Jurnal Dinamika Informatika
Volume 14, No 2, Oktober 2025
ISSN 1978-1660 : 73-84
ISSN *online* 2549-8517

Penanganan *Imbalance Data* Klasifikasi Teks Lowongan Pekerjaan: Komparasi *Algorithmic* vs *Data Level*

Nurrofiqi Ankisqiantari
Bisnis Digital, Universitas PGRI Yogyakarta
e-mail: **isqia@upy.ac.id**

Intisari

Klasifikasi otomatis tingkat pengalaman kerja (*experience level)* pada situs lowongan pekerjaan merupakan tantangan krusial dalam sistem *e-recruitment* untuk memetakan kebutuhan industri. Tantangan utama dalam penelitian ini adalah ketimpangan data (*imbalanced data*) dimana jumlah lowongan level '*Senior*' jauh lebih sedikit dibandingkan level '*Fresh Graduate*' dan level '*Junior/Mid*', serta ambiguitas semantik pada deskripsi pekerjaan. Penelitian ini bertujuan untuk membandingkan efektivitas penanganan *imbalanced data* menggunakan dua pendekatan berbeda, yaitu pendekatan *Algorithmic Level* menggunakan *Class Weighting* dan pendekatan *Data Level* menggunakan *Synthetic Minority Over-sampling Technique* atau SMOTE pada *Random Forest* dan SVM. Penelitian ini dilakukan pada dataset: 2001 lowongan kerja riil di Indonesia dengan ekstraksi fitur *TF-IDF*. Hasil penelitian menunjukkan bahwa pendekatan *Algorithmic Level* menggunakan SVM dengan *Class Weight* memberikan performa terbaik dalam mendeteksi kelas minoritas, mencapai *Recall* 47% untuk kelas Senior, mengungguli SVM dengan SMOTE yang hanya mencapai 34%. Temuan ini mengindikasi bahwa pada data teks berdimensi tinggi dengan *lexical overlap* yang signifikan, teknik *oversampling* sintetis (SMOTE) cenderung memperkenalkan *noise* yang mengaburkan batas keputusan (*decision boundary*), sehingga modifikasi bobot algoritma menjadi solusi yang lebih *robust.*

Kata kunci— *Imbalance Data,* Klasifikasi Teks, SVM, SMOTE, *Class Weighting*

Abstract

Auotomatic classification of job experience levels on job portals presents a crucial challenge in e-recuitment systems for mapping industry requirements. The primary challenges addressed in this study are data imbalance, where the number of 'Senior' level vacancies is significantly lower than 'Fresh Graduate' and 'Junior/Mid' levels, and the semantic ambiguity present in job descriptions. This study aims to compare the effectiveness of handling imbalanced data using two distinct approaches: the Algorithmic Level approach utilizing Class Weighting, and the Data Level approach utilizing the Synthetic Minority Over-sampling Technique (SMOTE), applied to Random Forest and Support Vector Machine (SVM) models. The research was conducted on a dataset comprising 2001 real-world job vacancies in Indonesia, employing TF-IDF for feature extraction. The results indicate that the Algorithmic Level approach using SVM with Class Weight yielded the best performance in detecting the minority class, achieving a Recall of 47% for the Senior level, outperforming SVM with SMOTE, which only achieved 34%. These findings indicate that in high-dimensional text data characterized by significant lexical overlap, synthetic oversampling techniques (SMOTE) tend to introduce noise that obscure the decision boundary, making algorithmic weight modification a more robust solution.

Keywords— *Imbalanced Data, Text Classification, SVM, SMOTE, Class Weighting*

Jurnal Dinamika Informatika
Volume 14, No 2, Oktober 2025
ISSN 1978-1660 : 73-84
ISSN *online* 2549-8517

PENDAHULUAN

Pertumbuhan pesat industri digital telah menghasilkan lonjakan volume data lowongan pekerjaan daring. Data ini memiliki potensi besar untuk digali informasinya melalui teknik *data mining* dan *machine learning* guna memahami tren keahlian pasar. Melalui klasifikasi otomatis, perusahaan dan pencari kerja dapat menyaring informasi secara efisien, namun proses ini dihadapkan pada tantangan yang signifikan terkait karakteristik data dunia nyata.

