8th Seminar Nasional Teknik Elektro, Informatika dan Sistem Informasi (SINTaKS)
26 Agustus 2023 - Yogyakarta, Indonesia
ISSN: 2961-7464

Perbandingan Algoritma Klasifikasi KNN, Gaussian Naive Bayes, Dan Random Forest Pada Dataset LCMS Tanaman Keladi Tikus Yang Diseimbangkan Dengan Metode Synthetic Minority Over-Sampling Technique

Vina Meriana[1], Ni Komang Ayu Wirayanti[2], Julius Rakha Bowo Laksono[3], Iwan Binanto [4*], Nesti F. Sianipar[5]

[1,2,3,4]*Jurusan Informatika, Universitas Sanata Dharma*
[5]*Biotechnology Department, Faculty of Engineering, Bina Nusantara University*
[1]felisiavina07@gmail.com
[2]komangayuwirayanti@gmail.com
[3]juliusrakha7@gmail.com
[4*]iwan@usd.ac.id
[5]nsianipar.binus.edu

***Abstrak*–Data yang tidak seimbang dapat mempengaruhi tingkat akurasi dalam klasifikasi, dan salah satu metode yang digunakan untuk menyeimbangkan data adalah Synthetic Minority Over-sampling Technique (SMOTE), yang merupakan teknik oversampling untuk menghasilkan data sintetis dari kelas minoritas. Pada penelitian ini digunakan algoritma Random Forest, KNN, dan Gaussian Naïve Bayes untuk klasifikasi. Metode penelitian yang digunakan meliputi pengumpulan data, preprocessing data, pemrosesan ketidakseimbangan data, pembagian data menjadi subset pelatihan dan pengujian, implementasi algoritma, evaluasi kinerja menggunakan metrik evaluasi klasifikasi, analisis hasil, uji statistik, kesimpulan, dan saran. Berdasarkan eksperimen didapat hasil bahwa algoritma Random Forest merupakan algoritma yang mempunyai akurasi tertinggi dibandingkan kedua algoritma yang lain, baik tu pada data tidak seimbang maupun data yang sudah diseimbangkan dengan metode Synthetic Minority Over-sampling Technique (SMOTE).**

***Kata kunci*– Synthetic Minority Over-sampling Technique, Klasifikasi, Imbalance Data, Keladi Tikus.**

I. PENDAHULUAN

Data tidak seimbang atau unbalance datset adalah suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (instance) yang satu lebih sedikit atau lebih banyak dibandingkan meneggunakan jumlah kelas data lainnya. Dalam mengklasifikasi sebuah dookumen dengan data yang tidak seimbang dapat menghasilkan tingkat accuracy yang berbeda. Salah satu data yang tidak seimbang adalah data LCMS dari tanaman Keladi Tikus yang mana data tersebut tidak seimbang[1].

Machine learning adalah bidang kecerdasan buatan atau artificial intelligence yang berkaitan dengan pengembangan teknik-teknik yang dapat diprogram dan dipelajari dari data masa lalu. Algoritma KNN adlah sebuah metode yang digunakan untuk melakukan klasifikasi terhadap objek berdasarkan data latih yang memiliki jarak paling dekat dengan objek tersebut. Klasifikasi pada imbalance class akan cenderung mengabaikan kelas yang memiliki jumlah sample yang sedikit sehingga dapat berpengaruh buruk terhadao performa dari algoritma klasifiksi. Salah satu cara untuk menangani unbalance class yaitu menggunakan teknik Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE adalah algoritma sintesis yang ditunjukkan untk memperbaiki data dengan mensintesis data latiih mengikuti teori interpolasi fraktal, sehingga data yang dihasilkan lebih representatif, dan menghasilkan kinerja lebih baik.

Data yang tidak seimbang dapat diseimbangkan dengan beberaoa algoritma sampling adalah SMOTE. Dimana SMOTE merupakan metode Random over Sampling yang mana Teknik Oversampling ini dilakukan dengan menggandakan atau mengulang sampel-samle dari kelas minoritas hingga mencapai proposi yang lebih seimbang dengan kelas mayoritas. Kemudian, Gaussian Nave Bayes diterapkan pada dataset yang telah diseimbangkan menggunakan teknik oversampling ini, Gaussian Nave Bayes adalah algoritma klasifikasi probabilistik yang populer karena sederhana dan efisien, tetapi performanya dapat menurun ketika diterapkan pada dataset dengan ketidakseimbangan kelas yang signifikasn[5-7].

