

Penerapan Teknik Penanganan Kelas Tidak Seimbang untuk Meningkatkan Performa Model Naive Bayes

Elvi Rahmi¹, Eva Yumami², Natasya Muliani³

Politeknik Negeri Bengkalis

elvirahmi@polbeng.ac.id¹, evayumami@polbeng.ac.id², natasyamuliani1@gmail.com³

Abstract

Choosing a major is a crucial decision in a student's life, often influenced by factors such as interests, aptitude, and external influences like parental or teacher advice. In Indonesia, high school students typically choose between Science (IPA) and Social Sciences (IPS) tracks, which significantly impacts their university major choices. Previous studies have shown promising results in predicting students' major choices using various machine learning algorithms, with some achieving 100% accuracy. However, such high accuracy rates should be interpreted cautiously due to potential overfitting issues. Moreover, these studies often overlook class imbalance, particularly between the IPA and IPS tracks. This imbalance can lead to biased models that favor the majority class.

This study evaluates the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) in addressing class imbalance in predicting high school students' major choices. The results indicate that SMOTE does not significantly improve the performance of a Naive Bayes classifier on our dataset. Both models with and without SMOTE achieved 100% accuracy, F1-score, and recall, suggesting that other factors may be more influential in this context.

Classification, Imbalance, Oversampling, SMOTE, Naive Bayes

1. PENDAHULUAN

Pemilihan jurusan merupakan salah satu keputusan penting dalam kehidupan siswa, yang seringkali dipengaruhi oleh minat, bakat, dan faktor eksternal seperti saran orang tua atau guru. Di Indonesia, siswa umumnya memilih antara jurusan Ilmu Pengetahuan Alam (IPA) atau Ilmu Pengetahuan Sosial (IPS) pada jenjang pendidikan menengah atas. Keputusan ini akan sangat menentukan pilihan jurusan di perguruan tinggi.

Penelitian sebelumnya yang mengkaji prediksi pemilihan jurusan siswa telah menunjukkan hasil yang menjanjikan. Beberapa penelitian melaporkan tingkat akurasi hingga 100% dalam memprediksi pilihan jurusan siswa menggunakan algoritma tertentu, seperti Naive Bayes. Namun, hasil akurasi yang sangat tinggi ini perlu diinterpretasikan dengan hati-hati. Kemungkinan besar, model-model tersebut mengalami overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data pelatihan sehingga kurang mampu menggeneralisasi pada data baru. Selain itu, penelitian-penelitian tersebut seringkali mengabaikan adanya ketidakseimbangan kelas dalam data, terutama terkait dengan pilihan jurusan IPA atau IPS. Ketidakseimbangan kelas dapat menyebabkan model cenderung memprediksi jurusan yang lebih populer, seperti IPA, sehingga mengabaikan preferensi siswa yang memilih jurusan IPS.

Penelitian ini bertujuan untuk mengatasi keterbatasan penelitian sebelumnya dengan fokus pada upaya mengevaluasi efektivitas berbagai teknik penanganan ketidakseimbangan kelas dalam meningkatkan akurasi dan generalisasi model prediksi pemilihan jurusan. Dengan

demikian, diharapkan penelitian ini dapat memberikan kontribusi yang lebih baik dalam pengembangan sistem pendukung keputusan untuk pemilihan jurusan siswa.

