JURNAL ILMIAH SISTEM INFORMASI (JUSI)
VOL 4, No. 3, November 2025, pp. 340 - 350
p-ISSN : 2803-1507 (media cetak) | e-ISSN : 2803-1531 (media online)
Link Publikasi: https://ejurnal.provisi.ac.id/index.php/JUISI/issue/view/151

Classification of Elementary School Children's Potential Based on Hobbies Using the Naive Bayes Method with Undersampling Technique

Umi Karolina*[1], Novan Wijaya[2], Derry Alamsyah[3]
[1,2,3] Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
Jl. Rajawali No. 14, 30113, Palembang, Sumatera Selatan, Indonesia
e-mail*[1]: umikarolina@mhs.mdp.ac.id

ARTICLE INFO

Article History:

Received: 09-06-2025
Revised: 30-06-2025
Accepted: 30-07-2025
Published: 30-11-2025

ABSTRACT

The identification of children's potential at an early age plays a crucial role in guiding educational strategies and parental decisions. This study aims to classify the potential of elementary school children, categorized into academic, artistic, and athletic domains, based on their hobbies using the Naive Bayes algorithm combined with a Random Undersampling technique to address class imbalance in the dataset. Data was sourced from a publicly available dataset and contextualized for Indonesian students through additional surveys. Preprocessing steps included data cleansing, categorical transformation, and balancing using Random Undersampling. A total of 1,230 records were used, evenly distributed across the three categories. The model was trained using an 80:20 train-test split and validated with 10-fold cross-validation. Evaluation metrics based on the confusion matrix show promising performance, with the highest precision and recall achieved in the artistic category (92.13%), and an overall classification accuracy of 90.24%. The results indicate that the integrated approach effectively reduces prediction bias towards the majority class, ensuring balanced model performance. This research demonstrates the applicability of probabilistic classification models in educational contexts and highlights the significance of balanced datasets when predicting multi-class potential in children. Future studies are encouraged to include more diverse features to enhance classification granularity and accuracy.

Keywords: *Child Potential, Naive Bayes, Random Undersampling, Classification.*

Abstrak

Identifikasi potensi anak sejak usia dini memiliki peran penting dalam merancang strategi pendidikan dan pengambilan keputusan oleh orang tua. Penelitian ini bertujuan untuk mengklasifikasikan potensi anak sekolah dasar dalam tiga kategori utama yaitu akademik, seni, dan olahraga berdasarkan hobi yang mereka miliki. Metode yang digunakan adalah algoritma Naive Bayes yang dikombinasikan dengan teknik Random Undersampling untuk mengatasi ketidakseimbangan data antar kelas. Data diambil dari sumber publik dan disesuaikan dengan konteks siswa Indonesia melalui survei tambahan. Proses pra-pemrosesan meliputi pembersihan data, transformasi kategorikal, serta penyeimbangan data menggunakan teknik Random

DOI: https://doi.org/10.51903/wxnxrm24

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Undersampling. Sebanyak 1.230 data digunakan dan dibagi merata ke dalam tiga kategori tersebut. Model dilatih menggunakan skema pembagian data 80:20 serta divalidasi dengan 10-fold cross-validation. Hasil evaluasi menggunakan confusion matrix menunjukkan kinerja yang menjanjikan, dengan presisi dan recall tertinggi pada kategori seni (92,13%), serta akurasi klasifikasi keseluruhan sebesar 90,24%. Temuan ini menunjukkan bahwa pendekatan kombinasi mampu mengurangi bias prediksi terhadap kelas mayoritas, sehingga menghasilkan model yang lebih seimbang. Penelitian ini membuktikan efektivitas model klasifikasi probabilistik dalam konteks pendidikan dan mendorong studi lanjutan untuk menyertakan atribut yang lebih beragam guna meningkatkan ketepatan klasifikasi.

Kata Kunci: *Potensi Anak, Nave Bayes, Random Undersampling, Klasifikasi.*

1. PENDAHULUAN

Di fase usia sekitar 9-12 tahun merupakan priode yang sanggat penting dalam perkembangan anak dimana minat dan potensi mulai terbentuk secara jelas. Anak-anak di usia sekolah dasar cenderung belajar lebih efektif melalui pendekatan yang menyenangkan seperti bermain dan aktivitas kreatif. Berdasarkan data yang dikumpulkan dari wawancara pribadi terhadap 50 siswa sekolah dasar, terlihat bahwa minat siswa terhadap hobi yang difasilitasi oleh kurikulum didominasi oleh olahraga bola kaki (44%), diikuti oleh kegiatan seni seperti menggambar (20%) dan menari (8%), sementara minat terhadap aktivitas akademik seperti belajar hanya sebesar 1%. Sebaliknya, dalam kegiatan non-kurikulum, siswa lebih tertarik pada aktivitas digital, di mana bermain game menjadi pilihan utama (58%), diikuti oleh makan (12%) dan membuat konten (8%). Data ini menunjukkan bahwa siswa lebih antusias terhadap aktivitas hiburan dan digital di luar kegiatan formal sekolah. minat terhadap kegiatan non-akademik seperti bermain game dan olahraga jauh lebih didominan dibandingkan dengan minat aktual anak [1].