Random forest seringkali digunakan dalam prediksi dan klasifikasi, karena tingkat akurasi dari prediksinya yang tinggi. Kita melakukan pengamatan dengan menggunakan data tanaman. Setelah dilakukan pemisahan data menjadi data latih dan data uji, akan dilakukan klasifikasi menggunakan random forest. Dalam random forest data langsung ditransformasi menggunakan standar scaler, selanjutnya hasil kalasifikasi dari random forest, kemudian dilanjutkan dengan melakukan perhitungan berdasarkan confusion matrix[4].

Pada paper ini akan dilakukan sebuah pengujian untuk dataset dengan menggunakan beberapa metode random forest, KNN, dan GNB. Sehingga kita akan mendapatkan hasil penggujian yang akan menampilkan nilai hasil accuracy, precission, recal dan f1-score untuk data yang tidaka seimbang dan data yang seimbang.

II. STUDI PUSTAKA

A. *K-Nearst Neighbor (KNN)*

K-Nearst Neighbor (KNN) termasuk kelompok instance-based learning. KNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. Algoritma KNN ini juga bisa digunakan untuk melakukan prediksi ataupun klasifikasi terhadap suatu data tergantung dari jenis data pada kumpulan daya yang ada. Algoritma ini melakukan klasifikasi pada suatu data berdarkan nilai k yang telah ditetapkan sebelumnya[1]. Nilai k pada K-NN harus menggunakan nilai ganjil jika digunakan untuk melakukan prediksi nilai k pada K-NN dapat berupa bilangan ganjil ataupun genap. Perhitungan jarak yang digunakan pada K-NN menggunakan Euclidean Distance[8-9].

B. *Gaussian Naïve Bayes*

Gaussian Naïve Bayes merupakan penegklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdarkan pengalaman di masa sebelumnya. Algoritma ini mengasumsikan bahwa atribut objek adalah independen. Pengklasifikasi Bayesian memiliki tingkat kesalahan minimal dibandingkan dengan klasifikasi lainnya dan NBC dipilih karena performa dan kesempatan yang tinggi dalam proses klasifikasi, dan mudah untuk menghasilkan probabilitas posterior data yang dites terhadap kelasnya, ada beberapa model yang digunakan untuk pengklasifikasian NBC, salah satunya adalah Gaussian Naïve Bayes[10-13].

C. *Random Forest*

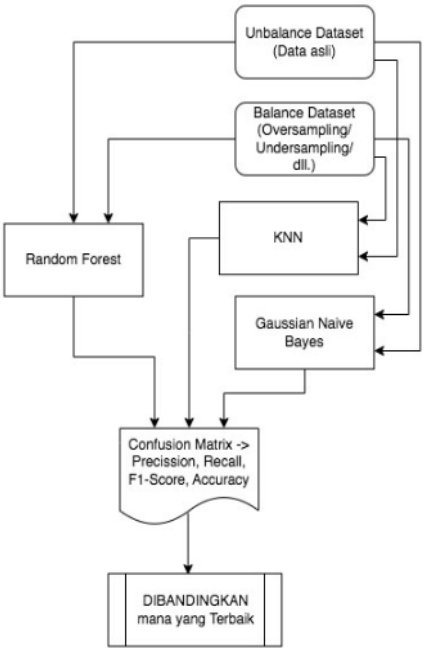
Random forest adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini merupakan kombinasi dari beberapa pohon keputusan yang independen yang dihasilkan dengan membagi data menjadi subset secara acak dan menggabungkan prediksi dari setiap pohoon untuk mencapai hasil akhir. Metode ini efekti untuk memperkirakan data yang hilang dan mempertahankan akurasi ketika sebagian besar data hilang dan memiliki metode menyimbangkan kesalahan kelas pada kumpulan data yang populasinya tidak seimbang. Peneliatan di bidang Random Forest bertujuan untuk meningkatkan akurasi, atau meningkatkan kinerja (mengurangi waktu yang dibutuhkan untuk pembelajaran dan klasifikasi), atau keduanya.

D. *Random Over Sampling*

Random Over Sampling (ROS) merupakan salah satu teknik reampling yang digunakan dalam penanganan ketidakseimbanagn kelas pada data dimana dengan cara bekerja dengan menduplikat secara acak sample-sample dari kelas minoritas hingga jumlah sampel dalam kelas minoritas seimbang dengan jumlah sample dalam kelas mayoritas.