Masalah inti yang dihadapi adalah ketidakseimbangan kelas (*Imbalanced Class*), dimana jumlah sampel dalam satu kelas jauh melebihi kelas lainnya [1][2]. Dalam konteks data lowongan pekerjaan, posisi level '*Senior*' secara inheren merupakan kelas minoritas dibandingkan dengan posisi level '*Fresh Graduate*'. Konsekuensinya algoritma klasifikasi standar, yang umumnya berasumsi pada distribusi kelas yang seimbang, cenderung menjadi bias terhadap kelas mayoritas dan gagal mengenali kelas minoritas [2][3]. Padahal, kelas minoritas seringkali merupakan kelas yang paling penting untuk dideteksi [4].

Secara umum, terdapat dua pendekatan utama untuk menangani masalah ini, yaitu pendekatan level data (*data level*) dan pendekatan level algoritma (*algorithmic level*) [5][6]. Pendekatan level data berfokus pada modifikasi distribusi data pelatihan melalui *resampling,* seperti SMOTE (*Synthetic Minority Over-sampling Technique*), yang bekerja dengan membangkitkan data sintetis. Namun, beberapa studi mengkritik potensi SMOTE dalam menghasilkan data "palsu" yang tidak realistis, terutama pada data kompleks, yang dapat menyesatkan proses pembelajaran model [2]. Disisi lain, pendekatan level algoritma, seperti *cost-sensitive learning* melalui *Class Weighting,* memodifikasi algoritma untuk memberikan penalti lebih besar pada kesalahan klasifikasi kelas minoritas tanpa mengubah data asli.

Penelitian ini bertujuan untuk mengisi celah literatur dengan melakukan komparasi antara efektivitas SMOTE (*Data Level*) dengan *Class Weighting* (*Algorithmic Level*). Fokus penelitian ini pada kasus klasifikasi teks lowongan pekerjaan yang tidak hanya tidak seimbang tetapi juga memiliki tingkat ambiguitas

semantik yang tinggi. Dengan demikian, penelitian ini akan menguji ketahanan (*robustness*) masing-masing pendekatan.

METODE PENELITIAN

Berdasarkan Gambar 1 Alur Penelitian ini berupa pengumpulan dataset, kemudian dilanjut dengan pembersihan data (pra-pemrosesan data), *data splitting,* lalu untuk penanganan *imbalanced data* dibagi menjadi dua, yaitu *Algorithmic Level,* dan *Data Level.* Setelah itu dilakukan pemodelan dengan metode *Random Forest* dan SVM. Terakhir dilakukan evaluasi model.

Gambar 1 Alur Penelitian

2.1 Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 2100 data lowongan pekerjaan khusus bidang IT yang dikumpulkan dari berbagai situs di Indonesia, yaitu Glints, JobStreet, Indeed, Kitalulus, Kalibrr, dan Prosple. Fitur yang digunakan dalam dataset ini adalah judul pekerjaan, syarat skill, deskripsi pekerjaan, dan label target penelitian ini adalah tingkat pengalaman kerja yaitu, *Fresh Graduate, Junior/Mid,* dan *Senior.*

Jurnal Dinamika Informatika
Volume 14, No 2, Oktober 2025
ISSN 1978-1660 : 73-84
ISSN *online* 2549-8517

2.2 Pra-Pemrosesan Data Teks (*Text Pre-Processing*)

Mempersiapkan data sebelum diproses lebih lanjut, dilakukan pra-pemrosesan data. Pra-pemrosesan yang dilakukan meliputi:

a. *Cleaning*

Proses ini menangani *missing values* untuk memastikan kualitas data. Selanjutnya menghapus URL, angka, dan karakter lainnya. Simbol teknis penting seperti "C++, C#, .NET, dan lainnya" dipertahankan karena merupakan fitur yang relevan. Kemudian dilakukan proses *case folding* untuk mengubah semua teks menjadi huruf kecil [7].

b. *Stopword Removal*

Proses ini menghilangkan kata-kata yang umum (sering muncul dalam kalimat) yang tidak membawa makna signifikan dalam data, seperti "yang, dan, juga, lainnya, dst"[7]. Penghapusan *stopword* bahasa Indonesia maupun Inggris.