2. TINJAUAN PUSTAKA

Beberapa kajian pada artikel hasil penelitian terdahulu yang menjadi rujukan dalam melakukan penelitian ini diantaranya artikel oleh Juan Fakhri dkk (2023) dengan judul “Perancangan Klasifikasi Algoritma Naive Bayes Pada Data Pemilihan Jurusan Siswa”. Penelitian ini berfokus pada pengembangan algoritma Naive Bayes untuk keperluan klasifikasi dalam konteks memprediksi pemilihan jurusan siswa. Mengevaluasi dampak ketidakseimbangan kelas terhadap kinerja model klasifikasi, serta menemukan kelebihan sampel data secara tidak sengaja, melakukan eksperimen dengan teknik SMOTE untuk mengatasi ketidakseimbangan kelas tersebut. Data pemilihan jurusan siswa dari SMA Negeri 2 Cikarang Selatan digunakan untuk mengidentifikasi ketidakseimbangan kelas. Penelitian ini membandingkan hasil klasifikasi sebelum dan setelah penerapan teknik SMOTE. Hasil analisis menunjukkan bahwa sebelum menggunakan teknik SMOTE, terdapat ketidakseimbangan kelas yang signifikan dalam data pemilihan jurusan siswa. Ketidakseimbangan ini memiliki dampak negatif terhadap kinerja model klasifikasi, terutama dalam mengenali kelas minoritas. Namun, setelah penerapan teknik SMOTE, ketidakseimbangan kelas berhasil dikurangi dan kinerja model klasifikasi mengalami peningkatan yang signifikan. Sebelum penerapan teknik SMOTE, akurasi model klasifikasi adalah 85,19% dan F1 Score adalah 33,3%. Namun, setelah penerapan teknik SMOTE, akurasi meningkat menjadi 85,71% dan F1 Score meningkat menjadi 92%. Dengan demikian, dapat disimpulkan bahwa penerapan teknik SMOTE berhasil mengurangi ketidakseimbangan kelas dan meningkatkan kinerja model klasifikasi dalam memprediksi pemilihan jurusan siswa. Metode Naive Bayes dapat digunakan sebagai alternatif yang efektif dalam memprediksi penjurusan siswa di SMA Negeri 2 Cikarang Selatan setelah menerapkan teknik SMOTE.

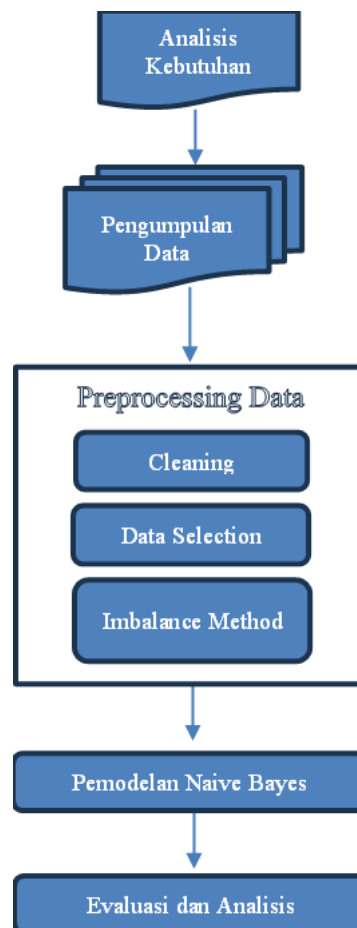
Penelitian selanjutnya oleh Azmi Aulia Rahman dkk (2023) membahas tentang analisis kinerja metode ketidakseimbangan data pada peningkatan akurasi klasifikasi metode hibrida machine learning. Penelitian ini menganalisis kinerja metode hibrid dalam meningkatkan akurasi pada data yang tidak seimbang menggunakan Data Kasus Demam Berdarah Dengue dari tahun 2017 hingga 2021 di Kota Bandung. Atribut yang digunakan dalam studi ini terdiri dari Total Populasi, Total Laki-laki, Kelulusan SD, Kelulusan SMP, Kelulusan SMA, Kelulusan Kuliah, Curah Hujan, Suhu Rata-rata, Kelembaban, Kasus Laki-laki, Jumlah Kasus, dan Kelas. Penelitian ini menggabungkan lima metode Pembelajaran Mesin, seperti Pohon Keputusan, Mesin Vektor Pendukung, Jaringan Saraf Tiruan, K-Tetangga Terdekat, dan Naive Bayes. Metode Hibrida yang digunakan dalam penelitian ini adalah metode Voting dan Stacking. Metode oversampling yang digunakan untuk menangani data yang tidak seimbang dalam penelitian ini adalah Oversampling Acak dan Adasyn. Hasil penelitian menunjukkan bahwa Voting dan Stacking tanpa Oversampling Acak dan Adasyn mendapatkan akurasi yang sama, yaitu 88,88%. Sementara menggunakan Oversampling Acak, voting mendapatkan akurasi sebesar 95,37% dan stacking mendapatkan akurasi sebesar 96,29%. Sementara menggunakan Adasyn, voting mendapatkan akurasi sebesar 94,44% dan stacking mendapatkan akurasi sebesar 97,22%. Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa Metode Oversampling Acak dan Adasyn dapat meningkatkan kinerja metode hibrid Pembelajaran Mesin pada data yang tidak seimbang. Kontribusi dari penelitian ini adalah memberikan informasi tentang studi dan analisis implementasi Metode Oversampling Acak dan Adasyn dalam meningkatkan kinerja metode Voting dan Stacking dalam klasifikasi hibrida.