III. METODE PENELITIAN

Pada paper ini digunakan metode menguji dataset asli yang masih tidak seimbang dengan algoritma Randim Forest, KNN, Gaussian Naïve Bayes Demikian juga dengan dataset yang sudah diseimbangkan dengan SMOTE. Metode atau alur diagram dalam penilitian ini terlihat pada gambar 1.



Gambar 1. Diagram alur program

- Pengumpulan data
  Data kladi tikus dikumpulkan dari sumber yang relevan dan data terdiri dari atribut dan label mewakili kelas kladi tikus yang berbeda.
- Preprocessing data
  Data yang terkumpul akan melalui proses preprocessing, temasuk penghapusan data yang hilang, normalisasi atribut, dan pemilihan firue yang relevan (jika diperlukan)
- Pemerosesan ketidak seimbangan data
  Untuk menguji kinerja pada data yang mengalami ketidak seimbangan data, teknik pemerosesan tidakseimbang data akan diterapkan, seperti undersampling, oversampling. Teknik-teknik ini akan diguanakan untuk menghasilkan versi dataset yang seimbang.
- Pembagain data
  Data kladi tikus akan dibagi menjadi subset pelatihan(training set) dan subset pengujian (testeing set). Pembagian ini akan mempertahankan proposi kelas yang

sama baik pada data yang seimbang maupun tidak seimbang.

- Implementasi algoritma
  Algoritma randim forest KNN, dan GNB akan dimplementasikan pada kedua versi data, yaitu ketidakseimbangan (unbalance dataset) dan data seimbang (balance dataset), parameter yang tepat akan diatur untuk setiap algortitma sesuai dengan kebutuhan.
- Evaluasi kinerja
  Kinerja algoritma akan dievaluasi klasifikasi seperti akurasi, presisi, recall, f1-score dan area under the curve (AUC). Evaluasi dilakukan pada kedua versi data untuk membandingkan kinerja algoritma pada kondisi unbalance dan balance dataset.
- Analisis hasil
  Hasil evaluasi akan dianalisi untuk mengamati perbedaan kinerja algoritma pada data yang mengalami ketidakseimbangan dan keseimbangan data, perbandingan kinerja antara algoritma juga akan dilakukan.

IV. HASIL DAN DISKUSI

Pengujian data tidak seimbang (unbalance) dan data seimbang (balance) menggunakan parameter yang sama, yaitu untuk perbandingan training dan testing data adalah 70 – 30, random_state = 42, nilai k pada KNN adalah akar dari jumlah data dan menggunakan Euclidean Distance.
Berikut adalah hasil pengujian unbalance dan balance data pada data keladi tikus menggunakan algoritma Random Forest, KNN, dan Gaussian Naïve Bayes:

Tabel 1. Hasil eksperimen

	Unbalance Data (Data Asli)			Balance Data dg SMOTE		
	Random Forest	KNN	Gaussian NB	Random Forest	KNN	Gaussian NB
Precission	1	0	0	1	0.345	0.874
Recall	1	0	0	1	1	0.014
F1-Score	0.99	0	0	1	0.513	0.028
Accuracy	0.99	0.985	0.985	1	0.972	0.119

Gambar 2. Hasil pengujian terhadap data unbalance dan balance dataset.

V. KESIMPULAN

A. Kesimpulan

Pada tabel hasil eksperimen di atas bisa dilihat untuk data unbalance algoritma random forest memperoleh performa yang sangat baik dengan presisi, recall, dan F1-score yang sempurna. Ini dapat disebabkan oleh kemampuan algoritma Random Forest dalam menangani ketidakseimbangan kelas. Untuk data balance, model KNN mengalami peningkatan yang signifikan dalam mengklasifikasikan kelas minoritas setelah resampling dengan SMOTE, terlihat dari peningkatan presisi, recall, F1-score, dan akurasi yang signifikan. Hal ini menunjukkan bahwa SMOTE berhasil meningkatkan kemampuan model KNN dalam mengenali dan mengklasifikasikan kelas minoritas. Model Gaussian NB, meskipun mengalami peningkatan dalam presisi dan F1-score setelah resampling, tetap memiliki kinerja yang rendah dalam mengklasifikasikan kelas minoritas, terlihat dari recall dan akurasi.