c. *Feature Engineering*

Menggabungkan kolom Judul Pekerjaan, Syarat Skill, dan Deskripsi Pekerjaan menjadi satu kolom. Dengan menggabungkan menjadi satu kolom untuk memperkaya dan menangkap hubungan semantik secara mendetail disetiap kalimat [8].

d. TF-IDF (*Term Frequency-Inverse Document Frequency*)

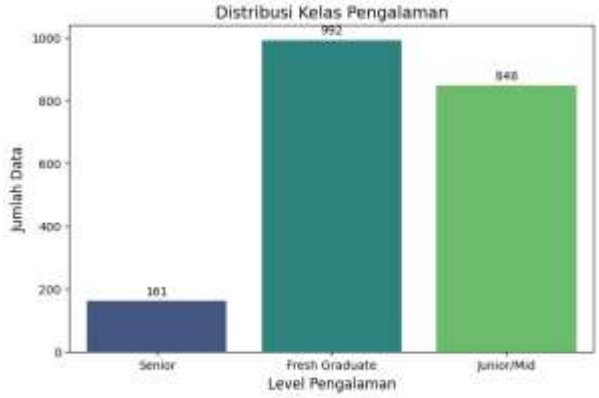
Proses ini mengubah teks menjadi representasi numerik. TF-IDF ini memberikan bobot lebih tinggi pada kata-kata yang banyak muncul dalam suatu dokumen tetapi jarang muncul di seluruh korpus, sehingga menyoroti istilah-istilah yang relevan [7].

2.3 Pembagian Data (*Data Splitting*)

Dataset yang telah bersih dibagi menjadi dua, yaitu data pelatihan (*data train*) dan data pengujian (*data test*) dengan perbandingan 80:20, itu merupakan perbandingan umum yang digunakan dalam *machine learning.* Data pelatihan (*data train*) digunakan untuk mengajari model mengenal pola dalam dataset [9]. Data pengajian (*data test*) digunakan untuk mengevaluasi seberapa baik model dapat menggeneralisasi data baru yang belum pernah dilihat sebelumnya [9].

2.4 Penanganan *Imbalanced Data*

Pada penelitian ini mengalami masalah ketidakseimbangan kelas (*class imbalance*). Lowongan pekerjaan pada level senior lebih sedikit dibandingkan level junior, yang menyebabkan model *machine learning* standar mengalami bias dalam memprediksi kelas mayoritas (junior) dan gagal mendeteksi level senior. Gambar 2 dibawah ini menunjukkan ketidakseimbangan kelas yang terjadi.



Gambar 2 Distribusi Kelas Pengalaman Kerja

Distribusi kelas *Senior* sejumlah 161, kelas *Fresh Graduate* 992, dan kelas *Junior/Mid* 848. Terlihat sekali terjadi ketidakseimbangan kelas, dan ini merupakan tantangan signifikan dalam pemodelan prediktif [6]. Dalam mengatasi hal tersebut, digunakan pendekatan Level Data dengan SMOTE (*Synthetic Minority Over-sampling Technique*) dan pendekatan Level Algoritma dengan *Class Weighting* [6].

a. Pendekatan Level Data: SMOTE

   SMOTE (*Synthetic Minority Over-sampling Technique*) adalah salah satu metode *oversampling* yang paling populer. Metode ini bekerja dengan menambah data (*oversampling*) pada kelas minoritas. Tujuannya untuk meminimalkan ketimpangan data atau membuat jumlah data setara dengan kelas target lainnya [6].

b. Pendekatan Level Algoritma: *Class Weighting*

   Class Weighting ini memberikan bobot atau nilai yang berbeda di setiap kelas tergantung jumlah sampelnya. Kelas mayoritas (banyak) diberi bobot lebih kecil dan kelas minoritas (sedikit) diberi bobot atau hukuman lebih

Jurnal Dinamika Informatika
Volume 14, No 2, Oktober 2025
ISSN 1978-1660 : 73-84
ISSN *online* 2549-8517

besar. Jika model salah menebak, maka hukumannya berat, jadi model lebih hati-hati dalam mempelajari [10].