Selanjutnya artikel oleh Hairani yang membahas tentang peningkatan kinerja metode Random Forest pada klasifikasi data diabetes yang tidak seimbang menggunakan SMOTE-TOMEK link. Penelitian ini bertujuan untuk penelitian ini bertujuan untuk menerapkan metode Smote-Tomeklink dan Random Forest dalam klasifikasi diabetes. Metodologi penelitian yang

digunakan adalah pengumpulan data diabetes yang diperoleh dari Kaggle sebanyak 768 data dengan 8 atribut input dan 1 atribut output sebagai kelas, pra-pemrosesan data menggunakan Smote-Tomeklink untuk menyeimbangkan dataset, klasifikasi menggunakan metode random forest, dan evaluasi kinerja berdasarkan akurasi, sensitivitas, presisi, dan skor F1. Berdasarkan pengujian yang dilakukan dengan membagi data menggunakan validasi silang 10-kali lipat, algoritma Random Forest dengan Smote-TomekLink mendapatkan akurasi, sensitivitas, presisi, dan skor F1 tertinggi dibandingkan dengan Random Forest dengan Smote. Algoritma Random Forest dengan Smote-Tomeklink memiliki akurasi 86,4%, sensitivitas 88,2%, presisi 82,3%, dan skor F1 85,1%. Dengan demikian, penggunaan SMOTE-Tomeklink dapat meningkatkan kinerja metode random forest berdasarkan akurasi, sensitivitas, presisi, dan skor F1.

3. METODE PENELITIAN

Penelitian ini dilakukan untuk mengevaluasi efektivitas penerapan teknik penanganan ketidakseimbangan kelas dalam meningkatkan performa model Naive Bayes dalam memprediksi pemilihan jurusan siswa SMA dan mengidentifikasi kontribusi teknik penanganan ketidakseimbangan kelas terhadap akurasi dan keandalan prediksi pemilihan jurusan siswa SMA. Untuk memenuhi tujuan tersebut, penelitian ini secara keseluruhan dibagi menjadi lima tahapan.



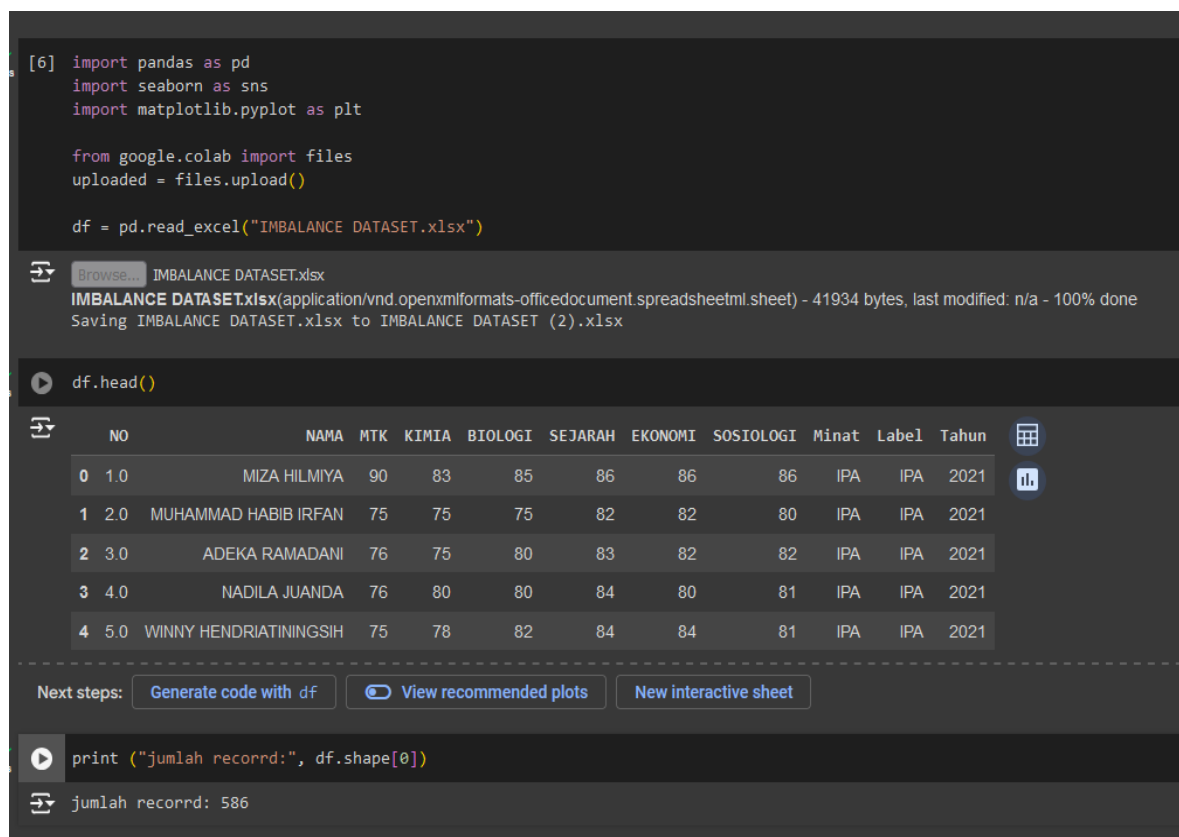
Gambar 1. Tahapan Penelitian

Analisis Kebutuhan

Tahapan penelitian dimulai dengan melakukan analisis kebutuhan akan penelitian yang dilakukan. Fokus penelitian ini adalah untuk meningkatkan akurasi dalam memprediksi pemilihan jurusan siswa SMA dan memahami bagaimana teknik penanganan kelas tidak seimbang dapat memperbaiki performa model Naive Bayes dalam konteks pemilihan jurusan siswa.

Pengumpulan Data

Penelitian ini menggunakan data peminatan siswa MAN 1 Bengkalis, yang mencakup informasi nilai akademis mata pelajaran Matematika, Kimia, Biologi, Sejarah, Ekonomi, dan Sosiologi, serta pilihan jurusan siswa, apakah IPA atau IPS. Data ini digunakan oleh pihak sekolah untuk menentukan siswa yang bersangkutan akan ditempatkan di jurusan IPA atau IPS. Dataset yang digunakan mencakup informasi dari periode tiga tahun terakhir dengan total 586 instances dan 7 atribut.



```
[6] import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from google.colab import files
uploaded = files.upload()

df = pd.read_excel("IMBALANCE DATASET.xlsx")
```

Browse... IMBALANCE DATASET.xlsx
IMBALANCE DATASET.xlsx(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 41934 bytes, last modified: n/a - 100% done
Saving IMBALANCE DATASET.xlsx to IMBALANCE DATASET (2).xlsx

df.head()

	NO	NAMA	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI	Minat	Label	Tahun
0	1.0	MIZA HILMIYA	90	83	85	86	86	86	IPA	IPA	2021
1	2.0	MUHAMMAD HABIB IRFAN	75	75	75	82	82	80	IPA	IPA	2021
2	3.0	ADEKA RAMADANI	76	75	80	83	82	82	IPA	IPA	2021
3	4.0	NADILA JUANDA	76	80	80	84	80	81	IPA	IPA	2021
4	5.0	WINNY HENDRIATININGSIH	75	78	82	84	84	81	IPA	IPA	2021

Next steps: [Generate code with df](#) [View recommended plots](#) [New interactive sheet](#)

```
print ("jumlah record:", df.shape[0])
```

jumlah record: 586

Gambar 2. Dataset Penelitian

Data ini kemudian dibagi menjadi dua, yakni Data Latih dan Data Uji. Untuk Data Latih, data yang digunakan adalah data peminatan siswa periode 2021-2022 dan 2022-2023. Sedangkan untuk Data Uji digunakan data peminatan siswa periode 2023-2024. Jumlah instance untuk masing-masing periode, deskripsi dari masing-masing atribut, dan data sampel yang digunakan dapat terlihat pada Tabel 1, Tabel 2, dan Tabel 3 berikut ini.

Tabel 1. Data Peminatan Siswa IPA dan IPS 2021-2023

No	Nama	Jumlah
1	IPA 2021-2022	186
2	IPS 2021-2022	58
3	IPA 2022-2023	99
4	IPS 2022-2023	40
5	IPA 2023-2024	65
6	IPS 2023-2024	138
Jumlah		586

Tabel 1. Deskripsi Atribut dari Dataset

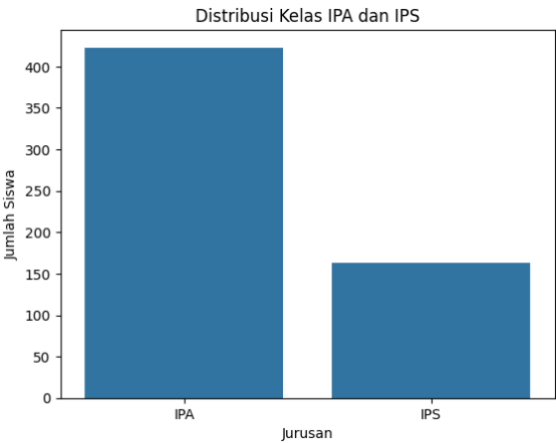
No	Atribut	Deskripsi	Label
1	Matematika	Nilai rapor Matematika semester sebelumnya	X1
2	Kimia	Nilai rapor Kimia semester sebelumnya	X2
3	Biologi	Nilai rapor Biologi semester sebelumnya	X3
4	Sejarah	Nilai rapor Sejarah semester sebelumnya	X4
5	Ekonomi	Nilai rapor Ekonomi semester sebelumnya	X5
6	Sosiologi	Nilai rapor Sosiologi semester sebelumnya	X6
7	Minat	Jurusan Pilihan Siswa IPA = 0 IPS = 1	X7
8	Jurusan	IPA = 0 IPS = 1	Y

Tabel 2. Sampel Dataset

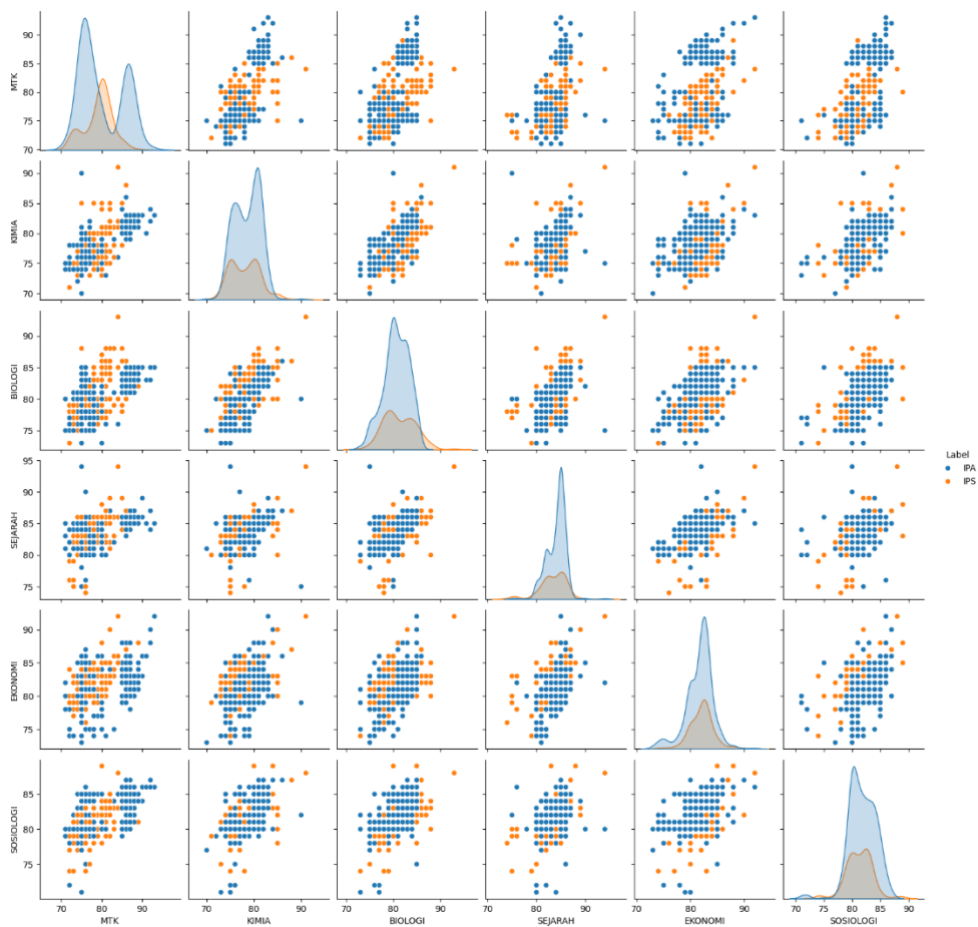
No	X1	X2	X3	X4	X5	X6	X7	Y
1	90	83	85	86	86	86	0	0
2	75	75	75	82	82	80	0	0
3	76	75	80	83	82	82	0	0
4	76	80	80	84	80	81	0	0
5	75	78	82	84	84	81	0	0
6	75	76	79	84	82	81	0	0
7	75	77	77	86	86	80	0	0
8	76	79	80	85	82	80	0	0
9	76	77	78	83	80	80	0	0
10	76	80	82	83	85	80	0	0
11	76	77	81	85	82	82	0	0
12	80	76	79	85	86	85	0	0
13	74	78	81	84	82	82	0	0
14	74	75	76	84	80	79	0	0
15	75	76	75	85	79	81	0	0
...	...	...	...	...	...	...	...	...
572	80	73	80	82	81	80	1	1
573	78	75	76	80	75	80	1	1
574	80	78	84	83	83	81	1	1
575	74	74	75	82	81	82	1	1
576	80	75	78	80	85	82	1	1
577	86	88	86	87	87	87	1	1
578	80	77	82	82	81	80	1	1
579	79	76	78	82	81	80	1	1
580	81	77	80	82	81	82	1	1
581	80	75	78	83	82	80	1	1
582	80	73	82	83	80	80	1	1
583	80	77	82	80	80	80	1	1
584	81	78	78	83	81	80	1	1
585	80	78	80	82	83	81	1	1
586	78	74	79	83	81	80	1	1

Preprocessing Data

Pre-processing Data adalah salah satu tahap penting dalam data mining untuk meningkatkan kualitas dataset. Penelitian ini berfokus pada penanganan data yang tidak seimbang dalam dataset. Dataset yang digunakan memiliki 350 instance IPA dan 236 instance IPS.



Gambar 3. Distribusi Kelas IPA dan IPS



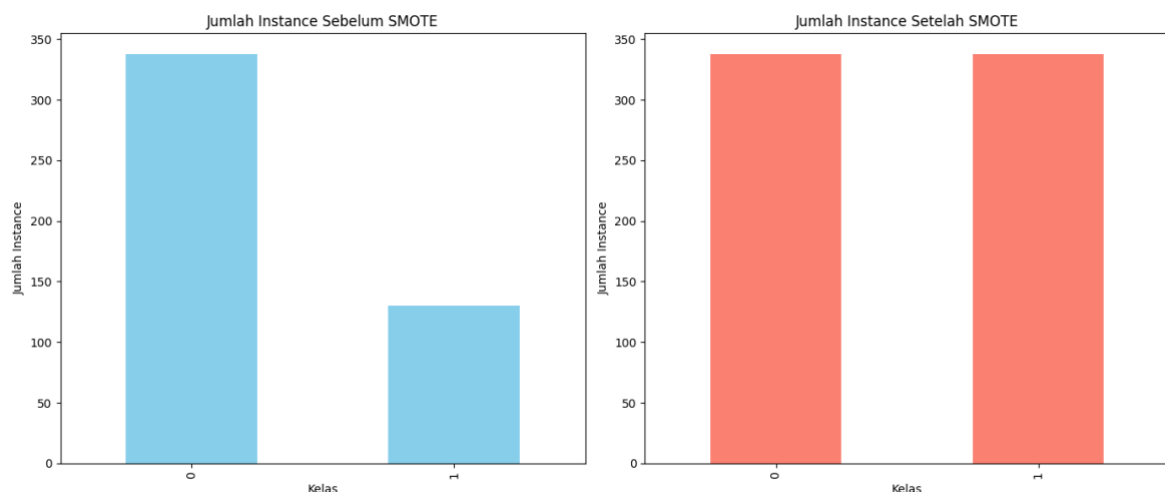
Gambar 4. Distribusi Nilai Mata Pelajaran dan Ketidakseimbangan Kelas

Algoritma yang digunakan untuk menangani data yang tidak seimbang dalam dataset ini adalah SMOTE (Synthetic Minority Oversampling Technique). Sebelum menerapkan teknik ini, langkah pertama yang dilakukan dalam mempersiapkan dataset adalah memeriksa adanya data yang hilang. Proses ini penting untuk memastikan bahwa kualitas data tidak terpengaruh oleh kekosongan yang dapat mempengaruhi analisis dan hasil akhir. Hasil pemeriksaan menunjukkan bahwa tidak ada nilai yang hilang pada setiap kolom dalam dataset. Kemudian dilakukan encoding fitur kategorikal. Pada tahap ini, fitur-fitur kategorikal dalam dataset perlu diubah menjadi format numerik agar dapat digunakan dalam proses analisis dan pemodelan. Encoding adalah teknik yang digunakan untuk mengonversi data kategorikal menjadi data numerik yang dapat diproses oleh algoritma *machine learning*.

Tahapan selanjutnya adalah penerapan SMOTE pada data latih. Teknik ini menghasilkan sampel sintetis untuk kelas minoritas dengan membuat interpolasi antara titik-titik data yang ada. Proses ini meningkatkan jumlah instance pada kelas minoritas dan membantu model untuk belajar dari data yang lebih seimbang. Data baru pada kelas minoritas dibuat menggunakan persamaan berikut.

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

Gambar berikut menunjukkan perbandingan jumlah instance antara kelas-kelas dalam data latih sebelum dan setelah penerapan teknik SMOTE.



Gambar 5. Penerapan Teknik SMOTE

Pemodelan Naive Bayes

Setelah menerapkan teknik SMOTE, tahapan selanjutnya adalah pembangunan model Naive Bayes. Naive Bayes adalah metode klasifikasi probabilistik yang didasarkan pada penerapan teorema Bayes dengan asumsi independensi fitur yang diberikan kelas. Berikut adalah langkah-langkah perhitungan untuk mengilustrasikan pemodelan Naive Bayes pada prediksi pemilihan jurusan SMA.

1. Menghitung Probabilitas A Priori

Dari total 468 data latih, terdiri dari 306 instance dengan label IPA dan 162 instance dengan label IPS. Tabel 4 menampilkan nilai probabilitas untuk data latih.

Tabel 3. Probabilitas Label Kelas

No	Nama	Nilai Probabilitas
1	IPA	0.653846153846154
2	IPS	0.346153846153846
Jumlah		1

2. Menghitung Probabilitas Kondisional

Untuk menghitung probabilitas kondisional ini, perlu untuk mengetahui nilai rata-rata (μ) dan deviasi standar (σ) untuk setiap fitur berdasarkan kelas.

Tabel 5. Nilai Perhitungan Rata-Rata

Label Kelas	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI
IPA	77,35	77,54	79,71	83,57	81,56	81,19
IPS	79,23	78,16	81,30	83,40	82,11	81,25

Tabel 6. Nilai Perhitungan Standar Deviasi

Label Kelas	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI
IPA	3,899	2,505	2,478	2,007	2,722	2,359
IPS	3,436	3,264	3,343	2,829	2,362	2,352

Dengan nilai rata-rata (μ) dan deviasi standar (σ) yang telah dihitung sebelumnya, langkah selanjutnya adalah melakukan perhitungan probabilitas kondisional untuk Data Uji ke-1 untuk setiap fitur guna menentukan kelasnya. Perhitungan dilakukan dengan menggunakan persamaan sebagai berikut.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(x_k - \mu_{ik})^2}{2\sigma^2}} \quad (2)$$

Hasil perhitungan probabilitas kondisional untuk data uji ke-1 dapat dilihat pada tabel 7 di bawah ini.

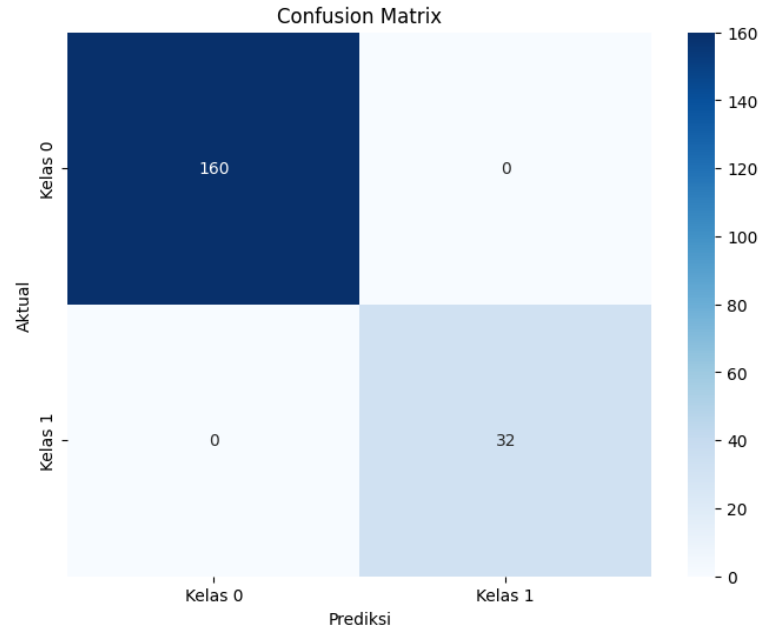
Tabel 4. Nilai Perhitungan Probabilitas Kondisional

