ANALISIS PERBANDINGAN ALGORITMA KNN, GAUSSIAN NAIVE BAYES, RANDOM FOREST UNTUK DATA TIDAK SEIMBANG DAN DATA YANG DISEIMBANGKAN DENGAN METODE TOMEK LINK UNDERSAMPLING PADA DATASET LCMS TANAMAN KELADI TIKUS

Ni Made Dina Aprilianti[1], Jeanytha Gein[2], Patricia Dian Paska[3], Iwan Binanto[4*] dan Nesti F. Sianipar[5]

[1,3,4]Informatika, Universitas Sanata Dharma, Yogyakarta

[2]Biotechnology Department, Faculty of Engineering, Bina Nusantara University, Jakarta

[*]Email: iwan@usd.ac.id

Abstrak

Data tidak seimbang adalah data yang mempunyai kelas mayoritas dan kelas minoritas dalam hal ini merupakan kelas target karena satu kelas melebihi jumlah kelas lain dalam dataset. Salah satu data tidak seimbang didapat pada penelitian Binanto, et. al. yang merupakan data LCMS dari tanaman Keladi Tikus hasil penelitian Sianipar et. al. Data ini tidak seimbang karena target biner yang menyatakan senyawa anti kanker dan senyawa biasa sangat kontras. Penelitian ini bertujuan untuk mengevaluasi potensi tanaman keladi tikus dalam pengobatan penyakit serta menjelaskan mekanisme yang mungkin terlibat. Untuk itu diperlukannya sebuah metode klasifikasi dokumen yang dapat mengelompokkan secara otomatis dan akurat. Terdapat banyak metode klasifikasi yang dapat digunakan. Metode yang digunakan dalam penelitian ini adalah Naive Bayes, Random Forest, dan KNN serta digunakan pula Algoritma Tomek Link Undersampling untuk menyeimbangkan data. Dari penelitian ini didapatkan bahwa Algoritma Random Forest merupakan algoritma yang paling tepat untuk menyelesaikan permasalahan Imbalanced Data maupun Balanced Data dengan menggunakan Tomek Links Undersampling karena algoritma ini memiliki nilai accuracy, precision, recall dan F1-Score yang tinggi dibanding algoritma lainnya.

Kata kunci *LCMS, imbalance data, oversampling/undersampling, Tomek Links Undersampling, klasifikasi*

1. PENDAHULUAN

Data yang tidak seimbang merujuk pada data yang memiliki kelas mayoritas dan kelas minoritas yang tidak sebanding. Salah satu kelas dalam dataset ini menjadi target karena jumlahnya melebihi kelas lainnya. Ketidakseimbangan data ini dapat mengakibatkan masalah pada beberapa algoritma klasifikasi dalam machine learning. Menangani ketidakseimbangan data merupakan tantangan tersendiri yang telah menjadi fokus penelitian selama dua dekade terakhir. Tujuannya adalah untuk mencapai hasil yang baik (Ancy & Paulraj, 2020; Rekha et al., 2021).

Penelitian yang dilakukan oleh Binanto et al. (2022) menggunakan data LCMS dari tanaman Keladi Tikus yang telah diteliti sebelumnya oleh Sianipar & Purnamaningsih (2018). Data ini tidak seimbang karena kontras antara senyawa anti kanker dan senyawa biasa yang direpresentasikan dalam target biner. Penggunaan data yang tidak seimbang dalam klasifikasi menggunakan machine learning akan menghasilkan hasil yang tidak akurat. Namun, terdapat algoritma klasifikasi yang mampu menangani data yang tidak seimbang, salah satunya adalah Random Forest.

Salah satu pendekatan untuk menyeimbangkan data yang tidak seimbang adalah melalui algoritma oversampling atau undersampling. Salah satu metode undersampling yang digunakan adalah algoritma Tomek Links. AT et al. (2016) melakukan penelitian yang menggabungkan metode Tomek Links dengan Random Undersampling (RUS) sebagai metode reduksi data. Penelitian ini membahas efektivitas kombinasi metode Tomek Links dalam menangani masalah ketidakseimbangan data. Metode tersebut telah diimplementasikan dan diuji pada dataset khusus yang digunakan dalam penelitian tersebut, sehingga memberikan pemahaman yang lebih baik tentang kinerja metode ini dalam konteks tertentu.

