. Untuk algoritma *K-Nearest Neighbor* terdapat parameter yang ingin dicari nilainya agar akurasi model yang diperoleh maksimal yaitu jumlah tetangga (k). Sehingga tujuan dari tuning hyperparameter pada penelitian ini untuk mencari nilai k yang optimal. Pemilihan parameter optimal pada *RandomSearch* dilakukan secara acak, sehingga memiliki keunggulan yaitu lebih efektif digunakan karena waktu komputasinya jauh lebih cepat.

Permasalahan yang dapat terjadi dalam proses klasifikasi yaitu data *imbalance*. Data *imbalance* merupakan kasus yang dapat terjadi dalam proses klasifikasi data riil. Permasalahan ini terjadi ketika jumlah data dalam satu kelas jauh lebih tinggi (*majority class*) atau lebih rendah (*minority class*) dibandingkan kelas lainnya. Data yang tidak seimbang menyebabkan algoritma tersebut gagal untuk mewakili karakteristik data distribusif secara tepat dan menghasilkan akurasi yang buruk di seluruh kelas data (He & Gracia, 2009).

Synthetic Minority Over Sampling Technique (SMOTE) merupakan teknik untuk mengatasi masalah data *imbalance* dengan melakukan *oversampling* pada kelas minoritas yaitu dengan membuat sampel sintesis. Pembangkitan sintetis data dilakukan dengan menggunakan Persamaan (2) (Choi, 2010).

$$x_{syn_j} = x_i + (x_{knn} - x_i) \times \gamma \quad (2)$$

x_{syn_j} adalah data hasil sintetis hasil dari SMOTE, x_i merupakan data pengamatan ke-i dari kelas minor, x_{knn} adalah data dari kelas minor yang memiliki jarak terdekat dengan x_i, dan gamma (γ) adalah bilangan random antara 0 dan 1.

Adaptive Syntethic Sampling Approach (ADASYN) merupakan teknik untuk mengatasi masalah data *imbalance* dengan melakukan *oversampling* pada kelas minoritas yaitu dengan menggunakan bobot distribusi untuk data pada kelas minoritas berdasarkan pada tingkat kesulitan pembelajaran model. Data sintesis dihasilkan dari data minoritas yang sulit untuk dipahami dibandingkan dengan data minoritas yang lebih mudah untuk dipahami (He *et al*, 2008). ADASYN memiliki parameter yang digunakan untuk menentukan tingkat keseimbangan yang diharapkan (β) dan sebuah batas yang ditetapkan sebagai derajat toleransi maksimum rasio ketidakseimbangan kelas (d_{th}).

Langkah pembangkitan data sintetis dengan ADASYN adalah sebagai berikut:

1. Menentukan nilai parameter dari ADASYN, yaitu nilai d_{th} (nilai dari maksimal toleran data *imbalance*) dan β (nilai level keseimbangan)
2. Menghitung derajat keseimbangan $d = \frac{m_{minoritas}}{m_{mayoritas}}$
3. Menghitung banyaknya *instance* data sintetis yang akan dibuat untuk kelas minoritas $G = (m_{mayoritas} - m_{minoritas}) \times \beta$
4. Menghitung rasio berdasarkan *K-Nearest Neighbor* menggunakan *Euclidean distance* $r_i = \frac{\Delta_i}{k}$, dengan Δ_i adalah banyaknya instance pada *nearest neighbors* yang termasuk kedalam kelas, dan i = 1, 2, 3,…,$m_{minoritas}$
5. Normalisasi r_i sehingga $\hat{r}_i$ adalah distribusi kerapatan, $\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_{minoritas}} r_i}$
6. Menghitung banyaknya *instance* data sintetis yang perlu dibangkitkan untuk setiap *instance* minoritas, $g_i = \hat{r}_i \times G$, i = 1, 2, 3,…,$m_{minorita}$
7. Pembangkitan sampel data sintetis dilakukan dengan Persamaan (3).

$$s_i = x_i + (x_{ui} - x_i) \times \lambda \quad (3)$$

x_i adalah data hasil sintetis hasil dari ADASYN, x_i merupakan data pengamatan ke-i dari kelas minor, x_{ui} adalah data ke-i dari data latih yang dipilih secara acak, dan lambda (λ) adalah bilangan random antara 0 dan 1