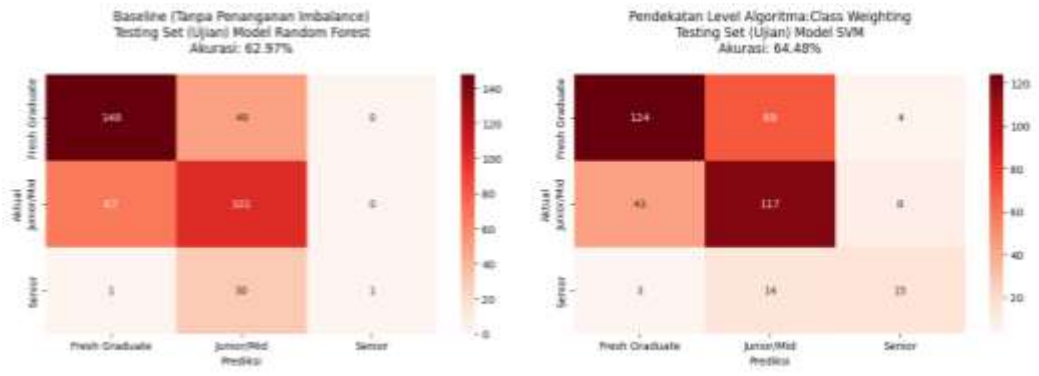
2.5 Evaluasi menggunakan *Confusion Matrix*

Confusion Matrix merupakan alat untuk mengevaluasi model pada *supervised learning* [11]. Fokus utama evaluasi kinerja model penelitian ini adalah pada kemampuannya untuk mengidentifikasi kelas minoritas yaitu *'Senior'*. Oleh karena itu, metrik utama yang digunakan adalah *recall* untuk kelas *'Senior'*, guna mengukur seberapa banyak dari total lowongan *'Senior'* yang sebenarnya berhasil diidentifikasi dengan benar oleh model. Selain itu, akurasi global juga diukur untuk melihat stabilitas dan kinerja model secara keseluruhan disemua kelas.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil kuantitatif dari perbandingan model yang dilakukan dan memberikan analisis mendalam tentang implikasi dari temuan penelitian, terutama terkait efektivitas masing-masing pendekatan dalam mengatasi masalah ketidakseimbangan dan ambiguitas data.

3.1 Perbandingan Kinerja Model



Kelas Senior (kanan bawah) hanya tertebak 1 *Kelas Senior (kanan bawah) berhasil ditebak lebih banyak (15)*

Gambar 3 Perbandingan *Confusion Matrix* pada *Data Test*

Hasil komparasi antara pendekatan Level Data dengan SMOTE (*Synthetic Minority Over-sampling Technique*) dan pendekatan Level Algoritma dengan *Class Weighting* pada *testing set* yang dirangkum pada Tabel 1 dibawah ini.

Tabel 1 Perbandingan Kinerja Model

Eksperimen	Algoritma	Akurasi Global	*Precision* (Senior)	*Recall* (Senior)	*Gap Overfitting*
Baseline (Tanpa Penanganan *Imbalance*)	SVM	65.24%	100%	19%	19.69%
	Random Forest	62.97%	100%	3%	36.97%
Algorithmic Level (Weighting Class)	**SVM**	64.48%	56%	**47%**	21.33%
	Random Forest	61.71%	61%	34%	30.35%
Data Level (SMOTE+nGram)	SVM	65.74%	69%	34%	28.62%
	Random Forest	62.72%	64%	28%	26.42%

Analisis Tabel 1 menunjukkan bahwa pendekatan *Algorithmic Level* pada SVM + *Class Weight* memberikan kinerja terbaik dalam mengenali kelas minoritas. Dengan nilai *Recall* 47%, model ini mampu mengidentifikasi hampir setengah dari seluruh lowongan 'Senior', yang merupakan peningkatan dibandingkan model SVM *Baseline* (*Recall* 19%). Kinerja ini secara signifikan juga mengungguli pendekatan *Data Level* yang menggunakan SMOTE, dimana kombinasi terbaik (SVM + SMOTE) hanya mencapai *Recall* 34%.

3.2 Analisis Kegagalan SMOTE pada Data Teks

Meskipun SMOTE secara teori dirancang untuk menyeimbangkan distribusi kelas, hasil penelitian ini menunjukkan bahwa penerapannya justru dapat

Jurnal Dinamika Informatika
Volume 14, No 2, Oktober 2025
ISSN 1978-1660 : 73-84
ISSN *online* 2549-8517

kontraproduktif pada data teks berdimensi tinggi. Terlihat bahwa penerapan SMOTE pada SVM menurun nilai *Recall* secara signifikan, yaitu dari 47% menjadi 34%. Temuan ini mendukung kritik dari Hassanat et al [2] yang menyatakan bahwa *oversampling* dapat menyuplai proses pembelajaran dengan instans yang "dipalsukan" dan tidak representatif.

