**ANALISIS UNBALANCE DENGAN TEKNIK UNDERSAMPLING DAN OVERSAMPLING UNTUK AKURASI NAÏVE BAYES**

## **Nama Tidak Ditemukan**

UNIVERSITAS MUHAMMADIYAH MAKASSAR 2024

e-mail: xxxx@xxxx.xxx\*

***Abstract***

*RENY RAHAYU.S. Unbalance Analysis with Undersampling and Oversampling Techniques for Naïve Bayes Accuracy (supervised by Titin Wahyuni ​​S.Pd., M.T. and Fahrim Irhamna Rachman S.Kom., M.T.).*

*Dataset imbalance is a condition where the distribution of classes in a dataset is unbalanced, with the majority class having a much larger amount of data compared to the minority class, which can cause errors in classification results where minority class data is often classified as the majority class. The aim of this research is to understand the impact of data imbalance on the accuracy of the Naïve Bayes model. The Naïve Bayes model referred to in this research is a simple but effective probability-based classification algorithm, which is based on Bayes' Theorem. This algorithm assumes that the features in the dataset are independent of each other, an assumption that is rarely true in reality but often produces quite good results in practice. The research results show that balance in the dataset greatly influences the accuracy of the Naïve Bayes model. The undersampling technique produces the highest accuracy of 96.6%, while the oversampling technique achieves a maximum accuracy of 92%, and the use of SMOTE (Synthetic Minority Over-sampling Technique) provides good results with the highest accuracy of 91.8%. The conclusion of this research is that undersampling techniques tend to reduce accuracy by removing important information from the majority class, while oversampling techniques risk causing overfitting. Therefore, the selection of balancing methods must be done carefully, adapting to the nature of the data and analysis objectives, and carefully evaluated to achieve optimal model performance.*

*KATA PENGANTAR iv*

*DAFTAR ISI vi*

*DAFTAR GAMBAR viii*

*DAFTAR TABEL ix*

*Keywords: : Unbalance Analysis, Naïve Bayes, Oversampling, SMOTE, Undersampling.*

**Abstrak**

RENY RAHAYU.S. Analisis Unbalance dengan Teknik Undersampling dan Oversampling untuk Akurasi Naïve Bayes (dibimbing Titin Wahyuni S.Pd.,M.T dan Fahrim Irhamna Rachman S.Kom.,M.T.).

Ketidakseimbangan dataset adalah kondisi dimana distribusi kelas dalam dataset tidak seimbang, dengan kelas mayoritas memiliki jumlah data yang jauh lebih banyak dibandingkan dengan kelas minoritas, yang dapat menyebabkan kesalahan dalam hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagai kelas mayoritas. Tujuan dari penelitian ini adalah untuk memahami dampak ketidakseimbangan data terhadap akurasi model Naïve Bayes. Model Naïve Bayes yang dimaksud dalam penelitian ini adalah salah satu algoritma klasifikasi berbasis probabilitas yang sederhana namun efektif, yang didasarkan pada Teorema Bayes. Algoritma ini mengasumsikan bahwa fitur-fitur dalam dataset bersifat independen satu sama lain, sebuah asumsi yang jarang benar dalam kenyataan tetapi sering kali menghasilkan hasil yang cukup baik dalam praktik. Hasil penelitian menunjukkan bahwa keseimbangan pada dataset sangat mempengaruhi akurasi model Naïve Bayes. Teknik undersampling menghasilkan akurasi tertinggi sebesar 96,6%, sementara teknik oversampling mencapai akurasi maksimal 92%, dan penggunaan SMOTE (Synthetic Minority Over-sampling Technique) memberikan hasil yang baik dengan akurasi tertinggi 91,8%. Kesimpulan dari penelitian ini adalah bahwa teknik undersampling cenderung mengurangi akurasi dengan menghilangkan informasi penting dari kelas mayoritas, sedangkan teknik oversampling berisiko menyebabkan overfitting. Oleh karena itu, pemilihan metode penyeimbangan harus dilakukan dengan hati-hati, menyesuaikan sifat data dan tujuan analisis, serta dievaluasi dengan cermat untuk mencapai performa model yang optimal.

Kata Kunci: : Analisis Unbalance, Naïve Bayes, Oversampling, SMOTE, Undersampling.

**1. Pendahuluan**

Ketidakseimbangan dataset adalah keadaan dimana distribusi kelas didalam dataset tidak seimbang. Sebuah kelas dikatakan tidakseimbang apabila ada suatu kelas yang memiliki data yang lebih banyak dibandingkan dengan kelas lainnya. Kelompok kelas dengan jumlah data yang banyak disebut dengan kelas mayoritas, sedangkan kelompok kelas dengan jumlah yang sedikit disebut dengan kelas minoritas. Dataset yang tidak seimbang menyebabkan misleading atau kesesatan dalam hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagaikelas mayoritas.Penerapan algoritma klasifikasi tanpa memperhatikan keseimbangan kelas mengakibatkan prediksi yang baik bagi kelas mayoritas dan kelas minoritas diabaikan.Apabila algoritma klasifikasi di implementasikan langsung terhadap dataset yang imbalance maka akan mengalami penurunan performa (Sulistiyono et al., 2021).

Masalah Ketidakseimbangan ini bisa menjadi masalah dalam pembelajaran mesin karena model yang dibuat mungkin cenderung lebih baik dalam memprediksi kelas mayoritas, sementara kelas minoritas diabaikan (Technology, 2023). Dalam hal ini Masalah ketidakseimbangan kelas dalam dataset dapat menjadi kendala serius dalam pengembangan model. Model Machine Learning yang digunakan yaitu Algoritma Naïve bayes. Naïve Bayes merupakan pengklasifikasi probabilitas sederhana untuk menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data yang diberikan (Ericha Apriliyani & Salim, 2022).

Dalam penelitian ini, peneliti menggunakan Teknik Under Sampling dan Oversampling dalam menangani Unbalance atau Ketidakseimbangan dataset sebelum melatih model Naïve Bayes. Pada Teknik Under Sampling, data yang mayoritas dikurangi sehingga jumlahnya sama dengan kelas minoritas, sedangkan oversampling mengenerate data baru untuk kelas minoritas sehingga

jumlahnya seimbang dengan kelas mayoritas (Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti et al., 2017).

Berdasarkan pernyataan diatas penelitian ini diharapkan dapat memberikan solusi dalam mengatasi ketidakseimbangan kelas dalam dataset dengan menerapkan teknik under sampling dan over sampling guna meningkatkan akurasi model Naïve Bayes.

Berdasarkan pada latar belakang masalah di atas, maka rumusan masalah pada penelitian ini yaitu “Bagaimana ketidakseimbangan (unbalance) pada dataset mempengaruhi akurasi model Naïve Bayes?”

Berdasarkan pada latar belakang masalah, rumusan masalah di atas, maka Adapun tujuan yang ingin dicapai dari penelitian ini yaitu : Adalah untuk memahami ketidakseimbangan data pada akurasi model Naïve Bayes?

Berdasarkan pada uraian latar belakang , rumusan masalah dan tujuan penelitian di atas, maka manfaat penelitian ini dibagi menjadi bebera aspek sebagai berikut :

1. Aspek Teoritis

Memberikan pemahaman yang lebih mendalam tentang konsep dasar Naïve Bayes dalam konteks klasifikasi data..

1. Aspek Praktis

2. Bagi Peneliti

3. Dapat menghasilkan publikasi ilmiah yang dapat meningkatkan reputasi peneliti dalam komunitas akademis.

4. Akan memperluas pemahaman peneliti tentang analisis data, model klasifikasi, dan teknik penanganan ketidakseimbangan data.

5. Dapat memperluas jaringan profesionalnya dengan berkolaborasi dengan rekan peneliti lainnya yang tertarik dalam bidang analisis data dan machine learning.

6. Bagi Universitas

7. Universitas dapat membantu mengembangkan keterampilan dan pengetahuan mereka dalam bidang analisis data dan machine learning.

8. Memberikan kontribusi pada pengetahuan dan pemahaman dalam bidang analisis data, klasifikasi, dan penanganan ketidakseimbangan data, yang merupakan aset bagi universitas dalam mendukung pengajaran, penelitian, dan layanan masyarakat.

Dari analisis rumusan masalah di atas dapat dirumuskan beberapa batasan masalah yaitu :

1. Fokus pada Model Naïve Bayes

2. Penggunaan Teknik Under Sampling dan Oversampling

3. Evaluasi pada Akurasi Model, pada penelitian ini akan membatasi analisis pada pengukuran akurasi model Naïve Bayes sebelum dan setelah penerapan teknik under sampling dan oversampling.

4. Dataset, dalam penelitian ini akan menggunakan dataset tertentu yang relevan dengan tujuan analisis dan memiliki masalah ketidakseimbangan kelas yang signifikan.

5. Data yang digunakan diambil dari data mahasiswa Fakultas Teknik Prodi Pengairan Universitas Muhammadiyah Makassar, dan penelitian ini membutuhkan data mengenai mahasiswa, seperti Jenis Kelamin, IPK, SKS, Lama Studi dan Mata kuliah wajib maupun tambahan yang berkaitan dengan Tingkat kelulusan mahasiswa. Data ini diperoleh dari Simak Fakultas.

6. Tidak Memperhitungkan Teknik Lain.

7. Tidak Memperhitungkan Variasi dalam Teknik

Untuk memberikan gambaran umum dari seluruh penulisan ini, Adapun sistematika penulisan yaitu :

**2. Metode Penelitian**

**3. Hasil dan diskusi**

**4. Kesimpulan**

Kesimpulan 55

Saran 55

DAFTAR PUSTAKA 56

Bab ini atau bab terakhir akan memuat Kesimpulan isi dari kesulurahan uraian bab sebelumnya dan saran – saran dari hasil yang telah diperoleh serta yang diharapkan daoat bermanfaat dalam pengembangan selanjutnya

1. Kesimpulan

Ketidakseimbangan (unbalance) pada dataset sangat mempengaruhi akurasi model Naïve Bayes. Dengan pengujian dataset sebanyak 1.155 data menunjukkan bahwa teknik undersampling menghasilkan akurasi tertinggi sebesar 96,6%, sementara teknik oversampling mencapai akurasi maksimal 92%. Penggunaan SMOTE memberikan hasil yang baik dengan akurasi tertinggi 91,8%. Teknik undersampling cenderung mengurangi akurasi dengan menghilangkan informasi penting dari kelas mayoritas, sedangkan oversampling berisiko overfitting. Oleh karena itu, pemilihan metode penyeimbangan harus dilakukan dengan hati-hati, menyesuaikan sifat data dan tujuan analisis, serta dievaluasi dengan cermat untuk mencapai performa model yang optimal.

Dari hasil penelitian yang dilakukan dapat diperoleh saran yaitu, Untuk melakukan evaluasi berkala terhadap performa model setelah penerapan teknik penyeimbangan dan prapemrosesan agar model tetap konsisten dan relevan dengan perubahan data dan dinamika bisnis. Hal ini diharapkan dapat memberikan kontribusi lebih dalam dalam mengatasi ketidakseimbangan data pada model klasifikasi Naïve Bayes.

**Referensi**

[1] Choirunnisa, S. (2019). Metode Hibrida Oversampling dan Undersampling Untuk Menangani Ketidakseimbangan Data Kegagalan Aakademik Universitas XYZ.

[2] Ericha Apriliyani, & Salim, Y. (2022). Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset. Indonesian Journal of Data and Science, 3(2), 47–54. https://doi.org/10.56705/ijodas.v3i2.45

[3] Hari, E., Prastyo, A., Studi, P., & Wiratsongko, R. (2020). Implementasi Teknik Web Scraping Pada Situs Berita Menggunakan Metode Supervised learning Implementasi Web Scraping Pada Situs Berita Menggunakan Metode Supervised learning IGL Putra Eka Prismana. https://republika.co.id

[4] Hendrawan, I. N. R., Arya, I. M., Saputra, B., Ayu, G., Cahya, P., & Gede, I. (2021). Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes. 50–56. https://doi.org/10.30864/eksplora.v11i1.606

[5] Heranova, O. (2019). Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 3(3), 443–450. https://doi.org/10.29207/resti.v3i3.1275

[6] Kawani, G. P. (2019). Implementasi Naive Bayes. Journal of Informatics, Information System, Software Engineering and Applications (INISTA), 1(2), 73–81. https://doi.org/10.20895/inista.v1i2.73

[7] Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, S., Chamidah, N., Mega Santoni, M., Matondang, N., Ilmu Komputer, F., & Pembangunan Nasional Veteran Jakarta, U. (2017). Terakreditasi SINTA Peringkat 2 Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN). Masa Berlaku Mulai, 1(3), 635–641.

[8] Mutmainah, S. (2021). Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke. SNATi, 1(1), 10–16. https://doi.org/10.20885/snati.v1i1.2

[9] Ningsi, B. A., & Arofah, I. (2021).

[10] …………………………………………………………………………………………

[11] ………………………………………………………………………..

[12] http://ejurnal.binawakya.or.id/index.php/MBI Vol.15 No.10 Mei 2021 Open Journal Systems. 15(10), 5097–5104.

[13] Prihandari, R. C. (2022). Data Mining: Konsep Dan Apikasi Menggunakan Rapidminer (Series: Supervised Learning Dan Unsupervised Learning). 8. http://repository.uin-suska.ac.id/63073/1/REGITA CAHYANI PRIHANDARI.pdf

[14] Rahmatullah, S., Utami, E., Informatika, T., Teknik, M., Dian, S., Cendikia, C., Yogyakarta, U. A., Negara, J., Candimas, N., & Lampung, K. (2019). Jurnal Informasi Dan Komputer Vol : 7 No : 1 2019 PREDIKSI TINGKAT KELULUSAN TEPAT WAKTU DENGAN METODE NAÏVE BAYES DAN K-

[15] NEAREST NEIGHBOR Jurnal Informasi Dan Komputer Vol : 7 No : 1 2019. 7–16.

[16] Septiani, Y., Aribbe, E., & Diansyah, R. (2020). ANALISIS KUALITAS LAYANAN SISTEM INFORMASI AKADEMIK UNIVERSITAS ABDURRAB TERHADAP KEPUASAN PENGGUNA MENGGUNAKAN

[17] METODE SEVQUAL (Studi Kasus : Mahasiswa Universitas Abdurrab Pekanbaru). Jurnal Teknologi Dan Open Source, 3(1), 131–143. https://doi.org/10.36378/jtos.v3i1.560

[18] Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. Sistemasi, 10(2), 445. https://doi.org/10.32520/stmsi.v10i2.1303

[19] Sutriani, E., & Octaviani, R. (2019). Keabsahan data. INA-Rxiv, 1–22.

[20] Technology, I. (2023). Media 2023 ANALISIS KETIDAKSEIMBANGAN KELAS DALAM PENGEMBANGAN MODEL KLASIFIKASI. 602–610.

[21] Wicaksana, A., & Rachman, T. (2018). Implementasi Metode Pembelajaran Reading Aloud dalam Meningkatkan Keaktifan Siswa pada Mata Pelajaran Qur’an Hadist. Angewandte Chemie International Edition, 6(11), 951–952., 3(1), 10–27. https://medium.com/@arifwicaksanaa/pengertian-use-case- a7e576e1b6bf

[22] Yustira, N., Witarsyah, D., & ... (2021). Implementasi Algoritma NaÏve Bayes Classification Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu. EProceedings ….

[23] https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/ article/view/16721%0Ahttps://openlibrarypublications.telkomuniversity.ac.id

[24] /index.php/engineering/article/view/16721/16429