e-ISSN: 2964-2531
DOI : 10.36499/psnst.v13i1.9005
p-ISSN: 2964-5131

KLASIFIKASI SENYAWA KELADI TIKUS MENGGUNAKAN ALGORITMA KNN, GAUSSIAN NAÏVE BAYES DENGAN MENERAPKAN IMBALANCE DATA BORDERLINE SMOTE

Frista Dea[1], Margareta Navi Primadani[2], Theresia Winnie Kartikasari[3], Iwan Binanto[4], Nesti F. Sianipar[5]
[1]Informatika, Universitas Sanata Dharma, Yogyakarta
[2]Biotechnology Department, Faculty of Engineering, Bina Nusantara University, Jakarta
[*]Email: iwan@usd.ac.id

Abstrak

Data seimbang atau imbalanced data merupakan keadaan di mana distribusi kelas data yang tidak seimbang yaitu jumlah data yang satu lebih sedikit atau lebih banyak dari kelas lainnya. Menangani data yang tidak seimbang telah menjadi tantangan besar selama dua dekade terakhir. Keseimbangan data merupakan faktor yang penting untuk diperhatikan, karena mempengaruhi hasil yang diperoleh. Penelitian ini bertujuan untuk melakukan perbandingan metode antara KNN, Gaussian Naïve Bayes, dan Random Forest untuk menentukan metode yang paling baik berdasarkan data tanaman keladi tikus. Data yang tidak seimbang akan diseimbangkan dengan menggunakan metode oversampling yaitu Borderline-SMOTE. Dari penelitian yang telah dilakukan, algoritma KNN, Gaussian Naïve Bayes, dan Random Forest pada data yang sebenarnya (belum seimbang) menghasilkan nilai akurasi berturut-turut sebesar 0.984, 0.985, dan 1. Sedangkan pada data yang sudah diseimbangkan menghasilkan akurasi berturut-turut adalah sebesar 0.967, 0.499, dan 0.984. Algoritma random forest dapat mengklasifikasikan data yang seimbang dan belum seimbang dengan baik dibandingkan dengan algoritma yang lain. Hal ini karena algoritma random forest menghasilkan score akurasi, recall, F1-score, dan Precission yang tinggi dibanding dengan algoritma KNN dan Gaussian Naive Bayes pada data yang unbalance maupun balance.

***Kata kunci**: LCMS, imbalance data, oversampling, Borderline-SMOTE*

1. PENDAHULUAN

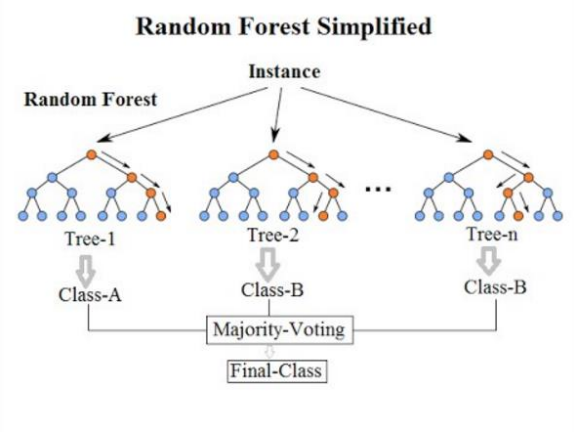
Ketersediaan data dan informasi semakin bertambah dari segi jumlah maupun komplektsitasnya karena perkembangan teknologi digital. Oleh karena semakin bertambahnya jumlah dan kompleks maka akan memunculkan masalah dalam mengklasifikasian data. Permasalahan tersebut ialah munculnya data tidak seimbang (Rahmatunissa, 2021). Data seimbang atau imbalanced data merupakan keadaan di mana distribusi kelas data yang tidak seimbang yaitu jumlah data yang satu lebih sedikit atau lebih banyak dari kelas lainnya. Untuk mengatasi data yang tidak seimbang ini dapat diatasi menggunakan algoritma oversampling atau *undersampling*. Didalam kasus ini data diseimbangkan dengan menerapkan salah satu algoritma dari oversampling yaitu *Borderline-SMOTE*. Algoritma ini dipilih karena menggabungkan informasi dari teknik SMOTE (Synthetic Minority Oversampling Techniuqe) dan border sampling (batas anatara kelas mayoritas dan minoritas. Borderline SMOTE bekerja dengan memilih sample *borderline* dan membuat sample sintetis di sekitarnya. Salah satu contoh kasus data tidak seimbang didapat pada peneliti Binanto, et al yang merupakan data LCMS dari tanaman Keladi Tikus hasil penelitian Sianipar, et.al. Data akan diklasifikasi menggunakan algoritma *K-Nearest Neighbour*, *Gaussian Naïve Bayes,* dan *Random Forest*.

Machine learning sering kali tidak dapat diandalkan ketika diterapkan pada data yang tidak seimbang. Namun, terdapat algoritma klasifikasi yang dapat mengatasi masalah ini, di antaranya adalah Random Forest T. M. Khoshgoftaar, et al (2015). KNN dan *Gaussian Naive Bayes*, dua metode klasifikasi *machine learning* yang populer, dapat mengidentifikasi data yang sudah seimbang dengan baik (Chandel et al., 2016). Dalam penelitian ini, eksperimen dilakukan pada data yang tidak seimbang dan data yang seimbang dengan menggunakan algoritma klasifikasi Random Forest, KNN, dan Gaussian NB.

2. DASAR TEORI

2.1. Random Forest

Random Forest didefinisikan sebagai prinsip umum suatu ansambel acak dari suatu pohon keputusan (Breiman, 2001). Bentuk umum dari model klasifikasi Random Forest dapat dilihat pada Gambar 1.



Gambar 1. Bagan Ilustrai Random Forest
(Sumber: Breiman, 2001)

Konstruksi Random Forest dapat dilakukan dengan tahapan sebagai berikut:

a. Menggambar sampel bootstrap n-tree dari data.
b. Menumbuhkan pohon untuk setiap kumpulan data bootstrap. Di setiap simpul pohon, pilih variabel entri secara acak untuk dipisahkan, lalu tumbuhkan pohon sehingga setiap no determinal memiliki tidak kurang dari kasus ukuran node.
c. Informasi agregat dari pohon n-tree untuk prediksi data baru seperti voting mayoritas untuk klasifikasi.
d. Hitung tingkat kesalahan out-of-bag (OOB) dengan menggunakan data bukan dalam sampel bootstrap. Metode penelitian ini didapatkan dari hasil perbandingan beberapa algoritma classifier dan ekstraksi fitur.

2.2. KNN

Alogaritma KNN merupakan salah satu metode yang dapat diterapkan dalam melakukan klasifikasi terhadap suatu data, dengan mencari data yang mempunyai jarak terdekat dengan suatu objek penelitian, sesuai dengan jumlah tetangga terdekatnya yang diinisialisasikan dengan K. Algoritma ini mengklasifikasikan data berdasarkan *similarity* atau kemiripan atau kedekatannya terhadap data lainnya. Secara umum, cara kerja algoritma KNN adalah sebagai berikut:

a. Tentukan jumlah tetangga (K) yang akan digunakan untuk pertimbangan penentuan kelas.
b. Hitung jarak dari data baru ke masing-masing data point di dataset.
c. Ambil sejumlah K data dengan jarak terdekat, kemudian tentukan kelas dari data baru tersebut.

Pencarian jarak terdekat biasanya dihitung menggunakan jarak Euclidean dengan persamaan sebagai berikut:

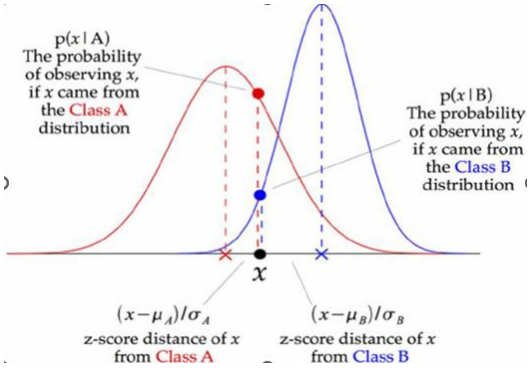
$$d\,(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - x_y)^2} \qquad (1)$$

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v13i1.9005
p-ISSN: 2964-5131

Keterangan:
$d\ (x, y)$ = Jarak Euclidean
x_i = Data Training Ke-i
x_y = Data Testing ke -i

2.3. Gaussian Naïve Bayes

Gausian Naïve Bayes adalah pendekatan dari algoritma Naïve Bayes, Naive Bayes sendiri merupakan salah satu algoritma klasifikasi yang berakar pada teorema bayes. Ciri utama algoritma ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing kondisi / kejadian (Wibawa et al., 2018). Metode Bayes Merupakan pendekatan statistic untuk melakukan inferensi induksi pada Teorema Bayes untuk melakukan klasifikasi pada Data Mining (Annur, 2018). Pendekatan Naive Bayes sendiri ada beberapa macam salah satunya adalah Gausian Naïve Bayes. Gaussian Naive Bayes mendukung fitur kontinu dan memodelkan setiap fitur sesuai dengan distribusi Gaussian (normal), saat membuat model sederhana, diasumsikan bahwa data dideskripsikan oleh distribusi Gaussian tanpa varian umum (tidak tergantung dimensi) antar dimensi. Model ini dapat dipasang dengan hanya mencari rata-rata dan standar deviasi dari peringkat untuk setiap label, yang cukup untuk mendapatkan distribusi yang baik (S A, 2021). Ilustrasi dari algoritma dapat dilihat pada gambar 2.



Gambar 2. Ilustrasi Gaussian Naïve Bayes
(sumber: (S A, 2021))

Ilustrasi diatas menunjukkan bagaimana pengklasifikasi Gaussian Naive Bayes (GNB) bekerja. Pada setiap titik data, jarak z-score antara titik tersebut dengan setiap rata-rata kelas dihitung, yaitu jarak dari rata-rata kelas dibagi dengan deviasi standar kelaster sebut, dengan demikian, kita melihat bahwa Gaussian Naive Bayes memiliki pendekatan yang sedikit berbeda dan dapat digunakan secara efisien. Langkah-langkah dalam melakukan algoritma ini adalah sebagai berikut:

a. Membaca data latih.
b. Menghitung jumlah dan probabilitas. Namun apabila dataset bersifat numerik maka :
   i. Mencari nilai mean dan standar deviasi dari masing masing parameter yang merupakan numerik
   ii. Mencari nilai probabilistic dengan menggunkan rumus sebagai berikut:

$$P(X_i = x_i | Y = y_j = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu j)^2}{2\sigma^2 ij}} \quad (2)$$

Keterangan:

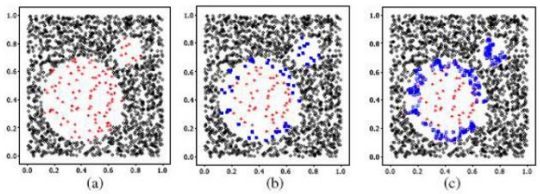
P: Peluang
X_i: Atribut ke i
x_i: Nilai atribut ke i
Y: Kelas yang dicari
y_j: Sub kelas Y yang dicari
μ : Mean menyatakan rata-rata dari seluruh atribut
σ : Deviasi standar

2.4. Borderiline SMOTE

Borderline-SMOTE adalah teknik oversampling data tidak seimbang yang menggabungkan informasi dari teknik SMOTE (Synthetic Minority Oversampling Techniuqe) dan border sampling (batas anatara kelas mayoritas dan minoritas. Borderline SMOTE bekerja dengan memilih sample borderline dan membuat sample sintetis di sekitarnya. Sample Borderline adalah sample yang terletak pada batas antara kelas mayoritas dan minoritas sehingga sangat penting untuk meningkatkan kinerja model pada data tidak seimbang (He & Garcia, 2009). Langkah-langkah pada borderline-SMOTE sebagai berikut (Abdurahman Baizal et al., 2009)

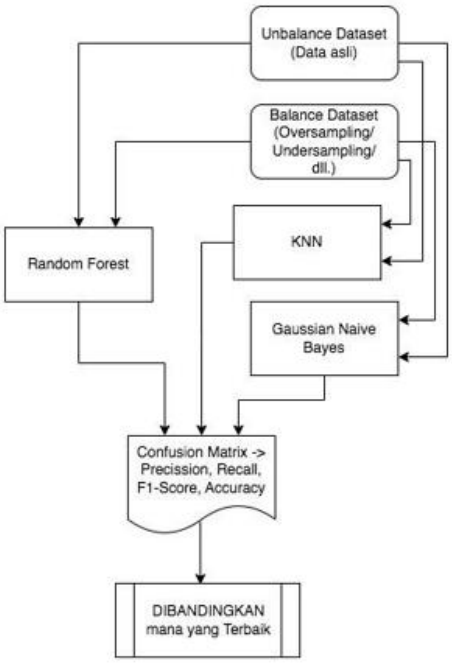
a. Tentukan k nearest neighbor untuk tiap data kelas minor
b. Periksa apakah data kelas minor masuk himpunan DANGER atau tidak. Himpunan DANGER adalah himpunan yang berisi data kelas minor dengan mayoritas nearest neighbornya adalah data kelas mayor.
c. Untuk tiap data di himpunan DANGER, lakukan proses SMOTE.

Ilustrasi Borderine-SMOTE dapat dilihat pada gambar 3.



Gambar 3. Ilustrasi Borderline-SMOTE
(Sumber: Abdurahman Baizal et al., 2009)

3. HASIL DAN PEMBAHASAN

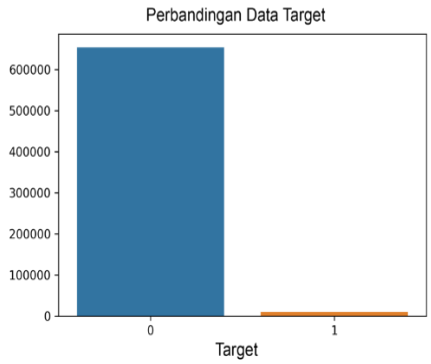


Gambar 4. Metode Penelitian

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v13i1.9005
p-ISSN: 2964-5131

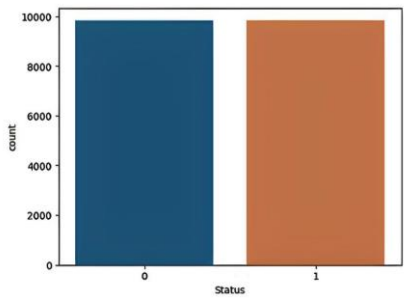
Penelitian ini dilakukan dengan menguji data tanaman keladi tikus yang berpotensi dalam pengobatan penyakit dengan jumlah data sebanyak 663.228 menggunakan metode KNN, *Gaussian Naïve Bayes*, dan *Random Forest.* Pengujian dilakukan dengan menguji dataset tidak seimbang (dataset asli) dan data yang sudah seimbang dengan membandingkan nilai accuracy, f1 score dan recall dari setiap algoritma untuk mendapatkan algoritma terbaik. Dengan tahapan-tahapan penelitian seperti gambar 4.

Pengujian dilakukan dengan data tidak seimbang (data asli). Data tidak seimbang disebabkan target biner yang menyatakan senyawa anti kanker dan senyawa biasa sangat kontras seperti terlihat pada gambar 5.



Gambar 5. Data Tidak Seimbang (data asli)

Setalah dilakukan penyeimbangan data menggunkan Borderline SMOTE data menjadi seimbang seperti pada gambar 6.



Gambar 6. Data sudah menerapkan Borderline SMOTE

Setelah dilakukan klasifikasi menggunkan KNN, *Gaussian Naïve Bayes*, dan *Random Forest* diperoleh hasil seperti pada tabel 1.

Tabel 1. Hasil Eksperimen

	Unbalance Data (Data Asli)			Balance Data dg metode yg dipilih		
	Random Forest	**KNN**	**Gaussian NB**	**Random Forest**	**KNN**	**Gaussian NB**
Precission	1	0,411	0	0,981	1	0
Recall	1	0,003	0	0,987	0,99	0
F1-Score	1	0.006	0	0,984	1	0
Accuracy	1	0,984	0,985	0,984	0,967	0,499

Dari tabel diatas algoritma yang menggunakan data tidak seimbang (data awal) algoritma KNN, *Gaussian Naïve Bayes*, dan *Random Forest* menghasilkan nilai akurasi berturut-turut *sebesar 0.984*, 0.985, dan 1, ini membuktikan bahwa Random Forest sedikit lebih baik dari algoritma KNN dan GNB dalam menangani data awal (data tidak seimbang) meskipun perbedaan akurasi tidak signifikan yang artinya ketiga algoritma ini dapat melakukan klasifikasi yang baik terhadap data. Pada klasifikasi yang dilakukan menggunkan data seimbang (borderline SMOTE) akurasi tertinggi diperoleh oleh algoritma Random Forest dan algoritma *Gaussian Naïve Bayes* memiliki akurasi yang buruk. Berdasarkan hasil tersebut algoritma Gaussian Naive Bayes sangat tidak disarankan untuk mengatasi data yang sudah diseimbangkan (Borderline SMOTE).

4. KESIMPULAN

Berdasarkan percobaan diatas dapat disimpulkan bahwa algoritma Gaussian NB menghasilkan score yang kurang baik dibandingkan dengan algoritma lain. Algoritma KNN hasil score lebih baik saat data balance. Sedangkan random forest memiliki tingkat akurasi, nilai recall, F1-score, dan Precission yang tinggi dibanding dengan algoritma yang lain pada data yang unbalance maupun balance. Random Forest juga tahan terhadap data unbalance, dapat kita lihat dari perubahan nilai scorenya yang tidak begitu jauh. Namun random forest membutuhkan waktu proses lebih lama dibanding KNN. Penelitian selanjutnya diharapkan untuk melakukan kombinasi algoritma pada saat melakukan balance data untuk klasifikasi menggunakan Gaussian NB.

DAFTAR PUSTAKA

Abdurahman Baizal, Z. K., Arif Bijaksana, M., & Sastrawan, A. S. (2009). ANALISIS PENGARUH METODE OVER SAMPLING DALAM CHURN PREDICTION UNTUK PERUSAHAAN TELEKOMUNIKASI. *Seminar Nasional Aplikasi Teknologi Informasi*.

Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., & Mukherjee, S. (2016). A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. *CSI Transactions on ICT*, *4*(2–4), 313–319. https://doi.org/10.1007/s40012-016-0100-5

Rahmatunissa, I. (2021). *Penanganan Data Tidak Seimbang Menggunakan Borderline Synthetic Minority Oversampling Technique (Borderline-SMOTE) pada Analisis Klasifikasi IKA RAHMATUNNISA, Drs. Zulaela, Dipl.Med.Stats., M.Si.*

Annur, H. (2018). KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES. In *Agustus* (Vol. 10, Issue 2).

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

S A, S. (2021). Comparative Study of Naive Bayes, Gaussian Naive Bayes Classifier and Decision Tree Algorithms for Prediction of Heart Diseases. *International Journal for Research in Applied Science and Engineering Technology*, *9*(3), 475–486. https://doi.org/10.22214/ijraset.2021.33228

Wibawa, A. P., Guntur, M., Purnama, A., Fathony Akbar, M., & Dwiyanto, F. A. (2018). Metode-metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1).

I. Binanto, H. L. H. S. Warnars, N. F. Sianipar, and W. Budiharto, "Webscraping Data Labeling System on Liquid Chromatography-Mass Spectrometry of Rodent Tuber for Efficiency of Supervised Learning Preprocessing," *ICIC Express Lett. Part B Appl.*, vol. 13, no. 1, pp. 107–114, 2022, doi: 10.24507/icicelb.13.01.107.

N. F. Sianipar and R. Purnamaningsih, "Enhancement of the contents of anticancer bioactive compounds in mutant clones of rodent tuber (Typhonium flagelliforme Lodd.) based on GC-MS analysis," *Pertanika J. Trop. Agric. Sci.*, vol. 41, no. 1, pp. 305–320, 2018.

N. F. Sianipar, M. Vidianty, Chelen, and B. S. Abbas, "Micropropagation of rodent tuber plant (Typhonium flagelliforme lodd.) from medan by organogenesis," *Pertanika J. Trop. Agric. Sci.*, vol. 40, no. 4, pp. 471–484, 2017.

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v13i1.9005 **p-ISSN: 2964-5131**

T. M. Khoshgoftaar, A. Fazelpour, D. J. Dittman and A. Napolitano, "Alterations to the Bootstrapping Process within Random Forest: A Case Study on Imbalanced Bioinformatics Data," 2015 IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 2015, pp. 342-348, doi: 10.1109/IRI.2015.59.