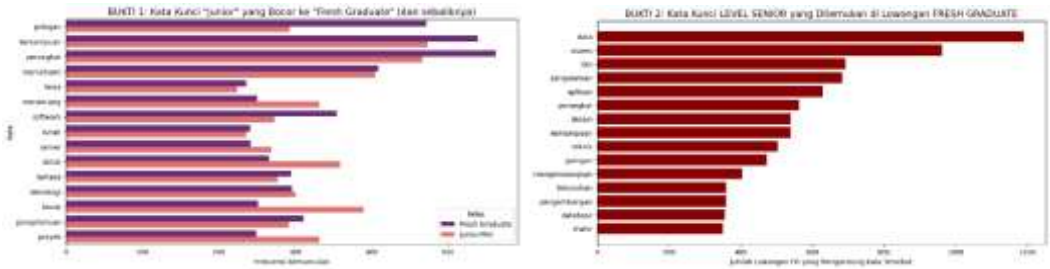
Dalam konteks teks lowongan pekerjaan, dimana ruang fitur sangat padat dan kompleks, proses interpolasi linear SMOTE menciptakan sampel sintesis yang kemungkinan besar jatuh di area yang ambigu. Sampel-sampel baru ini tidak membawa informasi semantik yang otentik, antara kelas '*Senior*' dan kelas lainnya seperti '*Junior/Mid'*. Akibatnya, kemampuan *classifier* untuk membedakan kelas-kelas ini justru menurun.

3.3 Dampak Ambiguitas Semantik Intrinsik

Salah satu temuan penting lainnya adalah hampir semua akurasi global tertahan disekitar angka 60-an di setiap eskperimen Tabel 1. Keterbatasan ini tampaknya bukan semata-mata disebabkan oleh kelemahan model, melainkan oleh ambiguitas intrinsik yang ada di dalam data itu sendiri.

Gambar 4 Visualisasi WordCloud pada Kelas Pengalaman yang Menunjukkan Dominasi Kata Kunci yang Identik

Gambar 4 menunjukkan adanya irisan leksikal (*lexical overlap*) yang sangat tinggi antar kelas. Kata kunci umum seperti "pengalaman", "sistem", "data", dan "tim" muncul secara dominan di deskripsi lowongan pekerjaan untuk semua level pengalaman, dari *Fresh Graduate,* hingga *Senior.* Hal ini menyebabkan model kesulitan menemukan fitur pembeda (*distinctive features*) yang unik disetiap kelas.



Gambar 5 Frekuensi kemunculan kata kunci disetiap Kelas

Kemudian ditemukan banyak inkonsistensi dalam data, seperti lowongan yang ditujukan untuk *Fresh Graduate* namun *skill* kompleks yang biasanya dimiliki oleh level *Junior.* Gambar 5 mengungkap anomali data dimana lowongan yang berlabel *Fresh Graduate* ternyata mengandung persyaratan *skill* kompleks yang lazimnya dimiliki oleh level 'Senior'. Kata kunci tingkat lanjut seperti "mahir", "mengembangkan", muncul ribuan kali pada deskripsi lowongan pemula. Fenomena "*Fresh Graduate* rasa *Senior*" ini menciptakan *noise* semantik yang signifikan.

Ambiguitas ini menginformasi bahwa kondisi dimana data asli memiliki *overlap* (memiliki kesamaan) dan batas kelas yang kabur menjadi penghambat utama bagi efektivitas metode *Data Level* seperti SMOTE [12]. SMOTE yang bekerja berdasarkan asumsi kedekatan fitur, akan kesulitan menciptakan sampel yang bermakna ketika batas-batas antar kelas secara inheren sudah kabur.