Dalam konteks pendidikan dasar, perkembangan anak mencakup enam aspek fundamental yang kemudian sering disederhanakan menjadi tiga domain utama: akademik, seni, dan olahraga. Ketiga domain ini saling melengkapi dalam membentuk potensi holistik anak melalui berbagai aktivitas pembelajaran tematik dan pembiasaan. Stimulasi kognitif yang tepat, seperti melalui permainan edukatif dan eksperimen sederhana, terbukti secara signifikan meningkatkan kemampuan berpikir kritis dan pemecahan masalah [2]. Adapun juga model pembelajaran kooperatif yang diharapkan agar siswa bisa meningkatkan pembelajaran siswa dengan aktif dengan membangun pola pikir peserta didik agar dapat berfikir kritis, inovatif, dan kreatif dalam menyelesaikan sebuah masalah [3].Paradigma pendidikan kontemporer menekankan bahwa keberhasilan pembelajaran tidak lagi semata diukur dari pencapaian akademik, tetapi juga mencakup perkembangan keterampilan non-akademik seperti kreativitas, komunikasi, dan kemandirian [4].

Adapun beberapa penelitian terkait tentang potensi anak, metode berbasis data telah dikembangkan dalam penelitian terdahulu untuk memahami potensi anak. [5] menemukan bahwa teknik K-Nearest Neighbor (KNN) memiliki tingkat akurasi 73,33% dalam mengidentifikasi minat berdasarkan nilai ekstrakurikuler. Dalam penelitian [6] digunakan algoritma C4.5 dan Random Forest untuk memprediksi minat anak berdasarkan hobi dan aktivitasnya. Tingkat akurasi algoritma Random Forest adalah 92,7%. [7]. juga melakukan klasifikasi minat olahraga anak sekolah dasar menggunakan metode Naive Bayes dengan hasil yang baik. Selanjutnya, [8] berhasil membuat model klasifikasi minat siswa dengan tingkat akurasi 99,47% menggunakan Naive Bayes. Penelitian [9] mengunakan algoritma *Random Fores* , *Naiv Bayes* , dan KNN untuk klasifikais tanaman keladi tikus dengan mengunakna Teknik Random undersampling, tingkat akurasi tertinggi mengunakan algoritma Random Fores 99.1%.

Berdasarkan penelitian sebelumnya, sebagian besar studi mengabaikan masalah ketidakseimbangan data dan hanya menggunakan satu algoritma untuk mengkategorikan potensi anak. Oleh karena itu, studi ini memberikan kontribusi dengan mengintegrasikan teknik Random Undersampling dengan metode Naive Bayes, yang diharapkan dapat meningkatkan kinerja klasifikasi multikelas potensi anak sekolah dasar dan mengurangi bias prediksi di kelas mayoritas. Penelitian ini meyorot pentingknya klasifikasi potensi anak dalam tiga kategori utama yaitu: Akademik, Seni, dan Olahraga. Oleh karena itu, tujuan utama penelitian ini adalah untuk mengevaluasi seberapa baik kinerja metode Naive Bayes dalam mengidentifikasi potensi anak dan memeriksa bagaimana strategi Random Undersampling berkontribusi terhadap peningkatan akurasi klasifikasi pada kumpulan data yang tidak seimbang.

2. TINJAUAN PUSTAKA

p-ISSN : 2803-1507 e-ISSN : 2803-1531

2.1. *Potensi dan Minat Anak Usia Sekolah dasar*

Masa kanak-kanak merupakan masa penting dalam perkembangan individu, terutama pada rentang usia 9–12 tahun. Anak pada usia ini menunjukkan minat dan kecenderungan bakat yang lebih spesifik, sehingga penting bagi pendidik dan orang tua untuk mengenali potensi yang dimilikinya sejak dini [2]. stimulasi kognitif melalui kegiatan edukatif dapat meningkatkan kemampuan berpikir kritis dan pemecahan masalah pada anak. Selain itu, paradigma pendidikan kontemporer mendorong pengembangan potensi tidak hanya dari aspek akademik, tetapi juga seni dan olahraga [4].

2.2. *Minat Sebagai Faktor Penggerak Potensi*

Salah satu komponen internal yang sangat memengaruhi jalur pertumbuhan anak di masa mendatang adalah minat. Menurut KBBI, minat adalah kecenderungan hati terhadap sesuatu yang jika dibimbing dengan tepat akan memaksimalkan potensinya. Pemahaman terhadap minat anak masih sering keliru dalam praktiknya, misalnya ketika harapan orang tua berbenturan dengan minat bawaan anak [10]. Adapun minat menurut penelitian [11] adalah potensial yang harus dimiliki oleh setiap siswa dalm peroses belajar 341ikelomp dapat mencapai tujuan yang diharapkan, minat belajar itu sendiri harus dibagun dan dikembangkan didalam diri setiap siswa .

2.3. *Klasifikasi Potensi Anak Berbasis Machine Learning*

Beberapa penelitian telah memanfaatkan metode Machine Learning untuk klasifikasi potensi anak, Penelitian yang dilakukan oleh [6], menunjukan bahwa hobi dan aktivitas anak dapat digunakan sebagai bahan referensi untuk memprediksi potensi minat anak dalam bidang akademik maupun non-akademik menggunakan model prediksi Algoritma C4.5 dan *Random Forest*. Kedua model prediksi menunjukkan kinerja yang sangat baik berdasarkan sampel data aktivitas anak. Hasil pengujian validasi dilakukan menunjukkan secara rata-rata algoritma *Random Forest* lebih unggul daripada C4.5 baik dari data testing dengan nilai akurasi, *f-score*, *presisi* dan *recall* adalah 92.7%, 92.8%, 92.8% dan 92.7%.