Penggunaan Tomek Links dengan teknik undersampling terbukti memberikan kinerja terbaik dalam menyeimbangkan data dan menghasilkan akurasi tertinggi dibandingkan dengan teknik lainnya (Alqaida et al., 2022).

Dengan mengurangi jumlah sampel dari kelas mayoritas yang tidak relevan atau ambigu menggunakan Tomek Links, metode ini membantu mengurangi dominasi kelas mayoritas dan meningkatkan kemampuan model untuk mempelajari pola dan fitur dari kelas minoritas. Penelitian tersebut menyajikan eksperimen dan analisis komprehensif untuk mendukung penggunaan Tomek Links sebagai metode reduksi data. Referensi jurnal tersebut memberikan bukti ilmiah dan temuan penelitian yang mendukung pilihan ini (AT et al., 2016).

Setelah data seimbang, klasifikasi dapat dilakukan dengan lebih baik menggunakan algoritma klasifikasi machine learning yang umum digunakan, seperti KNN dan Gaussian Naïve Bayes (Anand et al., 2022; Chandel et al., 2016; Safri et al., 2018). Penelitian ini melakukan eksperimen pada data yang tidak seimbang dan data yang seimbang menggunakan algoritma klasifikasi Random Forest, KNN, dan Gaussian NB.

2. LANDASAN TEORI

2.1. Random Forest

Random Forest adalah pengembangan dari metode Decision Tree yang menggunakan beberapa Decision Tree, dimana setiap Decision Tree telah dilakukan pelatihan menggunakan sampel individu dan setiap atribut dipecah pada pohon yang dipilih antara atribut subset yang bersifat acak. Random Forest memiliki beberapa kelebihan, yaitu dapat meningkatkan hasil akurasi jika terdapat data yang hilang, dan untuk resisting outliers, serta efisien untuk penyimpanan sebuah data. Selain itu, Random Forest mempunyai proses seleksi fitur dimana mampu mengambil fitur terbaik sehingga dapat meningkatkan performa terhadap model klasifikasi. Dengan adanya seleksi fitur tentu Random Forest dapat bekerja pada big data dengan parameter yang kompleks secara efektif. (Devella et al., 2020)

2.2. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah sebuah metode supervised yang berarti membutuhkan data training untuk mengklasifikasikan objek yang jaraknya paling dekat. Prinsip kerja K-Nearest Neighbor adalah mencari jarak terdekat antara data yang akan dievaluasi dengan k tetangga (neighbor) dalam data pelatihan (Nugroho Whidhiasih et al., 2013). Dekat maupun jauhnya letak (jarak) dapat dihitung dengan menggunakan rumus jarak seperti rumus jarak Euclidean serta jarak Minkowski. Dikarenakan memiliki tingkat akurasi yang besar, rumus jarak Euclidean sering digunakan dalam mengukur jarak antar data dengan rumus:

$$d(x_i, y_{i)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)\, 2}$$

Dimana:
d = jarak antar data
xi = sampel data
yi = data uji
i = variable data
n = dimensi data

2.3. Gaussian Naïve Bayes

Naive Bayes adalah algoritma yang mengklasifikasikan data berdasarkan perhitungan probabilitas kelompok dan menjumlahkan kombinasi nilai dari kumpulan data yang telah dikumpulkan (Irawan Saputra & Hakim, 2022; Raharja et al., 2021). Dalam penelitian ini data yang diolah adalah data numerik, untuk itu dalam menghitung nilai probabilitas kelas dapat menggunakan fungsi Probability Density Function (PDF). Dalam fungsi PDF dapat mewakili distribusi data yang diketahui, berikut adalah rumus PDF yang ditunjukkan pada persamaan 1 dan rumus standar deviasi yang ditunjukkan pada persamaan 2. Persamaan tersebut dikenal dengan formula Gaussian Naive Bayes Classifier sebagai berikut:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\bar{x})^2}{2\sigma^2}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n = 1}}$$

Dimana:

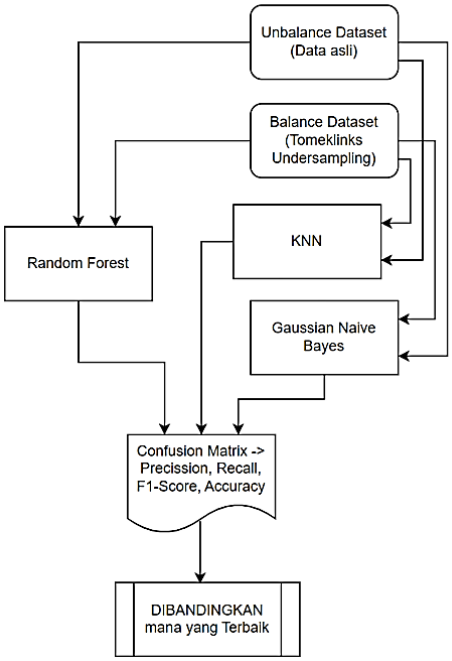
P : Probabilitas	yj : sub kelas yang berhubungan
Xi : Atribut	$\bar{x}$: rata-rata
xi : nilai atribut	σ : standar deviasi
Y : Kelas yang berhubungan	n : banyaknya data

2.4. Tomek Link Undersampling

Metode Tomek Links memilih sampel untuk dihapus. Metode ini memilih sampel dengan menemukan pasangan antar kelas (*cross-class pairs*) yang memiliki Euclidean Distance terkecil satu sama lain dalam ruang fitur. Metode ini dapat juga diterapkan untuk pembersihan data pasca pemrosesan untuk menghapus sampel baik dari kelas mayoritas maupun minoritas, karena wilayah batas yang kurang terdefinisi dengan baik (Swana et al., 2022). Jika sampel di kelas minoritas dianggap konstan, maka sampel tersebut dapat digunakan untuk menemukan semua contoh di kelas mayoritas yang paling dekat dengan kelas minoritas. Oleh karena itu, sampel yang dihasilkan oleh Tomek Links adalah sampel batas atau noise. Hal ini disebabkan fakta bahwa hanya sampel pada batas kelas dan noise yang memiliki tetangga terdekat yang berasal dari kelas lawan.

3. METODOLOGI PENELITIAN

Penelitian ini secara umum terdiri atas beberapa tahapan yang dimulai dari pengumpulan data, pemrosesan data yang tidak seimbang (*unbalance* data) diproses menggunakan algoritma K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, dan Random Forest. Data yang seimbang sebelumnya akan diseimbangkan terlebih dahulu menggunakan algoritma tomek link undersampling yang kemudian akan diproses menggunakan ketiga algoritma yang sama. Hasil pengolahan kedua data tersebut akan dievaluasi menggunakan confusion matrix yang nantinya akan menghasilkan nilai *accuracy, F1-Score, recall, dan precision* yang hasilnya akan dibandingkan. Tahap penelitian ini diilustrasikan pada gambar 1.



Gambar 1. Metode Penelitian

e-ISSN: 2964-2531
DOI : 10.36499/psnst.v13i1.9002
p-ISSN: 2964-5131

Data yang digunakan berjumlah 663.228. Data tersebut merupakan data tidak seimbang. Maka dari itu data akan diseimbangkan menggunakan algotirma Tomek Link Undersampling.

Dari data yang tidak seimbang dan data yang telah diseimbangkan akan dilakukan pemodelan menggunakan K-Nearest Neighbor (KNN), Gaussian Naïve Bayes dan Random Forest. Dari data tersebut akan dibagi menjadi data training dan data testing dengan perbandingan 80:20.

4. HASIL DAN DISKUSI

Pengujian data tidak seimbang dan data seimbang menggunakan parameter yang sama, yaitu untuk perbandingan training dan testing data adalah 80:20, random_state = 42, nilai k pada KNN adalah akar dari jumlah data dan menggunakan Euclidean Distance.

Tabel 1. Hasil Eksperimen Data Tidak Seimbang

	Random Forest	KNN	Gaussian NB
Precision	0.9956	0.4400	0.0000
Recall	0.3439	0.0111	0.0000
F1-Score	0.5112	0.0217	0.0000
Accuracy	0.9901	0.9850	0.9852

Tabel 2. Hasil Eksperimen Data Seimbang

	Random Forest	KNN	Gaussian NB
Precision	1.0000	0.4444	0.0000
Recall	0.9994	0.0184	0.0000
F1-Score	0.9997	0.0355	0.0000
Accuracy	0.9999	0.9851	0.9852

Dari hasil kedua tabel diatas dapat dilihat bahwa selisih hasil dari precision, recall, F1-Score, dan accuracy antara data tidak seimbang dan data seimbang tidak selisih jauh namun dapat dilihat bahwa algoritma yang terbaik adalah algoritma Random Forest. Dari kedua eksperimen juga dapat dilihat bahwa data yang sudah diseimbangkan lebih memiliki akurasi yang lebih tinggi.

5. KESIMPULAN

Algoritma Random Forest merupakan algoritma yang paling tepat untuk menyelesaikan permasalahan Imbalanced Data maupun Balanced Data dengan menggunakan Tomek Links Undersampling karena algoritma ini memiliki nilai accuracy, precision, recall dan F1-Score paling tinggi dibanding algoritma lainnya. Saran yang dapat diberikan untuk penelitian yang telah dilakukan adalah, perlu menambah dan melengkapi jurnal yang mendukung topik dalam makalah serta megembangkan pembahasan menggunakan bahasa sendiri.

Untuk penelitian kedepannya kami berencana untuk melakukan penyesuaian pada algoritma Tomek Link dengan memperkenalkan teknik pembobotan kelas atau menggabungkannya dengan metode undersampling lainnya. Tujuan kami adalah meningkatkan akurasi klasifikasi dan memperbaiki kinerja algoritma ini pada dataset yang tidak seimbang.

DAFTAR PUSTAKA

Alqaida, R. A., Ngurah, G., Wibawa, A., Yahya, I., Abapihi, B., Laome, L., & Oleo, U. H. (2022). *Combine Undersampling Untuk Menangani Data Tidak Seimbang Pada Lama Belajar Siswa di*

Rumah. *April*, 71–78.

Anand, M. V., Kiranbala, B., Srividhya, S. R., C., K., Younus, M., & Rahman, M. H. (2022). Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/2436946

Ancy, S., & Paulraj, D. (2020). Handling imbalanced data with concept drift by applying dynamic sampling and ensemble classification model. *Computer Communications*, *153*(January), 553–560. https://doi.org/10.1016/j.comcom.2020.01.061

AT, E., M, A., F, A.-M., & M, S. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Global Journal of Technology and Optimization*, *01*(S1). https://doi.org/10.4172/2229-8711.s1111

Binanto, I., Warnars, H. L. H. S., Sianipar, N. F., & Budiharto, W. (2022). Webscraping Data Labeling System on Liquid Chromatography-Mass Spectrometry of Rodent Tuber for Efficiency of Supervised Learning Preprocessing. *ICIC Express Letters, Part B: Applications*, *13*(1), 107–114. https://doi.org/10.24507/icicelb.13.01.107

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

Devella, S., Yohannes, Y., & Rahmawati, F. N. (2020). Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *7*(2), 310–320. https://doi.org/10.35957/jatisi.v7i2.289

Irawan Saputra, D., & Hakim, D. L. (2022). Implementasi Algoritma Gaussian Naive Bayes Classifier Untuk Prediksi Potensi Tsunami Berbasis Mikrokontroler. *EPSILON: Journal of Electrical Engineering and Information Technology*, *20*(2), 122–138. https://doi.org/10.55893/epsilon.v20i2.94

Nugroho Whidhiasih, R., Adi Wahanani, N., & Supriyanto. (2013). Klasifikasi Buah Belimbing Berdasarkan Citra Red-Green-Blue Menggunakan KNN Dan LDA | PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic. *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic*, *1*(1), 29–35. https://jurnal.unismabekasi.ac.id/index.php/piksel/article/view/288

Raharja, K. Y., Oktavianto, H., & Umilasari, R. (2021). *Perbandingan Kinerja Algoritma Gaussian Naive Bayes Dan K-Nearest Neighbor (Knn) Untuk Mengklasifikasi Penyakit Hepatitis C Virus (Hcv)*. 1–12.

Rekha, G., Tyagi, A. K., Sreenath, N., & Mishra, S. (2021). Class Imbalanced Data: Open Issues and Future Research Directions. *2021 International Conference on Computer Communication and Informatics, ICCCI 2021*. https://doi.org/10.1109/ICCCI50826.2021.9402272

Safri, Y. F., Arifudin, R., & Muslim, M. A. (2018). K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. *Scientific Journal of Informatics*, *5*(1), 18. https://doi.org/10.15294/sji.v5i1.12057

Sianipar, N. F., & Purnamaningsih, R. (2018). Enhancement of the contents of anticancer bioactive compounds in mutant clones of rodent tuber (Typhonium flagelliforme Lodd.) based on GC-MS analysis. *Pertanika Journal of Tropical Agricultural Science*, *41*(1), 305–320.

Swana, E. F., Doorsamy, W., & Bokoro, P. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors*, *22*(9). https://doi.org/10.3390/s22093246