KESIMPULAN

Penelitian ini membandingkan dua strategi utama, yaitu pendekatan *Algorithmic Level* (*Class Weighting*) dan pendekatan *Data Level* (SMOTE) untuk menangani masalah data tidak seimbang dalam klasifikasi teks lowongan pekerjaan, dan ditemukan ambiguitas semantik yang tinggi. Pendekatan *Algorithmic Level* yang diimplementasikan melalui SVM dengan *Class Weighting* terbukti secara signifikan lebih unggul dalam mengidentifikasi kelas minoritas "Senior" dibandingkan dengan pendekatan *Data Level* menggunakan SMOTE.

SMOTE meskipun populer, namun pada penelitian ini menunjukkan kinerja yang kurang optimal untuk data teks berdimensi tinggi. SMOTE membuat sampel sintesis cenderung memperkenalkan *noise* yang mengaburkan batas keputusan antar kelas, terutama ketika terdapat tumpang tindih semantik yang tinggi pada data asli. Pada kasus data teks berdimensi tinggi dengan *overlap* semantik yang signifikan, memodifikasi bobot internal algoritma untuk memberikan perhatian lebih pada kelas minoritas merupakan strategi yang lebih *robust* dan aman. Pendekatan ini menghindarai risiko penambahan *noise* dan distorsi pada distribusi data yang melekat pada teknik *oversampling* sintesis.

SARAN

Saran untuk penelitian kedepan, eksplorasi teknik *cost-sensitive* lainnya pada arsitektur model yang lebih canggih, seperti model berbasis *Transformer,* dapat menjadi arah yang menjanjikan untuk menangani ambiguitas semantik secara lebih efektif. Sekaligus menambah jumlah dataset tidak hanya lowongan kerja bidang IT saja, agar semakin banyak data yang diolah dan mesin semakin belajar dari data yang banyak.

DAFTAR PUSTAKA

[1] T. Maciejewski and J. Stefanowski, “Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data,” pp. 104–111, 2011.

[2] A. B. Hassanat, A. S. Tarawneh, G. A. Altarawneh, and A. Almuhaimeed, “Stop Oversampling for Class Imbalance Learning: A Critical Review,” 2022.

[3] S. Mishra, “Handling Imbalanced Data : SMOTE vs . Random Undersampling,” pp. 317–320, 2017.

[4] S. Ndaba, “REVIEW OF CLASS IMBALANCE DATASET HANDLING TECHNIQUES FOR DEPRESSION,” vol. 12, no. 2, pp. 31–45, 2023, doi: 10.5121/ijci.2023.120203.

[5] A. M. Elsobky, A. El Keshk, and M. G. Malhat, "A Comparative Study for Different Resampling Techniques for Imbalanced datasets," vol. 10, pp. 147–156, 2023.

[6] W. Chaipanha and P. Kaewwichian, "SMOTE VS . RANDOM UNDERSAMPLING FOR IMBALANCED DATA- CAR OWNERSHIP DEMAND MODEL," vol. 24, no. 3, pp. 105–115, 2022.

[7] E. Qais and M. N. Veena, "TxtPrePro : Text Data Preprocessing using Streamlit Technique for Text Analytics Process," *2023 Int. Conf. Network, Multimed. Inf. Technol.*, pp. 1–6, 2023, doi: 10.1109/NMITCON58196.2023.10275887.

[8] N. Sofa, F. S. Utomo, and R. E. Saputro, "Eksplorasi Model Hybrid Transformer-Latent Semantic Analysis (LSA) Untuk Pemahaman Konteks Teks Berita Berbahasa Indonesia Fakultas Ilmu Komputer , Universitas Amikom Purwokerto , Indonesia Exploration of Hybrid Transformer Model-Latent Semantic Analysis (LSA) for Context Understanding of Indonesian News Texts," vol. 5, no. 5, pp. 1239–1252, 2025.

[9] Y. D. Suwito and Y. A. Susetyo, "Prediksi Kepuasan Pelanggan Maskapai menggunakan Model Machine Learning," vol. 15, pp. 354–367, 2026.

[10] S. Ağırlıklandırma, T. Kullanımı, M. Araştırma, and B. Bakirarar, "Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning : Methodological Research Makine Öğrenmesinde Dengesiz Sınıf Problemiyle Başa Çıkmak İçin," pp. 19–29, 2023, doi: 10.5336/biostatic.2022-93961.

[11] F. Smote, "Klasifikasi Ulasan Konsumen Menggunakan Random," vol. 5, no. 1, pp. 66–77, 2024.

[12] M. Imani and A. Beikmohammadi, "The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction".