ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 12, Nomor 2, Tahun 2023, Halaman 231 - 241
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

K-NEAREST NEIGHBOR DENGAN *ADAPTIVE BOOSTING* DAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* UNTUK KLASIFIKASI DATA TIDAK SEIMBANG

Ria Sulistyo Yuliani[1*], Agus Rusgiyono[2], Rukun Santoso[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail** : riasulis2107@gmail.com

DOI: 10.14710/j.gauss.12.2.231-241

Article Info:
Received: 2022-11-08
Accepted: 2023-03-15
Available Online: 2023-07-28

Keywords:
Breast Cancer, K-Nearest Neighbor, Imbalanced Data, Synthetic Minority Oversampling Technique, Adaptive Boosting

Abstract: Breast cancer is non-skin cancer that is caused by several factors, including glandular ducts, cells, and breast support tissue, except for the skin of the breast. Breast cancer if not treated immediately will be fatal for the sufferer, so early detection of breast cancer is important for the patient's safety. The success of breast cancer detection depends on the right diagnosis. Measurement of the accuracy of a breast cancer diagnosis can be assisted by statistical methods, namely classification. K-Nearest Neighbor is a classification algorithm based on the nearest neighbor that is easy to implement. In the classification process, there are several problems including when faced with imbalanced data. Imbalanced data can cause classification algorithms to tend to focus on the majority class. Data imbalance can be overcome by using Synthetic Minority Oversampling Technique (SMOTE). Ensemble methods can be applied to improve the performance of imbalanced data classification, one of which is Adaptive Boosting. This study applies K-Nearest Neighbor combined with Adaptive Boosting and SMOTE for handling imbalanced data classification. The results of this study are, SMOTE can handle the problem of imbalanced data and the application of K-Nearest Neighbor with Adaptive Boosting can produce an accuracy of 80%, a sensitivity of 83,33%, a specificity of 66,67%, and a G-Mean value of 74,54%. So it can be concluded that K-Nearest Neighbor combined with Adaptive Boosting and SMOTE can be applied for handling imbalanced data classification.

1. PENDAHULUAN

Kanker payudara merupakan penyakit kanker non kulit berbahaya penyebab kematian nomer dua terbanyak pada wanita yang disebabkan oleh beberapa faktor di antaranya dari saluran kelenjar, sel serta jaringan penopang payudara, kecuali pada kulit payudara (Cahyanti *et. al.*, 2020). Kanker payudara umumnya dibagi menjadi dua kategori yaitu ganas dan jinak (Rejani dan Selvi, 2009). Kanker payudara dengan kategori ganas apabila tidak segera diobati akan berakibat fatal bagi penderitanya, sehingga pendeteksian dini kategori kanker payudara sangat penting bagi keselamatan penderita. Dengan mengetahui kategori kanker payudara sedini mungkin, maka dapat segera dilakukan penanganan yang sesuai dengan tingkat kanker yang diderita (Farahdiba dan Nugroho, 2016). `

Keberhasilan pendeteksian kanker payudara bergantung pada diagnosis yang tepat. Ketepatan diagnosis penyakit kanker payudara dapat dibantu dengan salah satu metode statistika yaitu klasifikasi. *K-Nearest Neighbor* merupakan metode klasifikasi berdasarkan ketetanggaan terdekat antara data satu dengan data lainnya yang mudah diterapkan (Nikhitha dan Jabbar, 2019). Terdapat beberapa permasalahan yang dijumpai dalam klasifikasi, salah satunya yaitu ketidakseimbangan data. Data tak seimbang terjadi ketika salah satu kelas memiliki jumlah jauh lebih besar dibandingkan dengan kelas lainnya sehingga mampu menyebabkan menurunnya performa klasifikasi pada kelas minoritas (Fitriani *et al.*, 2021).

Permasalahan data tidak seimbang dapat ditangani dengan menerapkan *oversampling*. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan salah satu metode *oversampling* yang paling efektif (Pangastuti, 2018).

Alternatif lain dalam meningkatkan performa klasifikasi kelas data tidak seimbang yaitu dengan menerapkan metode *ensemble*. *Adaptive Boosting* merupakan salah satu metode *ensemble* dengan dasar teori yang kuat, prediksi akurat, dan sederhana (Rais dan Subekti, 2019). Penelitian ini akan menerapkan algoritma *K-Nearest Neighbor* yang dikombinasikan dengan *Adaptive Boosting* serta terdapat penanganan data tidak seimbang dengan *Synthetic Minority Oversampling Technique* untuk klasifikasi data pasien biopsi insisi kanker payudara RSUD Kabupaten Nganjuk tahun 2019-2020.

2. TINJAUAN PUSTAKA

Kanker payudara merupakan penyakit kanker non kulit yang disebabkan oleh beberapa faktor di antaranya dari saluran kelenjar, sel serta jaringan penopang payudara, kecuali pada kulit payudara (Cahyanti *et. al.*, 2020). Kanker payudara umumnya dikategorikan menjadi dua, yaitu *benign* (jinak) dan *malignant* (ganas) (Rejani dan Selvi, 2009). Dengan mengetahui kategori kanker payudara sedini mungkin, maka dapat segera dilakukan penanganan yang sesuai dengan tingkat kanker yang diderita (Farahdiba dan Nugroho, 2016). Pendeteksian kanker payudara dapat dilakukan dengan beberapa metode di antaranya dengan biopsi. Biopsi merupakan salah satu metode operasi yang layak digunakan dalam pendeteksian kanker (Versaggi, S. dan Leucio, A., 2022). Adapun jenis biopsi yang sering digunakan yaitu biopsi insisi. Biopsi insisi merupakan metode untuk mendeteksi kanker dengan cara mengambil sebagian kecil dari jaringan yang dicurigai sebagai kanker, hasil biopsi insisi ini memiliki prediksi yang baik (Bromberg *et.al.*, 2018). Pendeteksian kategori kanker payudara bergantung pada ketepatan diagnosisnya. Ketepatan sebuah diagnosis atau prediksi dapat dibantu dengan salah satu metode statistika yaitu klasifikasi.

Klasifikasi merupakan teknik pada data *mining* yang memprediksikan kelas dalam suatu objek yang labelnya belum diketahui tetapi label pada data yang digunakan telah diketahui (Han *et al.*, 2012). Pada proses klasifikasi memerlukan data latih yang digunakan untuk membentuk model dalam mengelompokkan data uji (Prianti *et al.*, 2020). Salah satu cara untuk membagi data menjadi data latih dan data uji yaitu dengan *holdout validation*. *Holdout validation* memiliki cara kerja yaitu dengan membagi data kedalam dua bagian dengan proporsi tertentu, adapun proporsi yang sering digunakan oleh peneliti yaitu 60/40, 70/30, 80/20 (Raschka, 2018).

Proses *pre-processing* dilakukan sebelum data diolah. Salah satu proses dari *pre-processing* yaitu normalisasi data. Normalisasi digunakan untuk menyamakan rentang nilai pada setiap atribut dengan menggunakan skala tertentu (Nasution *et al*., 2019). Salah satu metode normalisasi yang dapat digunakan yaitu *Min-Max Normalization*. Metode *Min-Max Normalization* mampu mengubah data yang kompleks dengan tidak menghilangkan isinya, sehingga lebih mudah untuk proses pengolahannya (Wimmer, 2018). Perhitungan *Min-Max Normalization* dapat dilihat pada Persamaan 1.

$$X_{baru} = \frac{X_{lama} - X_{min}}{X_{max} - X_{min}} \quad (1)$$

dengan X_{baru} merupakan data atribut yang dinormalisasi, X_{lama} merupakan data atribut asli, X_{min} merupakan nilai terkecil atribut, dan X_{max} merupakan nilai terbesar atribut.

K-Nearest Neighbor (K-NN) merupakan metode klasifikasi yang didasarkan pada jarak antara data uji dengan data latih yang dapat dihitung salah satunya dengan jarak *euclidean* yang didefinisikan dalam Persamaan 1 (Han dan Kamber, 2006). K-NN merupakan metode yang cukup sederhana yaitu tidak memerlukan asumsi pada distribusi data dan mudah untuk diimplementasikan (Santosa, 2007).

$$d(\boldsymbol{x_i}, \boldsymbol{y_j}) = \sqrt{\sum_{l=1}^{L} dif\ (f_{(x_{il}, y_{jl})})^2} \qquad (2)$$

dengan x_{il} merupakan data uji ke-i pada variabel ke-*l*, y_{jl} merupakan data latih ke-j pada variabel ke-*l*, *L* merupakan banyaknya variabel bebas, *n* merupakan banyaknya data latih, *i* dan *j* berjalan dari 1, 2, … ,*n*, dan $dif\ f_{(x_{il}, x_{jl})}$ merupakan *difference* atau ketidaksamaan antara x_{il} dan y_{jl}. Perhitungan nilai ketidaksamaan berdasarkan pada tipe data untuk tiap variabel disajikan pada Tabel 1 (Prasetyo, 2014).

Tabel 1 Ketidaksamaan Dua Data dengan Satu Atribut

Tipe Atribut	Formula Jarak
Nominal	$dif\ (f_{(x_{il}, y_{jl})}) = \begin{cases} 0 \text{ apabila } x_{il} = y_{jl} \\ 1 \text{ apabila } x_{il} \neq y_{jl} \end{cases}$
Ordinal	$dif\ (f_{(x_{il}, y_{jl})}) = \frac{\lvert x_{il} - y_{jl} \rvert}{(q-1)}$ dengan *q* merupakan banyaknya pengkategorian dalam variabel bebas
Interval atau Rasio	$dif\ (f_{(x_{il}, y_{jl})}) = \lvert x_{il} - y_{jl} \rvert$

Imbalanced class data merupakan suatu kondisi data tak seimbang antara kelas satu dan kelas lainnya. *Imbalanced class* data terjadi jika jumlah data dalam satu kelas lebih tinggi (mayoritas) atau lebih rendah (minoritas) jika dibandingkan dengan kelas yang lain. Kondisi data tidak seimbang merupakan salah satu kasus dalam klasifikasi yang mampu mengakibatkan algoritma pembelajaran klasifikasi cenderung fokus memprediksi pada kelas mayoritas dibandingkan dengan kelas minoritas (Fitriani, 2021). Suatu data dikatakan tidak seimbang apabila kelas sampel yang lebih sedikit, proporsinya kurang dari 35% (Thammasiri *et.al*., 2014).

Synthetic minority over-sampling technique atau yang biasa dikenal dengan SMOTE merupakan salah satu metode untuk penanganan data tidak seimbang. SMOTE bekerja dengan cara menambah jumlah sampel data minoritas agar seimbang dengan sampel data mayoritas dengan cara pembangkitan data sintetis yang didasarkan pada tetangga terdekat yang dipilih berdasarkan jarak *euclidean* antara kedua data (Chawla *et.al*, 2002). Adapun rumus jarak *euclidean* terdapat pada Persamaan 2. Pembangkitan data sintetis dilakukan dengan Persamaan 3 (Choi, 2010).

$$x_{synj} = x_i + (x_{knn} - x_i) \times \gamma \qquad (3)$$

dengan x_{synj} merupakan data sintetis, x_i merupakan data ke-*i* dari kelas minoritas, x_{knn} merupakan data dari kelas minoritas yang memiliki jarak terdekat dari x_i, γ merupakan bilangan random antara 0 dan 1 (dengan γ dipilih secara acak oleh program).

Metode *ensemble* merupakan sebuah metode yang digunakan untuk menggabungkan beberapa algoritma klasifikasi guna menciptakan model baru yang memiliki kinerja lebih baik. Beberapa studi eksperimental dengan *machine learning* menunjukkan bahwa kombinasi output dari sejumlah algoritma pengklasifikasian dapat mengurangi kesalahan generalisasi (Quinlan, 1996). Salah satu metode *ensemble* yang dapat digunakan yaitu *stacking*. *Stacking* merupakan salah satu metode *ensemble* dengan cara kerja menumpuk dua atau lebih algoritma klasifikasi tunggal (Nurmasani dan Pristyanto, 2021). *Stacking* dapat memanfaatkan kemampuan dari beberapa model tunggal dan membuat prediksi menjadi lebih baik daripada model tunggal. Pada penerapan *stacking* terdapat dua level pembelajaran, di antaranya yaitu level pertama merupakan model pembelajaran level-0 atau biasa disebut dengan *base learner*, dan untuk level-1 merupakan model *meta learner* (Nugraha dan Rahman, 2019).

Adaptive boosting atau yang biasa dikenal dengan Adaboost merupakan salah satu metode *ensemble* jenis *boosting*. Cara kerja metode *boosting* ini yaitu menghasilkan prediksi yang akurat dengan mengombinasikan pengklasifikasi lemah. Adaboost menjadi metode *ensemble* yang sering digunakan serta mampu diterapkan di berbagai bidang karena memiliki dasar teori yang kuat dan prediksinya akurat (Rais dan Subekti, 2019). Penjabaran algoritma Adaboost (Zhu, *et.al*, 2009) yaitu:

1. Menginisialisasi bobot awal amatan $w_i = 1/n$ (4)
   dengan n merupakan banyaknya amatan data latih dan i = 1,2,3,…,n.
2. Iterasi (m) = 1,2,3,…,M
   a. Menetapkan fungsi klasifikasi $G_m(x)$ pada data latih. $G_m(x)$ memberikan hasil prediksi pada saat proses pelatihan menggunakan *stump* (pohon kecil) yang didasarkan pada *information gain* dan entropi. *Information gain* merupakan salah satu metode seleksi fitur yang sering digunakan olek peneliti untuk penentuan batas kepentingan dari suatu atribut (Novakovic, 2010). Pengukuran atribut dengan entropi dirumuskan dengan Persamaan 5 (Gallager dan Fellow, 2001):
   $$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (5)$$
   dengan D merupakan himpunan kasus, m merupakan banyaknya partisi D, dan p_i merupakan proporsi dari D_i terhadap D. Perhitungan entropi setelah pemisahan dilakukan dengan Persamaan 6.
   $$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} \times I(Dj) \quad (6)$$
   dengan A merupakan atribut, v merupakan banyaknya partisi atribut A, $|D_j|$ merupakan jumlah kasus partisi ke-j, dan $I\ (Dj)$ merupakan total entropi dalam partisi. Mencari *information gain* pada atribut A dengan Persamaan 7.
   $$Gain\ (A) = I(D) - I(A) \quad (7)$$
   dengan Gain (A) merupakan *information gain* atribut A, I(D) merupakan total entropi, dan I(A) merupakan entropi A. Data latih diprediksi guna mendeteksi klasifikasi yang salah diklasifikasikan dan data uji digunakan untuk mengambil suara terbanyak dalam menentukan kelas.
   b. Menghitung *error rate* pada klasifikasi yang salah diklasifikasikan dengan Persamaan 8.
   $$e_m = \frac{\sum_{i=1}^{n} w_i{}^{(m)}\ l(G_m\ (x_i) \neq y_i)}{\sum_{i=1}^{n} w_i\ {}^{(m)}} \quad (8)$$
   dengan e_m merupakan nilai error iterasi ke-m, $w_i\ {}^{(m)}$ merupakan bobot awal amatan iterasi ke-m, $G_m\ (x_i)$ merupakan nilai prediksi data latih, y_i merupakan kelas asli, dan $l(G_m\ (x_i) \neq y_i)$ merupakan indikator yang akan bernilai 1 jika kelas prediksi tidak

sama dengan kelas aslinya dan akan bernilai 0 untuk lainnya. Apabila nilai $e_m > 1 - \frac{1}{c}$ dengan c merupakan banyaknya kelas pada data, maka iterasi akan dihentikan. Sebaliknya, apabila $e_m \leq 1 - \frac{1}{c}$, maka proses dilanjutkan dengan menghitung bobot suara klasifikasi.

c. Menghitung bobot suara klasifikasi dengan Persamaan 9.

$$\alpha_m = ln\left(\frac{1-e_m}{e_m}\right) \quad (9)$$

d. Memperbarui nilai bobot dengan Persamaan 10.

$$w_i^{(m+1)} = \frac{w_i{}^{(m)}}{z_m} exp\left(\alpha_m l(G_m(x_i) \neq y_i)\right) \quad (10)$$

dengan $Z_m = \sum_{i=1}^{n} w_i{}^{(m)}$, Z_m merupakan jumlah dari seluruh bobot pada iterasi ke-m.

e. Menentukan prediksi kelas menggunakan fungsi klasifikasi akhir dengan persamaan 11.

$$T(x) = arg\ \max_c \sum_{m=1}^{M} \alpha_m I(G_m(x_i) = c) \quad (11)$$

dengan $T(x)$ merupakan prediksi kelas, *c* merupakan kelas prediksi data uji yaitu antara 0 dan 1, dan $l(G_m(x_i) = c)$ merupakan fungsi indikator yang akan bernilai 1 jika kelas prediksi data uji sama dengan kelas k dan bernilai 0 untuk lainnya.

k-fold cross validation merupakan salah satu metode statistik yang digunakan untuk mengevaluasi serta membandingkan algoritma pembelajaran dengan cara membagi data menjadi data latih dan data uji (Tan *et.al.*, 2006). Pada proses *k-fold cross validation*, data dibatasi secara acak ke dalam *k* bagian yang bersifat saling terbatas $D_1, D_2, \ldots, D_k$, dengan masing-masing memiliki ukuran hampir sama. Proses pelatihan dan pengujian dilakukan sebanyak *k* kali. Umumnya, 10-*fold cross validation* direkomendasikan karena bias dan variansinya relatif rendah (Han *et.al.*, 2006).

Evaluasi kinerja (*performance*) klasifikasi dapat dilakukan dengan beberapa cara, di antaranya dengan menggunakan tabulasi silang (*confussion matrix*). *Confussion matrix* menyajikan data asli dan data hasil prediksi dari sebuah model klasifikasi yang mengandung informasi kelas data asli direpresentasikan dalam baris matriks dan kelas data hasil prediksi dalam kolom (Han *et.al.*, 2006). Bentuk dari *confussion matrix* dapat dilihat pada Tabel 2.

Tabel 2 *Confussion Matrix* pada Klasifikasi Dua Kelas

	Predictive Positive Class	*Predictive Negative Class*
Real Positive Class	TP (*True Positive*)	FN (*False Negative*)
Real Negative Class	FP (*False Positive*)	TN (*True Negative*)

Akurasi merupakan kriteria umum yang sering digunakan untuk mengukur kinerja sebuah klasifikasi, tetapi dalam kasus data tidak seimbang kriteria akurasi kurang tepat karena kelas minoritas akan memiliki peranan kecil (Khasanah *et.al.*, 2019). Rumus akurasi dapat dilihat pada Persamaan 12.

$$Akurasi = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100\% \quad (12)$$

Pengukuran kinerja klasifikasi biner dilakukan dengan menghitung nilai sensitivitas dan spesifitas. Sensitivitas digunakan dalam pengukuran proporsi *true positive* yang secara tepat diprediksi sebagai positif pada seluruh titik data yang positif, sedangkan spesifitas digunakan

dalam pengukuran proporsi *true negative* yang secara tepat diprediksi sebagai negatif pada seluruh titik data yang negatif (Gorunescu, 2011).

$$Sensitivitas = \frac{TP}{(TP+FN)} \times 100\% \quad (13)$$

$$Spesifitas = \frac{TN}{(TN+FP)} \times 100\% \quad (14)$$

Geometric Mean (G-*Mean*) dapat digunakan untuk mengevaluasi kinerja sebuah metode secara keseluruhan. Suatu algoritma apabila semua kelas positif tidak dapat diprediksi maka nilai G-*Mean* akan bernilai nol (Kubat dan Matwin, 1997).

$$G - Mean = \sqrt{Sensitivitas \times Spesifitas} \quad (15)$$

3. METODE PENELITIAN

Jenis data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Rumah Sakit Umum Daerah Kabupaten Nganjuk. Data tersebut merupakan data pasien biopsi insisi kanker payudara tahun 2019-2020. Data tersebut terdiri dari 75 pasien yang terbagi ke dalam 62 pasien dengan kanker payudara kategori ganas (82,67%) dan 13 pasien dengan kanker payudara kategori jinak (17,33%). Variabel yang digunakan terdiri dari 1 variabel respon berupa kategori pasien kanker payudara (jinak dan ganas) dan 6 variabel bebas yang terdiri dari keseragaman bentuk dan ukuran inti sel, keberadaan anak inti sel, pertumbuhan tumor, mitosis, metastase, serta usia. Variabel penelitian dapat dilihat pada Tabel 3.

Tabel 3 Variabel Penelitian

Variabel	Deskripsi	Keterangan
Y	Kategori kanker payudara	0=jinak 1=ganas
$\boldsymbol{X_1}$	Usia	Tahun
$\boldsymbol{X_2}$	Bentuk dan ukuran inti sel	1=inti bulat monoton 2=inti bulat pleomorfik ringan 3=inti bulat pleomorfik sedang 4=inti bulat pleomorfik berat
$\boldsymbol{X_3}$	Keberadaan anak inti sel	0= tidak tampak anak inti 1= tampak anak inti
$\boldsymbol{X_4}$	Pertumbuhan tumor	0=potongan jaringan tumor jinak 1=potongan jaringan tumor ganas
$\boldsymbol{X_5}$	Mitosis	0=tidak ditemukan mitosis 1=ditemukan mitosis

Variabel	Deskripsi	Keterangan
$\boldsymbol{X_6}$	Metastase	0=tidak tampak metastase 1=tampak metastase

Pengolahan data dilakukan menggunakan bahasa R. Data yang telah diperoleh kemudian dilakukan analisis sebagai berikut:

1. Input data dengan *software* R.
2. Melakukan *pre-processing* data (normalisasi data) menggunakan metode *Min-Max Normalization*.
3. Membagi data menjadi data latih dan data uji dengan perbandingan 80:20.
4. Melakukan klasifikasi menggunakan algoritma *K-Nearest Neighbor* (K-NN) dengan metode *ensemble Adaptive Boosting* dan penanganan kelas data tidak seimbang menggunakan *Synthetic Minority Oversampling Technique* (SMOTE).
   a. Melakukan penanganan *imbalance* data dengan SMOTE pada data latih.
   b. Melakukan klasifikasi dengan algoritma KNN setelah dilakukan penanganan data *imbalance* dengan SMOTE.
   c. Melakukan klasifikasi dengan algoritma Adaboost.
   d. Memberikan bobot amatan sampel pada seluruh data latih $w_i = 1/n$
   e. Mencari *stump* dengan memilih prediktor yang memiliki nilai *information gain* yang paling tinggi.
   f. Menghitung kesalahan pelatihan (*error rate*). Apabila nilai $e_m > 1 - \frac{1}{c}$ dengan c merupakan banyaknya kelas kategori kanker payudara yaitu 2, $e_m > 0{,}5$ maka iterasi dihentikan. Jika nilai nilai $e_m \leq 1 - \frac{1}{c}$ dengan c merupakan banyaknya kelas kategori kanker payudara yaitu 2, $e_m \leq 0{,}5$ maka menghitung pembobot suara klafisikasi.
   g. Menetapkan bobot suara untuk komponen latih.
   h. Memperbarui komponen bobot latih. Hasil klasifikasi yang salah, bobotnya akan dinaikkan sehingga prediksi selanjutnya diharapkan tidak salah lagi.
   i. Melakukan normalisasi bobot agar jumlah total bobot sama dengan 1.
      Normalisasi bobot = bobot baru/total bobot
   j. Menentukan prediksi kelas menggunakan fungsi klasifikasi akhir.
   k. Melakukan *stacking* pada model K-NN dan Adaboost yang telah diseimbangkan oleh SMOTE.
   l. Mengklasifikasikan data uji menggunakan model K-NN yang telah dikombinasikan dengan metode *ensemble* Adaboost dan penanganan kelas data tidak seimbang menggunakan SMOTE.
5. Melakukan perhitungan performa klasifikasi model *K-Nearest Neighbor* (K-NN) yang telah dikombinasikan dengan metode *ensemble Adaptive Boosting* dan penanganan kelas data tidak seimbang menggunakan *Synthetic Minority Oversampling Technique* (SMOTE).

4. HASIL DAN PEMBAHASAN

Pre-processing dilakukan sebelum proses pengolahan data. *Pre-processing* dilakukan dengan normalisasi data menggunakan metode *Min-Max normalization* untuk mengatasi variabel yang dominan dalam proses pengolahan data. Perhitungan *Min-Max Normalization* menggunakan Persamaan 1. Setelah data dinormalisasi, kemudian dilanjutkan dengan proses pembagian data latih dan data uji sebelum dilakukan pengolahan data. Penelitian ini menggunakan pembagian 80% untuk data latih dan 20% untuk data uji. Pembagian data menggunakan *software* R menghasilkan data latih sejumlah 60 data dan data uji sejumlah 15 data. Data latih terdiri dari 10 data dengan kelas kategori kanker payudara jinak dan 50 data dengan kelas kategori kanker payudara ganas.

Penanganan kelas data tidak seimbang dilakukan dengan menerapkan *Synthetic Minority Oversampling* (SMOTE). SMOTE bekerja dengan mencari ketetanggan terdekat dari data sebanyak K untuk setiap data kelas minoritas. Penelitian ini menggunakan nilai ketetanggaan (K) sejumlah 5 seperti pada *default* program SMOTE pada *software* R yang didasarkan pada SMOTE *Algorithm for Unbalanced Classification Problems in Performance Estimation : An Infra-Structure for Performance Estimation of Predictive Models* (https://rdrr.io/cran/performanceEstimation/man/smote.html). Hasil SMOTE pada data latih dapat dilihat pada Tabel 4.

Tabel 4 Persentase Data Sebelum dan Sesudah SMOTE

Kelas (Kategori)	Banyaknya Data Awal (%)	Banyaknya Data Setelah SMOTE (%)
0=jinak	10 (18,03%)	20 (40%)
1=ganas	50 (81,97%)	30 (60%)
Jumlah	60 (100%)	50 (100%)

Nilai *perc. over* yang digunakan untuk *oversampling* data minoritas sebesar 100% dari kelas mayoritas, yang berarti bahwa kelas minoritas akan ditambahkan sejumlah 10 data. Nilai *perc. under* yang digunakan untuk *undersampling* data mayoritas yang digunakan sebesar 300% dari *oversampling* kelas minoritas, yang berarti bahwa kelas mayoritas akan dikurangi jumlahnya menjadi 30 data. Frekuensi data sebelum dan sesudah penerapan SMOTE pada kelas kanker payudara kategori jinak sebelumnya berjumlah 10 data bertambah menjadi 20 data, serta pada kelas kanker payudara kategori ganas sebelumnya berjumlah 50 data berkurang menjadi 30 data. Penerapan SMOTE mampu menghasilkan data yang hampir seimbang kelasnya dengan persentase 40% data dengan kelas kanker payudara kategori jinak dan 60% data kelas kanker payudara kategori ganas dari data latih.

Klasifikasi menggunakan *K-Nearest Neighbor* didasarkan pada jarak antara data uji dan data latih yang dihitung menggunakan jarak *euclidean* dengan Persamaan 2. Pada penelitian ini penentuan jumlah ketetanggaan (K) menggunakan tuning K optimal dari *packages* yang disediakan oleh *software* R untuk mempermudah proses penentuan K pada K-NN dan menggunakan 10-*fold cross validation* pada proses pelatihan. Nilai ketetanggaan yang paling optimal dilihat dari nilai akurasi yaitu sebesar 99,6% dengan nilai K=9. Sehingga digunakan nilai K=9 untuk model klasifikasi menggunakan algoritma K-NN setelah dilakukan penanganan kelas data tidak seimbang dengan SMOTE.

Klasifikasi menggunakan *Adaptive Boosting* (Adaboost) diawali dengan menempatkan bobot awal $w_1 = 1/n = 1/60 = 0{,}0167$, dengan n merupakan banyaknya data latih. Pada penelitian ini menggunakan *stump* dengan nilai *information gain* tertinggi dan batas iterasi maksimum untuk mencari nilai prediksi pada data latih. Pemilahan *stump* dilakukan dengan mencari nilai entropi dan *information gain* pada variabel bebas. Nilai

information gain tertinggi pada atribut akan digunakan sebagai pemilah *stump*. Prediksi yang telah diperoleh menggunakan *stump*, dilanjutkan dengan mencari nilai *error* (e_1) menggunakan Persamaan 8. Apabila nilai $e_m \leq 1 - \frac{1}{c}$ dengan *c* merupakan banyaknya kelas kategori kanker payudara yaitu 2, ($e_1 \leq 0{,}5$) maka perhitungan dilanjutkan dengan mencari bobot suara klasifikasi (α_1) dengan Persamaan 9. Nilai α_1 telah didapatkan, kemudian dilanjutkan dengan memperbarui bobot dengan Persamaan 10. Perhitungan berulang (iterasi) dan akan berhenti apabila $e_m > 1 - \frac{1}{c}$ dengan *c* merupakan banyaknya kelas kategori kanker payudara yaitu 2, ($e_m > 0{,}5$) atau perhitungan sampai dengan iterasi maksimum. Prediksi kelas dilihat dari jumlah α tertinggi pada masing-masing kelas. Pada proses Adaboost ini, model dibangun dengan data latih yang telah diseimbangkan dengan SMOTE. Penelitian ini menggunakan *packages* yang disediakan oleh *software* R dalam penentuan jumlah iterasi (M) optimal dan menggunakan 10-*fold cross validation* pada proses pelatihan. Hasil yang didapatkan yaitu nilai M=150 dengan akurasi tertinggi sebesar 93,2%. Sehingga jumlah iterasi maksimal yang digunakan untuk model klasifikasi menggunakan algoritma Adaboost yaitu 150 iterasi.

Stacking digunakan untuk mengombinasikan dua atau lebih model dengan algoritma yang berbeda guna mendapatkan model prediksi dengan akurasi yang tinggi. Tahap pertama metode *stacking* yaitu setiap pengklasifikasi dasar yang digunakan dilatih dengan menggunakan data yang sama sehingga memperoleh hasil prediksi masing-masing. Pada penelitian ini menggunakan pengklasifikasi dasar pada data latih dengan algoritma K-NN dan Adaboost dengan data yang sudah diseimbangkan dengan SMOTE. Tahap kedua metode *stacking* yaitu membangun model *meta classifier* yang digunakan untuk mengambil hasil prediksi dari masing-masing pengklasifikasi dasar kemudian digunakan sebagai input untuk penentuan kelas yang paling mungkin terhadap data uji coba. Pada penelitian ini menggunakan adaboost sebagai model *meta classifier*. Hasil model *meta classifier* yang telah dibangun pada data latih yang sudah di SMOTE menunjukkan bahwa nilai M (jumlah iterasi) optimal yang digunakan bernilai 100 dengan akurasinya sebesar 96,64%.

Hasil metode *stacking* K-NN dengan Adaboost dan SMOTE yang telah dibangun akan dievaluasi kinerja model klasifikasinya menggunakan *confussion matrix* disajikan pada Tabel 5.

Tabel 5 *Confussion Matrix* Klasifikasi *K-Nearest Neighbor* Kombinasi dengan Adaboost dan SMOTE

Kelas Asli	Kelas Prediksi		Total
	Jinak	Ganas	
Jinak	2	2	4
Ganas	1	10	11
Total	3	12	15

Ukuran kinerja klasifikasi dapat dihitung berdasarkan Tabbel 5 dengan perhitungan berikut:

$$akurasi = \frac{(TP + TN)}{Total} \times 100\%$$

$$akurasi = \frac{10+2}{15} \times 100\% = 80\%$$

$$sensitivitas = \frac{TP}{(TP + FN)} \times 100\%$$

$$sensitivitas = \frac{10}{(10+2)} \times 100\% \ = 83{,}33\%$$

$$spesifitas = \frac{TN}{(TN + FP)} \times 100\%$$

$$spesifitas = \frac{2}{(2+1)} \times 100\% = 66{,}67\%$$

$$G - mean(rata - rata\ geometri) = \sqrt{sensitivitas \ \times spesifitas}$$

$$G - mean(rata - rata\ geometri) = \sqrt{0{,}8333 \times 0{,}6667} = 74{,}54\%$$

Ukuran kinerja klasifikasi pada metode *K-Nearest Neighbor* yang dikombinasikan dengan *Adaptive Boosting* dengan penanganan kelas data tidak seimbang menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) diperoleh nilai akurasi sebesar 80% sehingga dapat dikatakan bahwa ketepatan algoritma dalam melakukan klasifikasi kategori kanker payudara adalah 80%. Nilai sensitivitas sebesar 83,33% menunjukkan bahwa 83,33% kanker payudara kategori ganas diprediksikan secara benar untuk kategori kanker payudara ganas. Sedangkan untuk nilai spesifitas sebesar 66,67% menunjukkan bahwa terdapat 66,67% kanker payudara kategori jinak yang diprediksikan secara benar untuk kategori kanker payudara jinak. Nilai *G-Mean* (rata-rata geometri) diperoleh sebesar 74,54% yang dapat disimpulkan bahwa sebesar 74,54% keseluruhan algoritma dapat melakukan klasifikasi secara tepat.

5. KESIMPULAN

Penerapan klasifikasi kategori kanker payudara menggunakan *K-Nearest Neighbor* yang dikombinasikan dengan *Adaptive Boosting* dan penanganan kelas data tidak seimbang menggunakan *Synthetic Minority Oversampling Technique* dapat diterapkan. Kelas data tidak seimbang dalam klasifikasi kategori kanker payudara ditangani dengan metode *Synthetic Minority Oversampling Technique* (SMOTE). Penanganan data tidak seimbang menggunakan SMOTE menghasilkan data yang hampir seimbang antara kelas kanker payudara kategori ganas dan kelas kanker payudara kategori jinak dengan persentase 60:40.

Metode klasifikasi *K-Nearest Neighbor* yang dikombinasikan dengan *Adaptive Boosting* dan penanganan kelas data tidak seimbang menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) menghasilkan nilai akurasi sebesar 80%, nilai sensitivitas sebesar 83,33%, nilai spesifitas sebesar 66,67%, dan nilai *G-Mean* (rata-rata geometri) sebesar 74,54%. Algoritma klasifikasi mencapai nilai *G-Mean* (rata-rata geometri) sebesar 74,54%, maka dapat disimpulkan bahwa algoritma klasifikasi secara keseluruhan mampu mengklasifikasikan semua amatan secara tepat sebesar 74,54%.

DAFTAR PUSTAKA

Bromberg, S.E., Moraes, P.R.A.d.F, dan Ades, F., 2018. *Prime incision: A minimally invasive approach to breast cancer surgical treatment-A 2 cohort retrospective comparison with conventional breast conserving surgery*. Tersedia: https://doi.org/10.1371/journal.pone.0191056 (diakses pada tanggal 21 juni 2022).

Cahyanti, D., Rahmayani, A., dan Husnair, S. 2020. Analisis Performa Metode KNN pada Dataset Pasien Pengidap Kanker Payudara. *Indonesian Journal of Data Science* Vol.1, No. 2, Hal: 39-43.

Chawla, N., Bowyer, K., Hall, L., dan Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* Vol. 16, No. 1, Hal: 321–357.

Choi, J. M. 2010. *A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines*. Graduate Theses and Dissertations, Paper 11529.

Farahdiba, B., dan Nugroho, Y. 2016. Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio. *Jurnal 1 Teknik Elektro* Vol. 8, No. 2.

Fitriani, R. D., Yasin, H., dan Tarno, T. 2021. Penanganan Klasifikasi Kelas Data Tidak Seimbang dengan Random Oversampling Pada Naive Bayes (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal). *Jurnal Gaussian* Vol. 10, No. 2, Hal:11-20.

Gorunescu, F. 2011. *Data mining: Concepts, Model, and Technique*. Jerman: Springer.

Han, J., Kamber, M., dan Pei, J. 2006. *Data mining: Concept and Techniques.* Waltham: Morgan Kaufmann Publisher.

Han, J., Kamber, dan M., Pei, J. 2012. *Data mining Concepts and Techniques 3rd Edition*. Kaufman Publisher, USA.

Khasanah, A., Muladi, Pujianto, U. 2019. Penerapan Teknik SMOTE untuk Mengatasi *Imbalance Class* dalam Klasifikasi Objektivitas Berita *Online* Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi* Vol.3, No. 2, Hal: 196-201.

Nikhitha, M. dan Jabbar, M.A. 2019. K Nearest Neighbor Based Model For Intrusion Detection System. *International Journal of Recent Technology and Engineering (IJRTE)* Vol. 8, No. 2, Hal: 2277-3878.

Novakovic, Jasmina. 2010. The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier Vol 2, Hal: 1113–16.

Nugraha, A. F., dan Rahman, L. 2019. Meta-algorithms for improving classification performance in the web-phishing detection process. *4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE* Vol. 6, Hal: 271–275.

Nurmasani, A., dan Pristyanto, Y. 2021. Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class. *Pseudocode* Vol. 8, No. 1, Hal: 21–26.

Prianti, A., Santoso, R., Hakim, A. 2020. Perbandingan Metode K-Nearest Neighbor dan Adaptive Boosting pada Kasus Klasifikasi Multi Kelas. *Jurnal Gaussian* Vol. 9, No. 3, Hal: 346-354.

Rais, N., Subekti, A. 2019. Integrasi SMOTE dan Ensemble Adaboost untuk Mengatasi Imbalance Class pada Data Bank Direct Marketing. *Jurnal Informatika* Vol. 6, No. 3, Hal: 278-285.

Raschka, S. 2018. *Model evaluation, model selection, and algorithm selection in machine learning*. arXiv preprint arXiv:1811.12808.

Rejani, Y. dan Selvi, S. 2009. Early Detection of Breast Cancer Using SVM Classifier Technique Vol. 1, No. 3, Hal: 127–130.

Versaggi, S.L., dan Leucio, A. d. 2022. *Breast Biopsy National Library of Medicine in StatPearls Publishing.* Tersedia: https://www.ncbi.nlm.nih.gov/books/NBK559147/ (diakses pada tanggal 21 Juni 2022).

Wimmer, H. 2018. Effect of Normalization Techniques on Logistic Regression in Data Science. *Proceeding of the Conference on Information Systems Applied Research* Hal: 1-9.