

Penggunaan Metode SMOTE pada Naïve Bayes Gaussian untuk Klasifikasi Mahasiswa Drop Out

Aliyah Kurniasih¹, Khaira Isyara²

Program Studi Ilmu Komputer

Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ

JL TB Simatupang, Cilandak Timur, Kec. Ps. Minggu, Kota Jakarta Selatan 12560

aliyah.kurniasih@esqbs.ac.id¹, khaira.isyara@students.esqbs.ac.id²

Abstrak. Lembaga institusi pendidikan di Indonesia saling berkompetisi untuk memberikan hasil terbaik dalam mencetak lulusan yang berkualitas. Dimana dalam prosesnya terdapat mahasiswa yang di *drop out*, dan ini merupakan satu indikator penting yang harus diperhatikan dalam proses pendidikan. Jika suatu sistem mampu memperkirakan atau memprediksi mahasiswa yang akan lulus tepat waktu, maka akan sangat membantu pihak kampus dalam mengambil tindakan pencegahan untuk mencegah terjadinya kasus *drop out*. Penelitian ini akan mengklasifikasikan kasus mahasiswa yang akan di *drop out* atau tidak menggunakan metode Naïve Bayes Gaussian dengan teknik *oversampling* SMOTE untuk mengatasi *imbalance class*. Nilai akurasi yang diperoleh dari model klasifikasi sebelum *oversampling* sebesar 0,84 atau 84%. Namun terdapat peningkatan akurasi sebesar 2% yang dihasilkan setelah melakukan *oversampling* SMOTE sebesar 0,86 atau 86%.

Kata Kunci: *drop out*, Naïve Bayes Gaussian, SMOTE, klasifikasi

1 Pendahuluan

Dalam dunia pendidikan Indonesia, baik Perguruan Tinggi Negeri (PTN) maupun Perguruan Tinggi Swasta (PTS) saling berkompetisi untuk memberikan hasil terbaik dalam mencetak lulusan yang berkualitas. Oleh karena itu, perguruan tinggi berusaha meningkatkan kualitas pendidikan dan memberikan pengalaman terbaik kepada mahasiswa sebagai penerima jasa mereka. Persaingan yang semakin ketat dalam mencari pekerjaan menuntut perguruan tinggi untuk menghasilkan sarjana yang berkualitas dan memiliki daya saing [1].

Salah satu indikator kegagalan dalam proses pendidikan di perguruan tinggi adalah jumlah mahasiswa yang *drop out* atau berhenti studi [2]. Keterlambatan atau ketidak tepatan waktu kelulusan beberapa mahasiswa dapat menjadi hambatan bagi perkembangan institusi pendidikan tinggi tersebut. Apabila sistem mampu mengantisipasi atau meramalkan mahasiswa yang akan menyelesaikan studi sesuai jadwal, ini akan secara signifikan mendukung upaya pihak kampus dalam mengambil langkah-langkah proaktif untuk mencegah terjadinya situasi di mana mahasiswa menghentikan studi [3]. Oleh karena itu, penelitian ini akan mengklasifikasikan kasus mahasiswa yang *drop out* dengan tujuan untuk mengidentifikasi mahasiswa yang berpotensi dikeluarkan berdasarkan faktor-faktornya.

Data mining atau penambangan data merupakan suatu proses untuk mengolah data dengan tujuan menggali atau mengambil inti dari pengetahuan dan kumpulan data hingga dapat diperoleh struktur atau susunan yang lebih mudah dipahami. Pada penelitian ini menggunakan metode penambangan data berupa klasifikasi. Klasifikasi ialah proses pengelompokan berdasarkan hubungan antara variabel fitur dan variabel target.

Dalam proses metode klasifikasi, terdapat beberapa algoritma yang dapat digunakan. Diantaranya dalam penelitian [4] yang menggunakan beberapa algoritma untuk klasifikasi data siswa untuk rekomendasi penjurusan, seperti C4.5, Naïve Bayes, KNN, Rule Induction, dan lain-lain dengan melakukan uji T-Test untuk mengetahui seberapa besar perbedaan antara algoritma-algoritma tersebut. Hasil dari analisis perbandingan dalam penelitian ini menunjukkan bahwa algoritma Naïve Bayes memberikan hasil terbaik daripada algoritma-algoritma lainnya, dengan tingkat akurasi sebesar 79,51% dan nilai AUC sebesar 0,861. Sedangkan pada penelitian [5] dilakukan perbandingan antara metode KNN dan Naïve Bayes Gaussian untuk klasifikasi penyakit stroke. Dalam penelitian ini, hasil yang diperoleh dari algoritma KNN menunjukkan akurasi sebesar 68,30%, presisi sebesar 67,20%, dan recall sebesar 73,34%. Sementara itu, algoritma Gaussian Naive Bayes menghasilkan akurasi sebesar 74,45%, presisi sebesar 74,01%, dan recall sebesar 75,71%. Dari perbandingan tersebut, terlihat bahwa terjadi peningkatan akurasi sebesar 6,15%, presisi sebesar 6,81%, dan recall sebesar

2,37% pada algoritma Gaussian Naive Bayes. Hal ini menunjukkan bahwa kinerja algoritma Naive Bayes Gaussian lebih baik.

Dalam penggunaan teknik *oversampling* SMOTE, telah dilakukan penelitian oleh [6] dengan melakukan perbandingan akurasi hasil klasifikasi performa rating iklan TV menggunakan metode ANN sebelum dan sesudah dilakukan *oversampling* menggunakan SMOTE. Hasil yang didapatkan yaitu metode ANN+SMOTE mencapai akurasi sebesar 87,06% dibandingkan ANN yang hanya sebesar 86,35%. Sedangkan dalam penelitian yang dilakukan oleh [7], SMOTE digunakan pada dataset *Credit Card Fraud* untuk mengatasi masalah data tersebut yaitu ketidakseimbangan kelas. Dengan menerapkan skema evaluasi 10-cross fold validation diperoleh hasil bahwa SMOTE menghasilkan peningkatan rata-rata G-Mean dari 53,4% ke 81,0% dan rata-rata F-Measure dari 38,7 ke 81,8%.

Dari beberapa penelitian tersebut dapat disimpulkan bahwa algoritma Naive Bayes Gaussian memiliki kinerja dan menghasilkan tingkat akurasi yang baik dalam melakukan klasifikasi. Dari beberapa penelitian tersebut, SMOTE juga menyelesaikan masalah ketidakseimbangan kelas sehingga memengaruhi tingkat akurasi yang didapat. Maka dari itu, akan dilakukan perbandingan antara penggunaan model Naive Bayes Gaussian dengan dan tanpa menggunakan teknik *oversampling* SMOTE dalam klasifikasi mahasiswa *drop out*. Tujuan dari penelitian ini yaitu untuk mendapatkan evaluasi yang akurat terhadap kedua model tersebut, sehingga hasilnya dapat menjadi panduan dalam pengambilan keputusan untuk mengurangi tingkat mahasiswa *drop out*. Serta dengan membandingkan performa kedua model, penelitian ini berharap dapat memberikan wawasan yang lebih baik tentang efektivitas penggunaan teknik *oversampling* dalam meningkatkan hasil klasifikasi mahasiswa *drop out*.

2 Tinjauan Pustaka

2.1 Drop out

Salah satu permasalahan pada perguruan tinggi yaitu *drop out*, *drop out* ialah situasi serius dimana beberapa mahasiswa memutuskan untuk tidak melanjutkan studi mereka. Hal ini merupakan masalah yang dialami oleh sebagian mahasiswa di perguruan tinggi [8].

2.2 Data Mining

Data mining adalah kegiatan atau proses pengumpulan data dalam skala besar dan ekstraksi informasi yang berguna dari data tersebut [9]. Data mining juga merupakan metode yang digunakan untuk mengatasi masalah dengan menganalisis objek yang ada dalam dataset. Proses ini dikenal juga sebagai *knowledge discovery in database* (KDD), yang melibatkan pengumpulan dan penggunaan data untuk mengidentifikasi pola atau hubungan yang terdapat dalam kumpulan data yang besar [5]. Hasil dari proses data mining dapat memberikan bantuan dalam pengambilan keputusan di masa depan.

2.3 Klasifikasi

Klasifikasi adalah salah satu metode data mining yang digunakan untuk mengalokasikan data rekaman baru ke salah satu kelas yang telah dipilih sebelumnya. Klasifikasi juga digunakan untuk memprediksi label untuk setiap kategori kelas. Selain itu, klasifikasi juga berfungsi untuk memodelkan data berdasarkan set data pelatihan dan nilai label kelas dalam atribut klasifikasi, yang kemudian digunakan untuk mengklasifikasikan data baru [10].

2.4 Naive Bayes Gaussian

Naive Bayes adalah salah satu algoritma klasifikasi yang menggunakan metode probabilitas dan statistik. Algoritma ini didasarkan pada Teorema Bayes yang dikemukakan oleh ilmuwan Thomas Bayes, yang memungkinkan prediksi peluang masa depan berdasarkan pengalaman sebelumnya. Dibandingkan dengan classifier lainnya, Naive Bayes memiliki kinerja yang baik dan tingkat akurasi yang tinggi. Dalam penggunaan algoritma Naive Bayes, nilai yang dipilih adalah yang memiliki persentase tertinggi, sehingga model dapat memberikan prediksi yang lebih akurat berdasarkan probabilitas yang diestimasikan [11]. Pada penelitian ini menggunakan Naive Bayes Gaussian karena data yang digunakan merupakan data kontinu. Seperti studi literatur yang digunakan, penggunaan algoritma Naive Bayes Gaussian mendapatkan hasil nilai akurasi yang baik.

$$P(X) = P(X|H)P(X)P(H)$$

Keterangan:

X : Data kelas yang belum diketahui

H : Hipotesis data kelas spesifik

$P(X)$: Probabilitas X

$P(H)$: Probabilitas hipotesis H (prior probabilitas)

$P(X|H)$: Probabilitas X kondisi hipotesis H

$P(H|X)$: Probabilitas hipotesis H kondisi X (posteriori probabilitas)

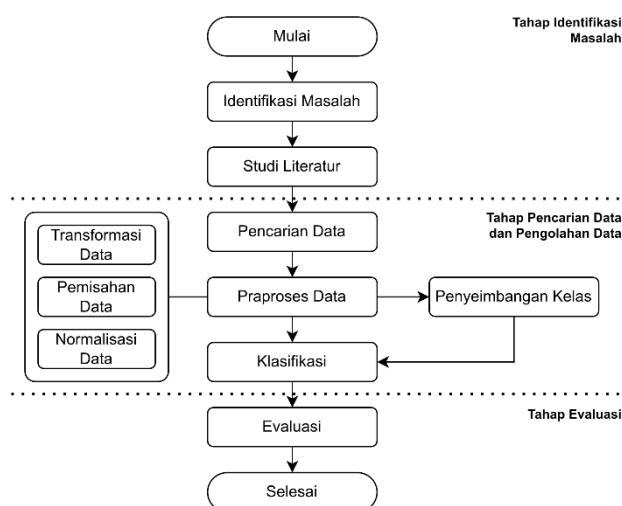
2.5 Imbalance Class

Ketidakseimbangan kelas (*imbalance class*) mengacu pada kondisi di mana jumlah sampel antara kelas-kelas yang berbeda dalam dataset tidak seimbang atau tidak proporsional. Ini berarti ada satu atau beberapa kelas yang memiliki jumlah sampel yang jauh lebih sedikit atau jauh lebih banyak daripada kelas lainnya. Masalah ketidakseimbangan kelas dapat mempengaruhi kinerja model pembelajaran mesin. Beberapa dampak yang mungkin terjadi adalah model dapat cenderung memprediksi secara dominan kelas mayoritas karena penyebaran yang tidak seimbang. Ini mengakibatkan kinerja model yang buruk dalam mengidentifikasi sampel dari kelas minoritas. Kemudian model mungkin tidak memiliki cukup sampel kelas minoritas untuk mempelajari pola dan karakteristiknya dengan baik [12]. Untuk mengatasi masalah ketidakseimbangan kelas, beberapa teknik dapat diterapkan, seperti *oversampling* yaitu dengan meningkatkan jumlah sampel pada kelas minoritas dengan cara menduplikasi atau menghasilkan sampel sintetis tambahan dari kelas minoritas.

2.6 SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) ialah salah satu metode yang banyak digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset. Metode ini menghasilkan sampel baru dari kelas minoritas dengan cara mensintesis data baru, sehingga mengimbangi dataset dengan melakukan *oversampling* pada kelas minoritas. *Oversampling* dalam metode SMOTE melibatkan pengambilan contoh dari kelas minoritas dalam dataset dan mencari tetangga terdekat untuk setiap contoh tersebut. Pendekatan ini membantu menghindari masalah *overfitting* yang berlebihan yang mungkin terjadi jika hanya mengulang-ulang contoh minoritas yang ada [13].

3 Metode Penelitian



Gambar 1. Alur Penelitian

3.1 Identifikasi Masalah

Proses identifikasi masalah dari penelitian ini mengangkat permasalahan yang ada di sebuah perguruan tinggi.

Pada tahap ini dilakukan pengamatan terhadap faktor-faktor apa saja yang memengaruhi mahasiswa yang berpotensi *drop out*. Rumusan masalah dari penelitian ini ialah mengenai bagaimana penggunaan *data mining* dalam melakukan klasifikasi menggunakan algoritma Naïve Bayes Gaussian dan SMOTE terhadap penyebab *drop out*-nya mahasiswa. Kemudian sebagai elemen pendukung dalam mengatasi tantangan dalam penelitian, digunakan sebagai basis referensi dengan menghimpun berbagai jurnal dan buku yang berkaitan dengan topik penelitian. Materi literatur penelitian dihimpun dari berbagai sumber termasuk buku, situs web, serta publikasi jurnal yang terkait dengan data mining atau klasifikasi dengan menggunakan algoritma Naïve Bayes Gaussian dan SMOTE.

3.2 Pencarian Data dan Pengolahan Data

Pada tahap ini dilakukan pencarian dan pengolahan data untuk penelitian. Data yang digunakan dalam penelitian ini adalah dataset yang bersifat publik, yang didapatkan dari situs Kaggle. Langkah selanjutnya yaitu tahap pengolahan data meliputi praproses data untuk menjadikan data menjadi suatu data yang lebih mudah untuk dibaca dan digunakan. Tahapan-tahapan pada praproses penelitian ini sebagai berikut.

1. Transformasi data, pada tahap ini dilakukan agar data yang digunakan sesuai dengan algoritma yang digunakan yaitu mengubah data kategorikal menjadi numerikal menggunakan Label Encoder.
2. Pemisahan data, tahap ini dilakukan dengan membagi data menjadi dua bagian yaitu data latih dan data uji menggunakan *train-test-split*.
3. Normalisasi data, proses ini mengubah skala data menjadi rentang yang konsisten atau standar menggunakan *MinMaxScaler*. Tujuan normalisasi adalah untuk menghilangkan perbedaan skala antara fitur-fitur dalam dataset.

Setelah praproses data, dilakukan pengolahan data dengan menggunakan algoritma Naïve Bayes Gaussian dan teknik *oversampling*, yaitu SMOTE. Kemudian di proses klasifikasi ini dilakukan dengan dua data yang berbeda, yang pertama dengan menggunakan data yang belum seimbang, dan yang kedua diterapkan SMOTE yaitu *oversampling* untuk mengatasi ketidakseimbangan kelas pada data.

3.3 Evaluasi

Setelah model dilakukan *training* menggunakan algoritma model klasifikasi Naïve Bayes Gaussian pada data yang telah di SMOTE dan tidak di SMOTE, model tersebut di evaluasi dengan menggunakan *performance* metriks evaluasi model yaitu *confusion matrix*, akurasi, *precision*, *recall* dan *f1-score* untuk menyatakan dalam menentukan model yang baik pada penelitian ini.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall}$$

Keterangan:

- TP : True Positive
 TN : True Negative
 FP : False Positive
 FN : False Negative

4 Hasil dan Pembahasan

4.1 Data

Tahap awal pada penelitian ini adalah pencarian data klasifikasi mahasiswa *drop out*. Data yang digunakan adalah dataset yang bersifat publik, yang didapatkan dari situs Kaggle [14]. Dataset yang digunakan berjumlah

318 data yang mempunyai 8 variabel yang terdiri dari 7 kelas fitur dan 1 kelas target. Kelas fitur yang terdapat dalam dataset yaitu Gender, Age, SleepingHours, Music, GamingHours, Distance, AverageScore. Kelas target pada data ini yaitu Dropout berada pada hasil klasifikasi yaitu “Yes” dan “No”. Dari hasil 318 data terdapat hasil klasifikasi “Yes” sebanyak 91 orang dan “No” sebanyak 227 orang, hal ini menunjukkan bahwa adanya ketidakseimbangan kelas. Berikut ini merupakan keterangan dari variabel yang ada pada data.

Tabel 1. Data

No	Variabel	Keterangan	Jenis	Nilai
Kelas Fitur				
1	Gender	Jenis kelamin mahasiswa	Kategorika 1	1. Female 2. Male
2	Age	Umur mahasiswa	Numerik	18-35
3	SleepingHours	Jam tidur mahasiswa	Numerik	4-9
4	Music	Kesukaan mahasiswa terhadap musik	Kategorika 1	1. Yes 2. No
5	GamingHours	Jam bermain mahasiswa	Numerik	0-4
6	Distance	Jarak dari rumah ke kampus yang ditempuh mahasiswa	Numerik	1820-96450
7	AverageScore	Nilai rata-rata yang dimiliki mahasiswa	Numerik	6.3-9.7
Kelas Target				
8	Dropout	Hasil klasifikasi	Kategorika 1	1. No 2. Yes

4.2 Praproses Data

Praproses data merupakan langkah yang diperlukan untuk menghapus gangguan agar menghasilkan data yang siap digunakan dalam proses klasifikasi. Dalam penelitian ini, langkah-langkah praproses data meliputi transformasi data, pembagian data, dan normalisasi data.

4.2.1 Transformasi Data

Tahap ini variabel fitur yang berjenis kategorikal akan diubah menjadi numerik. Variabel yang berjenis kategorikal yaitu Gender, Music, dan Dropout akan dilakukan proses inialisasi ke dalam bentuk numerik (integer). Atribut Gender mengubah label ‘F’ menjadi 0 dan label ‘M’ menjadi 1, sedangkan atribut Music dan Dropout mengubah label ‘No’ menjadi 0 dan label ‘Yes’ menjadi 1. Berikut tabel perbandingan 5 data teratas antara data sebelum *encoding* dan data setelah *encoding* menggunakan Label Encoder.

Tabel 1. Tabel perbandingan transformasi data

No	Gender	Music	Dropout
Data sebelum <i>encoding</i>			
1	F	Yes	No
2	M	No	No
3	M	Yes	Yes
4	F	Yes	No
5	M	No	Yes

Data sebelum <i>encoding</i>			
1	0	1	0
2	1	0	0
3	1	1	1
4	0	1	0
5	1	0	1

4.2.2 Pembagian Data

Proses ini membagi dataset menjadi subset yang berbeda untuk digunakan dalam tahap pelatihan (*training*) dan pengujian (*testing*) pada pembangunan model. Pada teknik ini, dataset asli dibagi menjadi dua subset yang saling terpisah, yaitu data latih (x_{train} dan y_{train}) yang digunakan untuk melatih model, dan data uji (x_{test} dan y_{test}) yang digunakan untuk menguji kinerja model yang telah dilatih. Dalam kode tersebut, data latih dan data uji akan dibagi dengan perbandingan 80:20 dari total data dengan hasil pembagian data latih sebanyak 115 data dan data uji sebanyak 29 data.

4.2.3 Normalisasi Data

Pada tahap normalisasi data dilakukan proses mengubah skala data menjadi rentang yang konsisten atau standar. Penggunaan teknik normalisasi *MinMaxScaler* mengganti nilai-nilai dalam dataset ke rentang tertentu, biasanya antara 0 dan 1. Dengan melakukan normalisasi ini, data akan memiliki rentang nilai yang terbatas dan terdefinisi secara konsisten.

4.3 Imbalance Class

Pada tahap ini digunakan teknik *oversampling*, yaitu SMOTE. SMOTE bekerja dengan cara menciptakan sampel sintetis baru dari kelas minoritas dengan melakukan interpolasi linier antara sampel-sampel yang ada dalam kelas minoritas. Berikut hasil penyeimbangan kelas dengan tabel perbedaan jumlah data latih sebelum dan sesudah dilakukan *oversampling*.

Tabel 2. Tabel perbandingan jumlah data latih sebelum dan sesudah *oversampling*

Jenis Data	0 (No)	1 (Yes)	Total
Data sebelum SMOTE	73	42	115
Data setelah SMOTE	73	73	146

4.4 Klasifikasi

Pemodelan klasifikasi dilakukan dengan menggunakan library dari *scikit learn* yaitu *Naïve Bayes Gaussian*. Lalu data latih akan dimasukkan pada fungsi *GaussianNB*. Setelahnya akan dilakukan hasil prediksi dari data uji untuk menentukan probabilitas hasil prediksi. Tahap klasifikasi ini dilakukan dengan dua data yang berbeda, data yang belum seimbang dan data yang sudah seimbang dengan menggunakan teknik SMOTE.

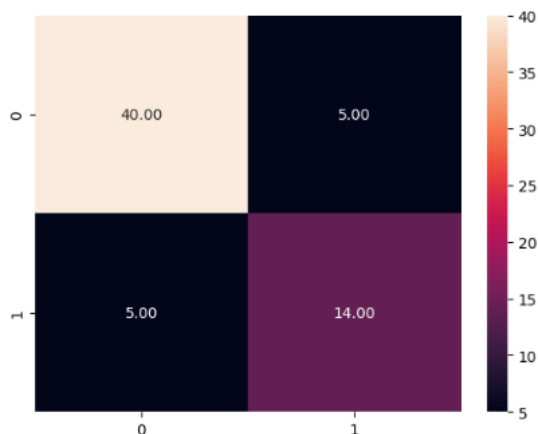
4.5 Hasil Evaluasi

Sesudah melalui proses klasifikasi, dilakukan evaluasi untuk mengetahui bagaimana performa model yang telah dilakukan, yaitu dengan cara membandingkan antara data yang belum seimbang dan data yang sudah seimbang

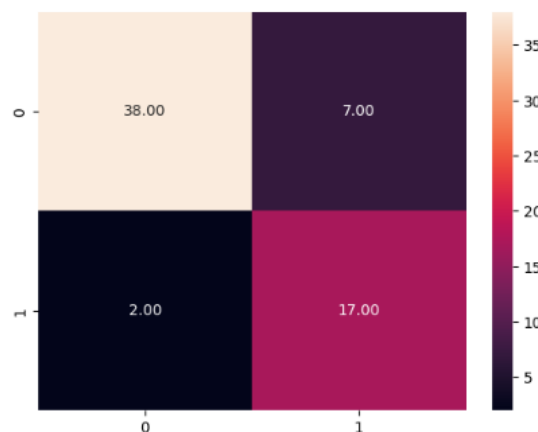
dengan teknik SMOTE. Berikut adalah hasil perbandingan akurasi presisi, sensitivitas/*recall*, dan *f1-score* dari pengujian terhadap kedua data tersebut menggunakan metode *confusion matrix*.

Tabel 3. Hasil evaluasi model GaussianNB dengan dan tanpa SMOTE

Model	Accuracy	Precision	Recall	F1-score
GaussianNB tanpa SMOTE	84%	0,81	0,81	0,81
GaussianNB dengan SMOTE	86%	0.83	0,87	0,84



Gambar 2. Hasil *confusion matrix* GaussianNB tanpa SMOTE



Gambar 3. Hasil *confusion matrix* GaussianNB dengan SMOTE

Dari tabel 3, hasil yang diperoleh menunjukkan nilai-nilai yang baik. Terlihat bahwa nilai akurasi yang dihasilkan dari klasifikasi yang dilakukan sebelum *oversampling* sebesar 0,84 atau 84%. Hal ini menunjukkan bahwa terdapat 84% jumlah tuple dalam data uji yang diklasifikasikan dengan benar oleh *classifier*. Namun terdapat peningkatan akurasi yang dihasilkan oleh klasifikasi yang dilakukan setelah *oversampling* sebesar 0,86 atau 86%.

Nilai presisi pada klasifikasi sebelum dan sesudah dilakukan *oversampling* mendapatkan hasil sebesar 81% dan 83%, yaitu terdapat sekian tuple yang dilabeli sebagai positif benar pada hasil prediksi yang dilakukan. Lalu didapatkan nilai *recall* sebesar 81% dan 87%. Nilai *f1-score* yang dihasilkan sebelum *oversampling* sebesar 81%. Namun setelah *oversampling* mendapat nilai *f1-score* sebesar 84%, 89% untuk kelas “No” dan 79% untuk kelas “Yes”.

5 Kesimpulan dan Saran

Berdasarkan hasil evaluasi yang diberikan, terlihat bahwa terdapat peningkatan nilai akurasi setelah penerapan *oversampling* menggunakan SMOTE. Sebelum *oversampling*, model klasifikasi memiliki akurasi sebesar 84%, sedangkan setelah *oversampling*, akurasi meningkat menjadi 86%. Hal ini menunjukkan bahwa terdapat peningkatan besarnya nilai akurasi setelah penerapan *oversampling* menggunakan SMOTE sebesar 2%.

Dengan nilai akurasi dan metrik evaluasi lain yang meningkat, dapat disimpulkan bahwa penggunaan *oversampling* dengan SMOTE berhasil meningkatkan kemampuan model dalam mengklasifikasikan dengan benar jumlah tuple dalam data uji. Hal ini menunjukkan bahwa penggunaan SMOTE efektif dalam mengatasi masalah ketidakseimbangan kelas dan meningkatkan kinerja model dalam mengenali dan memprediksi sampel-sampel dari kelas minoritas.

Saran untuk penelitian berikutnya yaitu menggunakan metode Naïve Bayes lainnya. Untuk mengatasi ketidakseimbangan kelas yang ada, bisa menggunakan teknik selain *oversampling* seperti *undersampling* atau kombinasi antara keduanya

Referensi

- [1] A. Budiyantra, Irwansyah, E. Prengki, P. Ahmad Pratama, and N. Wiliani, "Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu," *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, vol. 5, pp. 265–270, 2020, [Online]. Available: www.bri-institute.ac.id
- [2] I. H. Wijaya, "Prediksi Mahasiswa Drop Out Berdasarkan Klasifikasi Administratif," vol. 2, 2018.
- [3] Salmawati, Yuyun, and Hazriani, "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma Naive Bayes Dan Decision Tree," vol. 8, no. 2, 2022, [Online]. Available: <http://ejournal.fikom-unasman.ac.id>
- [4] A. R. Kadafi, "Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA," *Jurnal ELTIKOM*, vol. 2, no. 2, pp. 67–77, Dec. 2018, doi: 10.31961/eltikom.v2i2.86.
- [5] D. Ulfatul, M. Rachmad, H. Oktavianto, and M. Rahman, "Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke," *Jurnal Smart Teknologi*, vol. 3, no. 4, pp. 2774–1702, 2022, [Online]. Available: <http://jurnal.unmuhjember.ac.id/index.php/JST>
- [6] E. Sutoyo and M. Asri Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 6, pp. 379–385, 2020.
- [7] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor," *Jurnal ISD*, vol. 3, no. 1, pp. 2528–5114, 2018.
- [8] F. Imran, B. Susetyo, and A. H. Wigena, "Identifikasi Faktor-Faktor Yang Berhubungan Dengan Mahasiswa Putus Kuliah Di Ipb Angkatan 2008 Menggunakan Analisis Survival," 2013.
- [9] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Citec Journal*, vol. 2, pp. 207–217, 2015.
- [10] S. A. Nulhaqim, R. D. Heryadi, A. Pancasilawan, and M. Fedransyah, "Peranan Perguruan Tinggi Dalam Meningkatkan Kualitas Pendidikan Di Indonesia Untuk Menghadapi Asean Community 2015," *Social Work Journal*, vol. 6, pp. 154–272, 2015.
- [11] N. H. Alfianty and S. Mulyati, "Penerapan Naïve Bayes untuk Klasifikasi Data Penyakit Pada Anak."
- [12] N. Sulistiyowati and M. Jajuli, "Integrasi Naïve Bayes Dengan Teknik Sampling Smote Untuk Menangani Data Tidak Seimbang," vol. 14, no. 1, 2020, [Online]. Available: <https://journal.uniku.ac.id/index.php/ilkom>
- [13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [14] "Factors Contributing to Student Dropout," *Kaggle*, Mar. 2023. <https://www.kaggle.com/datasets/zaimeusoff/factors-contributing-to-student-dropout> (accessed Jun. 11, 2023).