B. Saran

Pada penelitian in dapat dikembangkan lebih lanjut dengan mencoba atau menerapkan metode klasifikasi lainnya.

REFERENSI

[1] G. Rekha, A.K. Tyagi, N. Sreenath, S. Mishra, “Class Imbalanced Data: Open Issues and Future Research Directions,” 2021 International Conference on Computer Communication and Informatics, ICCCI 2021, 2021, doi:10.1109/ICCCI50826.2021.9402272.

[2] Anis Nikmatul Kasanah, Muladi, Utomo Pujianto, U. Negeri Malang, " Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class Dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN", Vol. 3No. 2(2019)196-201ISSN Media Elektronik: 2580-0760

[3] R. Perangin-angin, E. J. G. Harianja, dan I. K. Jaya, “Pendekatan Level Data untuk Menangani Ketidakseimbangan Data Menggunakan Algoritma K-Nearest Neighbor”, *JTM*, vol. 9, no. 1, hlm. 22–32, Mei 2020.

[4] A. Saifudin, U. Pamulang, R. S. Wahono, U. Dian, and N. Semarang, “Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software,” J. Softw. Eng., vol. 1, no. 2, pp. 76–85, 2015

[5] Lukhia Britanthia Christina Tanujayaa, ,Bambang Susanto, Asido Saragih, U. Kristen Satya Wacana, " Perbandingan Metode Regresi Logistik dan R andom Forest untuk Klasifikasi Fitur Mode Audio Spotify," J. Indonesian Journal of Data and Science (IJODAS) ISSN: 2715-9930 Vol 1, No 3, Desember 2020, pp. 68-78.

[6] I. Binanto, H.L.H.S. Warnars, N.F. Sianipar, W. Budiharto, “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, 2022, doi:10.24507/icicelb.13.01.107.

[7] Abid Isha, U. Khwaja Fareed, Saima sadiq,U. Torrens Australia, Muhammad Umer, U. Yonsei, Saleem Ullah, U. Islamia Bahawalpur, Seyedali Mirjalili, U. King Abdulaziz, Vaibhav Rupapara, U. Florida International, Michele Nappi, U. Salemo, " Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques", Digital Object Identifier 10.1109/ACCESS.2021.3064084

[8] A. Syukron and A. Subekti, “*Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit,*” *JURNAL INFORMATIKA*, vol. 5, no. 2, 2018.

[9] S. Bagui, K. Li, “Resampling imbalanced data for network intrusion detection datasets,” Journal of Big Data, **8**(1), 2021, doi:10.1186/s40537-020-00390-x.

[10] Henny Leidiyana, U. r STMIK Nusa Mandiri, "Penerapan Algoritma K-Nearest Neighbor untuk penentu resiko kredit kepemilikan kendaraan bermotor", Jurnal Penelitian Ilmu Komputer, System Embedded & Logic 1(1) : 65-76 (2013)

[11] Ahmad Khairi, Achmad Fais Ghozali, Ach Darul Nur Hidayah, U. Nurul Jadid, " IMPLEMENTASI K-NEAREST NEIGHBOR (KNN) UNTUK KLASIFIKASI MASYARAKAT PRA SEJAHTERA DESA SAPIKEREP KECAMATAN SUKAPURA", P-ISSN: 2774-4574 ; E-ISSN: 2774-4582 TRILOGI, 2(3), September-Desember 2021 (319-323)

[12] Ansfridus Bravo, Tursina, Helen Sastypratiwi, U. Tanjungpura, " Penerapan Metode Naive Bayes Untuk Penentuan Bibit Kelapa Sawit Berdasarkan Kondisi Daerah Tanam dan Perawatan Tanaman", Vol. 11, No. 1, Januari 2023, p-ISSN : 2460-3562 / e-ISSN : 2620-8989.

[13] S. Ghosh dan S. Kumar, "Analisis Perbandingan Algoritma K-Means dan Fuzzy C-Means," Int. J. Adv. Comput. Sci. Appl., 2013, doi: 10.14569/IJACSA.2013.040406

[14] Zahra Putri Agusta, U. Surya, Adiwijaya, U. Telkom,Bandung, "Modifikasi hutan acak seimbang untuk meningkatkan prediksi data yang tidak seimbang", Vol. 5, No. 1, Maret 2019, hal. 58-65.