Label Kelas	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI	MINAT	LABEL
	87	81	83	85	83	83	IPA	?
IPA	0,0095	0,0972	0,1055	0,2186	0,2103	0,1937	0,997	5,66E-07
IPS	0,0167	0,1515	0,1919	0,2023	0,2418	0,1973	0,006	9,81E-09

3. Berdasarkan perhitungan di atas, nilai probabilitas minat 'IPA' untuk kasus data uji 1 adalah $5,7 \times 10^{-7}$, sedangkan nilai probabilitas minat 'IPS' adalah $9,8 \times 10^{-9}$. Dengan nilai probabilitas minat 'IPA' yang lebih tinggi dibandingkan 'IPS', dapat diprediksi bahwa siswa dengan nilai atribut tersebut cenderung termasuk dalam kategori 'IPA'.

4. HASIL PENELITIAN DAN PEMBAHASAN

Hasil penelitian menunjukkan bahwa tidak terdapat perbedaan signifikan dalam kinerja model yang menggunakan SMOTE dan model yang tidak menggunakan SMOTE. Baik model dengan SMOTE maupun tanpa SMOTE mencapai akurasi sebesar 100%, F1-score sebesar 100%, dan recall sebesar 100%. Hal ini mengindikasikan bahwa teknik SMOTE tidak memberikan peningkatan kinerja yang signifikan dalam kasus ini.



Gambar 6. Confusion Matrix Prediksi Jurusan

5. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis yang telah dilakukan, dapat disimpulkan bahwa penerapan teknik oversampling SMOTE pada dataset yang diteliti tidak memberikan peningkatan signifikan pada kinerja model prediksi pilihan jurusan. Temuan ini mengindikasikan bahwa faktor-faktor lain, seperti pemilihan fitur yang relevan, penyetelan hyperparameter, dan kompleksitas model, memiliki pengaruh yang lebih dominan terhadap kinerja prediksi.

Beberapa saran dapat diajukan untuk penelitian selanjutnya. Pertama, perlu dilakukan analisis fitur yang lebih mendalam untuk mengidentifikasi fitur-fitur yang memiliki kontribusi paling besar terhadap prediksi. Kedua, eksplorasi terhadap algoritma klasifikasi lain yang memiliki mekanisme penanganan ketidakseimbangan kelas yang lebih canggih dapat dilakukan. Ketiga, optimasi hyperparameter model secara lebih cermat dapat meningkatkan kinerja model secara keseluruhan.

6. DAFTAR PUSTAKA

- Bass, L., Weber, I., & Zhu, L. (2015). DevOps: A software architect's perspective. Upper Saddle River, NJ: Addison-Wesley.
- D. Sanwal, "DevOps advantages and disadvantages: Is it worth the hype?," 2021. [Online]. Available: <https://www.jaxenter.com/devops-advantages-and-disadvantages-173668.html>. [Accessed: 16-Mar-2023].
- J. Kim, P. Debois, J. Willis, and J. Humble, The DevOps Handbook: How to Create World-Class Agility, Reliability, and Security in Technology Organizations. IT Revolution Press, 2016.

- Kim, G., Humble, J., & Debois, P. (2016). The DevOps handbook: How to create world-class agility, reliability, and security in technology organizations. IT Revolution.
- M. A. Alqudah, M. T. Alsmadi, and A. H. Alabdallat, "A Systematic Literature Review on DevOps Adoption and Practices," in 2019 IEEE 4th Middle East Conference on Computer Science and Engineering (MEC), 2019, pp. 191–196.