Pada proses pelatihan tidak seluruh data digunakan, salah satu cara untuk melakukan pembagian data train dan data test adalah dengan *Holdout validation*. *Holdout validation* bekerja dengan membagi data menjadi 2 bagian, yaitu data train dan data test secara sederhana dengan proporsi yang telah ditentukan oleh peneliti. Selain *holdout validation*,

data juga dapat dibagi dengan metode *K-fold cross validation*. Secara umum, 10-*fold cross-validation* direkomendasikan karena bias dan variasi yang relatif rendah (Han et al., 2006). Pada setiap *fold* akan didapatkan ukuran kinerja yang selanjutnya digunakan untuk menghitung rata-rata dari ukuran kinerja klasifikasi.

Evaluasi hasil diperlukan guna melihat dan menilai kinerja dari model berdasarkan data yang digunakan. Cara untuk evaluasi ukuran kinerja algoritma yaitu dengan *confusion matrix* seperti pada Tabel 2.

Tabel 2 *Confusion Matrix* Klasifikasi Biner

		Kelas Prediksi	
		+	-
Kelas Sebenarnya	+	TP (*True Posistive)*	FN (*False Negative*)
	-	FP (*False Positive*)	TN *(True Negative)*

Ukuran kinerja model dihitung berdasarkan akurasi, spesifitas, sensitivitas, dan *g-mean* dengan rumus sebagai berikut:

$$accuracy = \frac{TP+TN}{Total} \tag{4}$$

$$specificity = \frac{TN}{TN+FP} \tag{5}$$

$$sensitivity = \frac{TP}{TP+FN} \tag{6}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{7}$$

Akurasi adalah nilai sebagai pengukur model untuk menentukan seberapa akurat dalam melakukan prediksi. Spesifikasi merupakan perhitungan yang digunakan untuk mengevaluasi seberapa besar keberhasilan suatu model dalam memprediksi kelas negatif. Sensitivitas merupakan proporsi data positif yang terprediksi dengan benar sebagai data positif. G-mean merupakan rata rata geometric dari sensitivitas dan spesifisitas, apabila semua amatan dapat diklasifikasikan secara tepat maka G-Mean akan bernilai satu (Kubat & Matwin, 1997)

3. METODE PENELITIAN

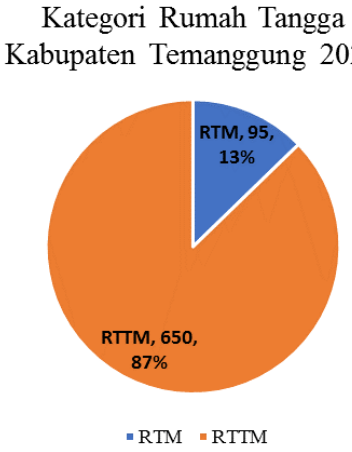
Data yang digunakan dalam penelitian ini adalah data kategori rumah tangga yaitu data Survey Sosial Ekonomi Nasional (Susenas) Kabupaten Temanggung periode Maret 2020 yang diperoleh melalui website silastik.bps.go.id. Data berjumlah 745 amatan rumah tangga, dan terdiri dari dua variabel, yaitu variabel dependen dan variabel independent. Variabel dependen yaitu kategori rumah tangga yang terbagi menjadi dua kategori yaitu rumah tangga miskin dan rumah tangga tidak miskin. Variabel independent terdiri dari 12 variabel yaitu jumlah anggota rumah tangga, usia kepala keluarga, pendidikan kepala keluarga, pekerjaan kepala keluarga, status kepemilikan rumah, luas lantai (m2/ kapita), jenis dinding, sumber air minum, bahan bakar memasak, jenis lantai, pemilikan aset, fasilitas buang air besar. Dataset yang telah diperoleh kemudian dilakukan pengolahan dengan langkah langkah sebagai berikut

Pengolahan data pada penelitian ini menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*), ADASYN (*Adaptive Syntethic Sampling Approach)*, dan *K-Nearest Neighbor*. Langkah analisis data penelitian ini yaitu sebagai berikut:

1. Memasukkan dataset
2. Melakukan analisis deskriptif
3. Membagi data latih dan data uji
4. Melakukan klasifikasi dengan *K-Nearest Neighbor*
5. Melakukan pengukuran kinerja klasifikasi *K-Nearest Neighbor*
6. Melakukan klasifikasi dengan SMOTE-KNN
7. Melakukan pengukuran kinerja klasifikasi SMOTE-KNN
8. Melakukan klasifikasi dengan ADASYN-KNN
9. Melakukan pengukuran kinerja klasifikasi ADASYN-KNN
10. Melakukan perbandingan kinerja klasifikasi KNN, SMOTE-KNN, dan ADASYN-KNN untuk memilih model terbaik

4. HASIL DAN PEMBAHASAN

Data kategori rumah tangga yang terbagi menjadi kategori rumah tangga miskin dan rumah tangga tidak miskin dengan proporsi seperti pada Gambar 1.



Gambar 1. Persentase Kategori Rumah Tangga.

Kategori rumah tangga tidak miskin yaitu sebanyak 650 rumah tangga (87%), sedangkan rumah tangga miskin sebanyak 95 rumah tangga (13%) dari jumlah keseluruhan amatan rumah tangga

Dataset dibagi menjadi 2 bagian yaitu data latih dan data uji dengan perbandingan 80:20. Dalam penelitian ini penentuan jumlah ketetanggaan menggunakan *tuning hyperparameter RandomsearchCV* dengan rentang *n_neighbors* antara 1 hingga 25 ketetanggaan dan menggunakan 10-foldCV. Berdasarkan *RandomsearchCV* diperoleh nilai k yang paling optimal adalah k=20, sehingga dilakukan klasifikasi rumah tangga menggunakan KNN dengan k=20. Hasil ukuran kinerja klasifikasi disajikan pada Tabel 4.

Permasalahan data *imbalance* ditangani dengan SMOTE dan ADASYN yang diterapkan pada data latih. Penelitian ini menggunakan nilai ketetanggaan (k) sebanyak 5 pada teknik SMOTE dan pada ADASYN digunakan nilai ketetanggaan (k) sebanyak 5, serta parameter ADASYN yaitu d_{th}=0.75 dan β=1. Hasil penanganan data *imbalance* dengan SMOTE dan ADASYN disajikan pada Tabel 3.

Tabel 3 Persentase Data Sebelum dan Sesudah Penanganan Data *imbalance*

Kategori	Banyaknya Data Awal (%)	Banyaknya Data setelah SMOTE (%)	Banyaknya Data setelah ADASYN (%)
1= Rumah Tangga Miskin	73 (12,2%)	523 (50%)	523 (49,7%)
0= Rumah Tangga Tidak Miskin	523 (87,8%)	523 (50%)	529 (50,3%)
Jumlah	596 (100%)	1046 (100%)	1052 (100%)

Proses setelah penanganan data *imbalance* selanjutnya dilakukan klasifikasi menggunakan algoritma KNN dengan *tuning hyperparameter RandomsearchCV* dengan range `n_neighbors` antara 1 hingga 25 dan menggunakan 10-foldCV. Berdasarkan *RandomsearchCV* diperoleh nilai k yang paling optimal adalah k=3 untuk SMOTE-KNN maupun ADASYN-KNN. Performa klasifikasi kategori rumah tangga dengan algoritma KNN setelah penanganan data *imbalance* disajikan pada Tabel 4.

Tabel 4. Perbandingan Ukuran Kinerja Klasifikasi

Metode Klasifikasi	Ukuran Kinerja Klasifikasi			
	Akurasi	Spesifisitas	Sensitivitas	G-*mean*
K Nearest Neighbor (k=20)	85%	100%	0%	0%
SMOTE K Nearest Neighbor (k=3)	77,2%	83,5%	41%	58,5%
ADASYN K Nearest Neighbor (k=3)	74,5%	80,3%	41%	57,3%

Tabel 4 merupakan perbandingan ukuran kinerja klasifikasi dan diperoleh bahwa nilai akurasi dari ketiga metode berkisar di angka 74% hingga 85%. Nilai sensitivitas KNN sebelum dilakukan penanganan data *imbalance* tidak muncul atau bernilai 0, setelah dilakukan penanganan dengan SMOTE dan ADASYN meningkat menjadi 41%. Pada pengukuran spesifisitas klasifikasi rumah tangga sebelum dilakukan penanganan imbalance menghasilkan nilai 100%, setelah penanganan imbalance dengan SMOTE berubah menjadi 83,5%, sedangkan pada penanganan data *imbalance* dengan ADASYN menjadi 80,3%. Pada nilai G-*mean* diperoleh hasil KNN setelah dilakukan penanganan data *imbalance* dengan SMOTE menghasilkan nilai yang paling tinggi diantara ketiga metode yaitu 58,5%.

Perbandingan ketiga metode menghasilkan bahwa, SMOTE-KNN dengan k=3 merupakan metode klasifikasi yang lebih tepat digunakan karena menghasilkan nilai sensitivitas yang baik, nilai spesifisitas yang cukup dibandingkan dengan KNN dan ADASYN-KNN, serta menghasilkan nilai G-*mean* yang paling tinggi yaitu 58,5%.

5. KESIMPULAN

Hasil dari klasifikasi kategori rumah tangga di Kabupaten Temanggung diperoleh kesimpulan bahwa klasifikasi menggunakan *K-Nearest Neighbor* menghasilkan nilai akurasi sebesar 85% dan spesifisitas sebesar 100%, tetapi nilai sensitivitas dan g-*mean* tidak dapat terdefinisi sehingga klasifikasi dianggap tidak berhasil dalam mengklasifikasi kelas minoritas (rumah tangga miskin). Dataset penelitian kategori rumah tangga menunjukkan adanya data *imbalance*, dengan proporsi kelas minoritas (rumah tangga miskin) yaitu sebesar 12,2%. Penanganan data *imbalance* pada klasifikasi kategori rumah tangga dilakukan dengan metode SMOTE dan ADASYN. Klasifikasi KNN setelah dilakukan penanganan data *imbalance*, diperoleh bahwa penerapan SMOTE pada KNN menghasilkan ukuran klasifikasi

yang paling baik, yaitu dapat meningkatkan nilai G-*mean* hingga 58,5%. Penerapan SMOTE mampu meningkatkan kinerja algoritma klasifikasi sehingga dapat mengklasifikasi data dari kelas minor dan mayor meskipun belum menunjukkan hasil yang sangat baik.

DAFTAR PUSTAKA

Badan Pusat Statistik (BPS). 2021. Kemiskinan dan Ketimpangan. (www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html#subjekViewTab1)

Myong Choi, J. 2010. A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines Recommended Citation "A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines', pp. 4–17. Available at: https://core.ac.uk/download/pdf/38924689.pdf.

Han, J. dan Kamber, M. 2006. *Data Mining Concepts and Techniques Second Edition*. San Francisco: Morgan Kaufmann. (http://hanj.cs.illinois.edu/bk2/bib/ch6bib.pdf)

He, H., Bai, Y., Garcia, E. A., & Li, S. .2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE. (https://ieeexplore.ieee.org/abstract/document/4633969)

He, H., E.A. Gracia. 2009. *Learning from Imbalanced Data, IEEE Trans*. Knowl. Discov. 21(9) 1263–1284. (https://ieeexplore.ieee.org/abstract/document/5128907)

He, H and Y. Ma. 2013. *Imbalanced Learning - Foundations, Algorithms, and Applications, 1st ed*. New Jersey: The Institute of Electrical and Electronics Engineers, Inc

Ispriyanti, D., Prahutama, A., & Mustafid, M. 2019. Analisis Klasifikasi Kemiskinan di Kota Semarang Menggunakan Algoritma Quest. *Jurnal Statistika Universitas Muhammadiyah Semarang*, 7(1). (https://jurnal.unimus.ac.id/index.php/statistik/article/view/4805)

Kubat, M., Holte, R. and Matwin, S.1997. Learning when Negative Examples Abound. In European Conference on Machine Learning (pp. 146-153). *Springer*, Berlin, Heidelberg. (https://link.springer.com/chapter/10.1007/3-540-62858-4_79)

Patgiri, C., & Ganguly, A. 2021. Adaptive thresholding technique based classification of red blood cell and sickle cell using Naïve Bayes Classifier and K-Nearest Neighbor classifier. *Biomedical Signal Processing and Control*, 68, 102745. (https://www.sciencedirect.com/science/article/)

Prasetyo, E. 2012. *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI Yogyakarta.