Adapun Penelitian terdahulu yang dilakukan oleh Kustanti dan teman-teman mengenai klasifikasi berdasarkan nilai ekstrakurikuler dengan penerapan metode KNN (*K-Nearest Neighbor*) berhasil diterapkan. Dengan penerapan metode KKN ini dapat lebih mempermudahkan siswa dalam memilih kriteria ekstrakurikuler pada data yang ada. Maka didapat hasil pengujian Metode KNN untuk tingkat keakuratan data bahwa nilai keakuratan data adalah 73.333%. selanjutnya Metode machine learning berikutnya yang digunakan adalah Artificial Neural Networks (ANN) [5]. Studi internasional Çetinkaya & Baykan, membahas permasalahan terkait keterampilan minoritas pada siswa sekolah menengah dalam bidang pemrograman sebagai lintasan karier di masa depan. Di antara berbagai metode pelatihan ANN, pengujian dengan regularisasi *Bayesian* memberikan hasil terbaik, dengan akurasi mencapai 97,2% [12]. Metode machine learning berikutnya yang digunakan adalah Jaringan Syaraf Tiruan dan Multilayer Perceptron (MLP). Penelitian yang dilakukan oleh [13] bertujuan memprediksi potensi pemain tenis berdasarkan profil prasyarat kinerja anak untuk meraih kesuksesan di masa depan.

Penelitian ini menggunakan data hasil tes fisik dan tes kompetensi motorik. Hasil terbaik diperoleh dengan mengombinasikan metode non-linier (Jaringan Saraf Tiruan) dengan salah satu metode linier. Namun, 39,2% pemain tenis berprestasi tinggi tetap tidak dapat diprediksi menggunakan salah satu metode tersebut [13]. Penelitian internasional lainya Jauhiainen Penelitian menerapkan metode SVM untuk mendeteksi pemain sepak bola berbakat. SVM digunakan sebagai pendekatan klasifikasi satu kelas (*one-class classification*) untuk mengidentifikasi pola khas dari pemain berbakat. Namun, hasil yang diperoleh belum optimal, dengan performa terbaik hanya mencapai akurasi 76,3%. [14]. Hasil tersebut memperkuat pilihan metode Naive Bayes untuk penelitian ini karena metodologi probabilistiknya yang lugas namun fleksibel. Berbeda dengan penelitian sebelumnya, penelitian ini membahas bias dalam dataset yang tidak seimbang dengan menggabungkan pendekatan Naive Bayes dan Random Undersampling, sebuah metodologi yang jarang digunakan dalam kategorisasi potensial siswa sekolah dasar.

2.4. *Algoritma Naive Bayes*

Berdasarkan *Teorema Bayes*, algoritma *Naive Bayes* merupakan sistem klasifikasi probabilistik yang menggabungkan independensi fitur. Pendekatan ini efisien secara komputasi dan sesuai untuk kategori data dan numerik [7]. Menurut [15] teori pengambilan keputusan berdasarkan teorima bayes ialah pendekatan statistika yang *fundamental* dalam data mining Pendekatan ini didasarkan pada kuantifikasi *trade-of* antara

p-ISSN : 2803-1507 e-ISSN : 2803-1531

berbagai keputusan klasifikasi dengan menggunakan *probabilitas* karena itu model *Naive Bayes* sering digunakan untuk klasifikasi data *mining*. Metode Naïve bayes juga menjadi suatu pilihan yang sanggat popular dalam klasifikasi dan analisis prediktif karena tergolong sederhana dan mudah dipahami serta mendapatkan hasil klasifikasi yang memuaskan [16]. Dari [17] yang menggunakan Algoritma *Naive Bayes* dapat mengklasifikasikan korosivitas tanah mendapatkan akurasi sebesar 83.3%.

2.5. *Random Undersampling*

Pengumpulan data yang tidak seimbang, di mana kuantitas data dari setiap kelas tidak merata, merupakan salah satu kesulitan dalam kategorisasi machine learning. Bias model yang menguntungkan kelas mayoritas dapat terjadi akibat hal ini. Strategi undersampling acak, yang mencakup penghapusan data dari kelas dominan untuk menyeimbangkan distribusi, sering digunakan untuk mengatasi masalah ini [18]. Adapun menurut penelitian [19] Undersampling ialah menghasilkan sub sampel acak dari setiap kelas mayoritas dan merupakan metode sampling secara acak memilih sample dikelas mayoritas untuk di hilangkan agar mendapatkan dataset baru yang seimbang. Di dalam penelitian [20] menggunakan metode CNN dengan Teknik Random undersampling untuk deteksi CT Scan Kista Ginjal mendapatkan akurasi sebesar 99% terbilang hamper mendekati sempurna. Selain metode Undersampling, [21] memperkenalkan SMOTE (Synthetic Minority Over-sampling Technique) yang efektif menangani data tidak seimbang pada dataset berukuran besar. Namun, teknik Random Undersampling tetap relevan untuk dataset berskala kecil-menengah karena lebih sederhana dan mengurangi risiko overfitting [22].

3. METODOLOGI PENELITIAN

3.1. *Desain Penelitian*

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan metode pendekatan klasifikasi berbasis *Supervised Learning* model pembelajaran yang menggunakan data berlabel sebagai dasar prediksi atau klasifikasi [23]. Algoritma yang digunakan adalah *Naive Bayes*, untuk mengkategorikan ketiga kelas potensi anak utama: Akademi, Seni, dan Olahraga. Untuk mengatasi permasalahan ketidak seimbangan dataset yang digunakan penulis menggunakan teknik *Random Undersampling*.

3.2. *Sumber Data Penelitian*

Metodologi Pengumpulan data merupakan langkah penting untuk memastikan ketersediaan dataset, pengumpulan data dalam penelitian ini bertujuan untuk memperoleh informasi yang relevan mengenai potensi anak berdasarkan hobi yang di minati serta dapat mengembangkan potensi tersebut. Metode pengumpulan data yang di lakukan adalah kuesioner terstruktur serta data yang telah di validasi. Dataset yang digunakan diambil dari Kaggle karena telah menyediakan data terkait aktivitas, hobi, dan minat anak sekolah dasar yang bersifat umum dan dapat dijadikan acuan awal. Meskipun dataset berasal dari luar negeri, peneliti melakukan penyesuaian atribut dan validasi melalui kuesioner tambahan pada 50 siswa SD di Indonesia agar konteksnya relevan. Dengan demikian, karakteristik hobi, waktu luang, serta kegiatan ekstrakurikuler tetap mencerminkan kondisi anak-anak SD di Indonesia. (https://www.kaggle.com/datasets/abtabm/hobby-prediction-basic) diperoleh dengan mematuhi standar etika penelitian dan kebijakan yang berlaku. Data yang terkumpul melalui 342ikelo diolah untuk mendapatkan dataset yang seimbang dengan menggunakan teknik undersampling untuk menyesuaikan data guna diterapkan untuk mengklasifikasi potensi anak SD. Penjelasan mengenai atribut yang terdapat dalam dataset disajikan pada Tabel 1.

Tabel 1. Penjelasan Atribut pada Data Hobi Anak

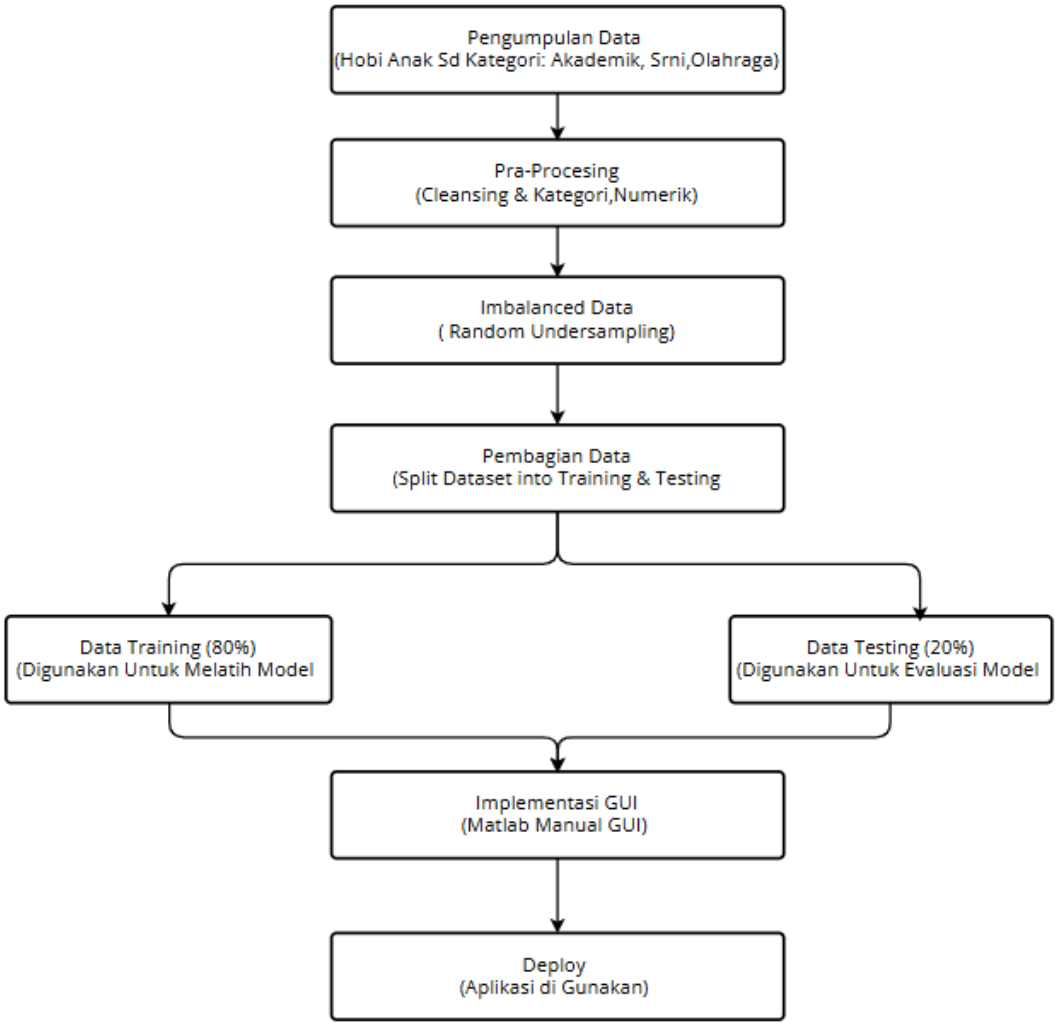
Atribut	Question	Type Data	Skala
Olympiad_Participation	Has your child participated in any Science/Maths Olympiad?	Boolean	Ya/Tidak
Scholarship	Has he/she received any scholarship?	Boolean	Ya/Tidak
School	Love's going to school?	Boolean	Ya/Tidak
Fav_sub	What is his/her favorite subject?	Kategori	Matematika, Sains, Bahasa Apapun, Sejarah/Geografi
Projects	Has done any projects under academics before?	Boolean	Ya/Tidak
Time_sprt	How much time does he/she spend playing outdoor/indoor games?	Numerik	1,2,3,…12 jam/hari

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Medals	Medals won in Sports?	Boolean	Ya/Tidak
Career_sprt	Want's to pursue his/her career in sports?	Boolean	Ya/Tidak
Act_sprt	Regular in his/her sports activities?	Boolean	Ya/Tidak
Fant_arts	Love creating fantasy paintings?	Boolean	Ya/Tidak
Won_arts	Won art competitions?	Boolean	Ya/Tidak
Time_art	Time utilized in Arts?	Numerik	1,2,3,…12 jam/hari

3.3. *Perancangan Sistem*

Pada tahapan ini dilakukan perancangan sistem yang dibutuhkan untuk memprediksi potensi anak SD, sehingga dalam hal ini menggunakan metode pembelajaran *Supervised Learning* (Pembelajaran Terawasi) yang terdiri dari dua tahapan *Training* dan *testing,* Berikut diagram alir yang digunakan untuk perancangan sistem. Diagram alir yang digunakan untuk perancangan sistem ditampilkan pada Gambar 1.



Gambar 1. Diagram Alir

Pada Gambar 1 Perancangan Model, menjelaskan dengan tahap – tahap berikut :

1. *Pengumpulan Data*

diawali dengan pengumpulan dataset sejumlah 1601 data, Fokus data yang di ambil adalah minat yang dilakukan siswa disekolahan dalam hal akademik maupun non akademik. Adapun informasi tentang hobi mereka, yang dapat 343ikelompokkan dalam tiga kategori utama yaitu Olahraga, Seni, Akademik.

2. *Pra-prosessing*

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Tahap ini adalah tahapan yang paling penting sebelum data akan digunakan dalam pelatihan model, data harus diclensing dan mengubah datah mentah menjadi data yang konsisten agar dapat diproses oleh model Naiv bayes. Langkah-langkah pembersihan data mencakup:

3. *Cleasing Data*

Setelah melakukan konversi string ke double, perlu dipastikan jika ada data kosong dapat diubah menjadi 0 sebagai penanganan data yang hilang. Untuk menghidari error pada pemrogresaan pada model Di mana data yang tidak memiliki nilai akan di jadikan NAN, berikut contoh data yang sudah dicleansing dengan sebelum dicleansing: Gambar 2 menampilkan kondisi data sebelum dilakukan proses *cleansing*, sedangkan Gambar 3 menunjukkan hasil setelah data dibersihkan.

Sebelum Cleansing:

	Time_sprt	Time_art
0	2	1
1	1	2
2	4	1
3	1	2
4	5	1

Gambar 2. Sebelum Cleansing

Setelah Cleansing:

	Time_sprt	Time_art
0	2	1
1	1	2
2	4	1
3	1	2
4	5	1

Gambar 3. Sesudah Cleansing

Pada gambar diatas dilakukan *Cleansing* data dibagian fitur Time_sprt dan time_art bisa diatritakan dataset yang didapatkan sudah sanggat bersih dan siapj digunakan untuk Klasifikasi kedalam model *Naive Bayes*.

4. *Transformasi Kategorikal*

Tujuan melakukan transformasi kategorikan untuk mengubah non-numerik menjadi numerik agar bisa dimasukan ke dalam model. Tujuan melakukan transformasi kategorikan untuk mengubah non-numerik menjadi numerik agar bisa dimasukan ke dalam model contoh pada Gambar 4.

1	2	3	4	5
[1;2]	[1;2]	[1;2]	[1;2;3;4]	[1;2]

Gambar 4. Transformasi Kategorikal

Dari gambar di atas dapat di lihat perubahan kategorikal menjadi numerik di mana [1;2] berisikan YA dan TIDAK sedangkan [1;2;3;4] berisikan matematika, sains, bahasa, sejarah, ini disebut label encoding yang memberikan angka untuk tiap kategori seperti diatas.

5. *Penangan Imbalanced Data*

Pada tahapan ini jika jumlah data pada suatu kelas lebih besar dibandingkan kelas yang lainya. Kondisi dapat menimbulkan bias pada model prediksi, dimana model akan lebih cenderung memprediksi kelas yang mayoritas dibanding kelas minoritas. karena itu dilakukan lah teknik *Random Undersampling* untuk penyetaraan data agar data seimbang.

6. *Pembagian Data*

Pembagian dataset merupakan suatu langkah yang penting dilakukan dalam proses untuk pengembangan model. Tujuanya agar dapat memisahkan data menjadi dua *subset* itama yaitu data Training dan data Testing yang masing-masing memiliki peran tersendiri dalam proses platihan dan evaluasi model. Pada penelitian ini penulis membagi 80% untuk data *Trening* dan 20% untuk data *Testing* dari 1601 dataset klasifikasi hobi anak

p-ISSN : 2803-1507 e-ISSN : 2803-1531

yang mana pembagian dilakukan secara acak untuk memastikan distribusi data yang seimbang. Penelitian yang dilakukan oleh Herianto dan rekan-rekan menggunakan pengujian splitting data yang menghasilkan prediksi dari model *Naïve Bayes*, seperti ditunjukkan pada Gambar 5.

N0	Split Data	Accuracy
1	90:10	92.30%
2	80:20	96.15%
3	70:30	82.05%
4	60:40	84.61%
5	60:40	84.61%
6	50:50	85.93%

Gambar 5. Hasil Pengujian Naïve Bayes

Hasil dari gambar tabel di atas dapat di simpulkan menggunakan pembagian data 80:20 mendapatkan nilai akurasi yang paling tinggi dengan akurasi 97.6% [24].

7. Evaluasi Model

Selanjutnya melakukan evaluasi kinerja model mengunakan data testing, evaluasi dilakukan dengan tujuan untuk mengetahui sejauh mana model dapat melakukan prediksi secara benar, dalam penelitian ini proses evaluasi dilakukan dengan membandingkan label hasil prediksi model dengan label aktual dari data *testing.*

4. HASIL DAN PEMBAHASAN

4.1 *Analisis Data*

Dataset yang digunakan dalam penelitian ini terdiri dari 1.601 data siswa anak sekolah dasar yang mencerminkan aktivitas dan minat anak dalam tiga kategori potensi: Akademik, Seni, dan Olahraga. Setelah dilakukanya proses data *proprocessing*, bentuk data dikonver menjadi data *numerik* dan *kategorikal* yang disesuaikan untuk algoritma *Naive bayes*. Namun, dataset memiliki ketidak semimbangan data yang mana akan mengakibatkan model bias terhadap kelas mayoritas karena itu dibutuhkan teknik Keseimbangan data dengan menggunakan teknik *Random Undersampling*, yaitu dengan cara pengurangan jumlah data pada kelas mayoritas secara acak agar distribusi antar kelas menjadi seimbang. Jumlah data akhir dari masing-masing kelas setelah proses *undersampling* dapat dilihat pada Tabel 2.

Tabel 2. Dataset Undersampling

Kategori	Jumlah Data
Akademik	410
Seni	410
Olahraga	410
Total	1.230

Data kemudian dibagi menjadi dua bagian, yaitu data pelatihan sebesar 80% dan data pengujian sebesar 20%. Hasil distribusi kelas pada kedua bagian data tersebut menunjukkan bahwa dataset telah berhasil diseimbangkan. Rincian distribusi data setelah *pre-processing* disajikan pada Tabel 3.

Tabel 3. Distribusi Data Setelah Pre-processing

Jenis Data	Akademik	Seni	Olahraga
Training	334	318	332
Testing	76	92	78

Dari tabel di atas dapat dilihat bahwa jumlah data pada masing-masing kategori hobi cukup seimbang, baik pada data pelatihan maupun data pengujian. Pada data pelatihan, kategori akademik dan olahraga masing-masing memiliki 327 data, sedangkan seni memiliki 330 data. Sementara itu, data pengujian terdiri dari 83 data untuk kategori akademik, 80 data untuk seni, dan 83 data untuk olahraga.

3.3 Metrik Evaluasi

Model menggunakan confusion matrix sebagai alat ukur untuk mengevaluasi performa model berdasarkan metrik *accuracy, precision,* dan *recall* pada setiap data uji. Untuk memastikan validitas hasil pengujian model, dilakukan teknik pre-processing berupa *random undersampling* guna menyeimbangkan distribusi data. *Confusion matrix* adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah. Contoh *confusion matrix* untuk klasifikasi biner ditunjukkan pada Tabel 4.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Tabel 4. Confusion Matrix

		Kelas Prediksi	
		1	**0**
Kelas Sebenarnya	**1**	TP	FN
	0	FP	TN

Keterangan tabel 4:

TP (*True Positive*) : Jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1
TN (*True Negative*) : Jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0
FP (*False Positive*) : Jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1
FN (*False Negative*) : Jumlah dokumen dari kelas 1yang salah diklasifikasikan sebagai kelas 0

Adapun rumus *confusion matrix* untuk menghitung *accuracy, precision,* dan *recall* seperti berikut [25].

$$accuracy = \frac{TP+TN}{Total} \quad (1)$$

$$precision = \frac{TP}{TP+FP} \quad (2)$$

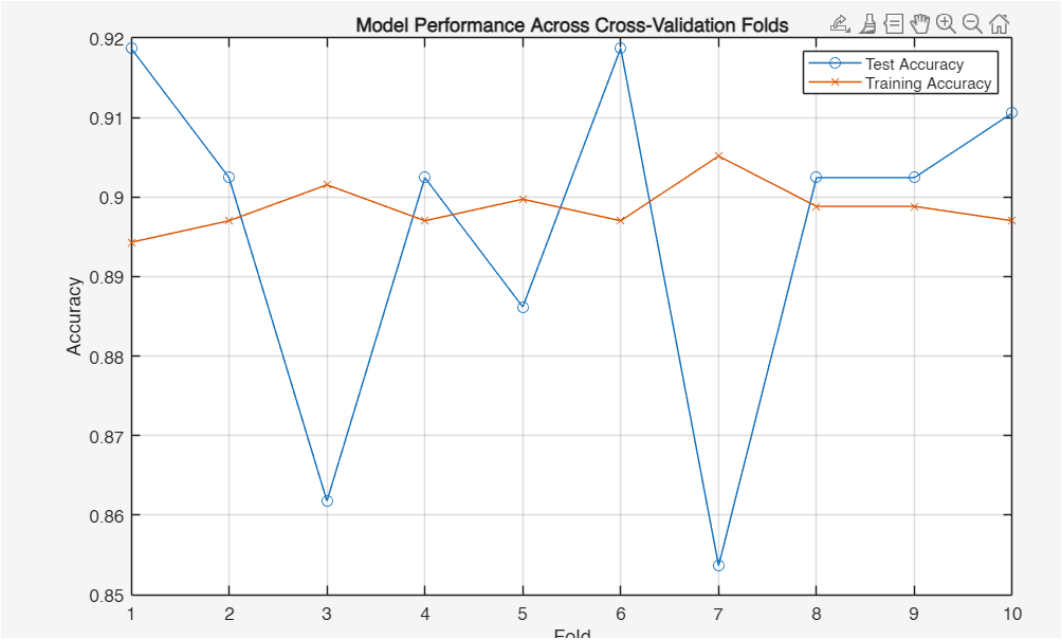
$$recall = \frac{TP}{TP+FN} \quad (3)$$

4.2 *Hasil Pelatihan Model*

Model Naive Bayes dilatih menggunakan data hasil *Undersampling* dan diuji menggunakan metode berikut:

A. 10-Fold Cross Validation

Pengujian model dilakukan menggunakan data yang telah melalui tahap pra-pemrosesan dan penyeimbangan. Sebelum dilakukan evaluasi akhir dengan *Confusion matix,* model juga diuji menggunakan metode *10-Fold-Cross-Validation* untuk memastikan bahwa model *Naive Bayes* memiliki performa yang stabil. metode *10-Fold-Cross-Validation* menggunakan pendekatan validasi silang dengan membagi data menjadi 10 bagian (*Fold*), kemudian melakukan pelatihan dan pengujian sebanyak 10 kali bergiliran sehingga setiap bagian data memperoleh kesempatan sebagai data uji. Berikut gambar model *perfomance cross-validation-folds*. Visualisasi performa model berdasarkan *cross-validation folds* ditampilkan pada Gambar 6.



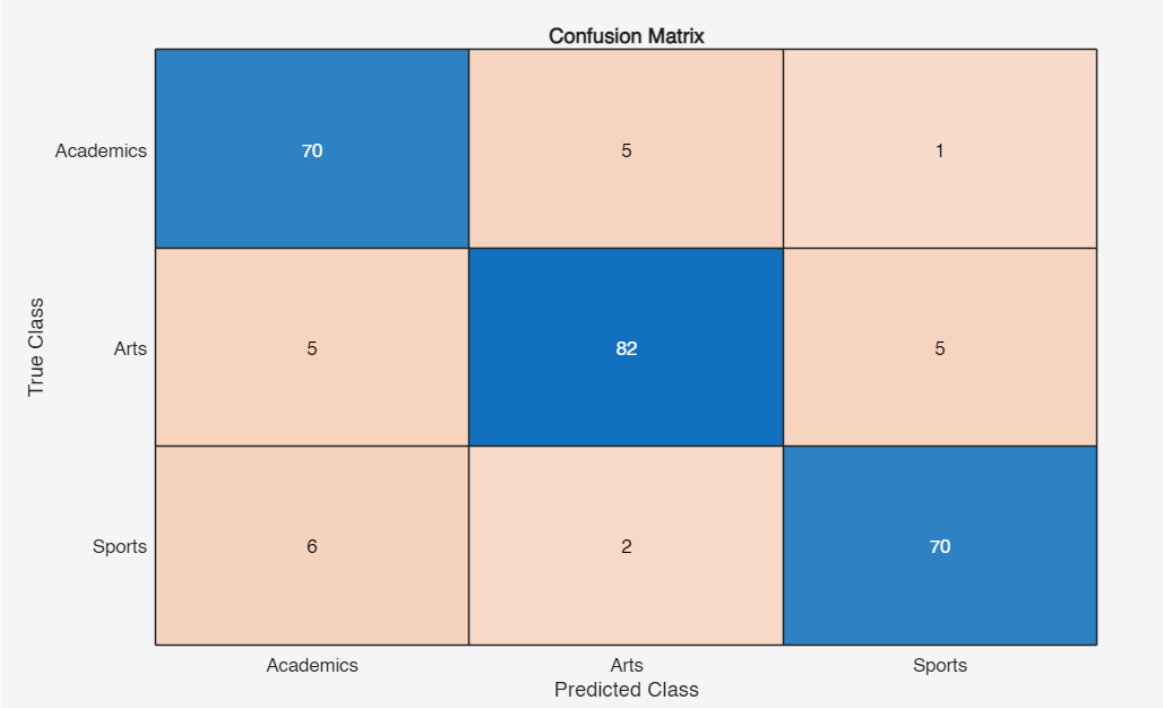
Gambar 6. 10-Fold Cross Validation

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Berdasarkan pengujian validasi silang 10 kali lipat, akurasi pelatihan sebagian besar tetap konstan di kisaran 89% hingga 91%, sementara akurasi pengujian berkisar antara 86% hingga 92%. Berkat kinerjanya yang konsisten pada data validasi silang dan minimnya overfitting yang signifikan, pola ini menunjukkan bahwa model *Naive Bayes* sangat andal. Dalam konteks kategorisasi multikelas potensi siswa sekolah dasar dalam domain Akademik, Seni, dan Olahraga, model ini sangat andal, terutama setelah menggunakan strategi Random Undersampling untuk menjaga distribusi data yang seimbang.

B. Confusion Matrix

Untuk melihat prforma model secara lebih mendlalam, dilakukan evaluasi mengunakan confusion matrix. Evaluasi ini bertujuan untuk mengetahui akurasi prediksi model terhadap masing-masing kategori kelas yaitu Akademik, Seni dan Olahraga. Evaluasi ini akan dilakukan menggunakan data yang telah melalui tahap pra-pemrosesan dan penyeimbangan. Dataset yang digunakan terdiri dari 1.230 data, yang terbagi secara merata ke dalam tiga kelas, masing-masing berjumlah 410 data. Adapun ketiga kelas tersebut meliputi kelas Akademik, Seni, dan Olahraga. Hasil evaluasi dalam bentuk *confusion matrix* ditampilkan pada Gambar 7.



.Gambar 7. Confusion Matrix

Confusion matix diatas menunjukkan hasil perfom model klasifikasi *Naive Bayes* dalam memprediksi potensi anak berdasarkan tiga kategori hobi anak yaitu Akademik, Seni, dan Olahraga. Setiap baris menunjukan kelas sebenarnya (*True Class*), dan setiap kolom menunjukkan kelas prediksi (*Predicted Class*).

Adapun cara evaluasi perkelasnya sebagai berikut:

1. Kelas Akademik

Menggunakan rumus perhitungan Precision dan Recall

True Positive (TP) : 70
False Positive (FP) : 5 (dari seni) + 6 (dari olahraga) = 11
False Negative (FN) : 5 + 1 = 6

2. Kelas Seni

Menggunakan rumus perhitungan Precision dan Recall

True Positive (TP) : 82

p-ISSN : 2803-1507 e-ISSN : 2803-1531

False Positive (FP) : 5 (dari Akademik) + 2 (dari olahraga) = 7
False Negative (FN) : 5 + 5 = 10

Menggunakan rumus Akurasi

3. Kelas Olahraga

Menggunakan rumus perhitungan Precision dan Recall

True Positive (TP) : 70
False Positive (FP) : 1 (dari Akademik) + 5 (dari Seni) = 6
False Negative (FN) : 6 + 2 = 8

Confusion Matrix yang telah disebutkan sebelumnya digunakan untuk menentukan nilai *Presisi*, Ingatan, dan Skor F1 pada Tabel 5 di bawah ini. Metrik evaluasi ini diturunkan dari perhitungan Positif Benar (TP), Positif Palsu (FP), dan Negatif Palsu (FN) untuk setiap kelas. Oleh karena itu, kinerja model *Navie Bayes* dalam mengidentifikasi tiga kelompok (Akademik, Seni, dan Olahraga) dapat diinterpretasikan secara kuantitatif melalui Tabel 5.

Tabel 5. Hasil Evaluasi Setiap Kelas

Hasil	Kelas		
	Akademik	Seni	Olahraga
Presisi	89.61%	92.13%	88.75%
Recall	85.19%	92.13%	92.42%
FI-Score	87.34%	92.13%	91.03%
Support	31.30%	36.18%	32.52%

Berdasarkan tabel diatas dapat dilihat, model naive bayes menunjukkan performa terbaik pada kategori Seni, dengan *presisi* dan *recall* yang sama tinggi, yaitu 92.13%. Sementara itu, meskipun kategori Akademik sedikit lebih rendah dari sisi *recall*, nilai *F1-score* masih tetap tinggi, menandakan keseimbangan antara prediksi benar dan salah. Dengan model *naive bayes* memberikan hasil yang baik dengan distribusi perfoma yang cukup merata pada setiap kelas. Teknik *random undersampling* terbukti mampu mengurangi bias terhadap kelas mayoritas dan meningkatkan perfoma keseluruhan model.

Akurasi sebesar 90,24% dalam penelitian ini merupakan hasil yang cukup kompetitif jika dibandingkan dengan akurasi algoritma Random Forest sebesar 92,7% yang digunakan dalam penelitian [6]. dan akurasi metode KNN sebesar 73,33% yang digunakan dalam penelitian [5]. Selain itu, penelitian [8] menggunakan Naive Bayes dan memperoleh akurasi yang lebih tinggi, yaitu 99,47%, meskipun tidak dilengkapi metode untuk menangani data yang tidak seimbang. Penelitian ini mencegah bias model terhadap kelas mayoritas dengan menggunakan Random Undersampling untuk menciptakan distribusi kinerja yang lebih merata di setiap kelas. Hasil ini mendukung gagasan bahwa teknik penyeimbangan data dapat memainkan peran penting dalam meningkatkan akurasi klasifikasi multikelas potensi anak.

5. KESIMPULAN DAN SARAN

5.1 *Kesimpulan*

Studi ini menunjukkan bahwa strategi *Random Undersampling* meningkatkan kinerja klasifikasi dan menyeimbangkan distribusi data secara efektif. Dengan akurasi keseluruhan 90,24%, model *Naive Bayes* mampu membagi potensi siswa sekolah dasar menjadi tiga kategori utama: akademik, artistik, dan atletik. Keandalan model dalam mengelola klasifikasi multikelas pada dataset yang awalnya tidak seimbang dikonfirmasi oleh evaluasi *metrik* per kelas, yang juga menunjukkan nilai presisi dan recall di atas 85% untuk semua kelas. Evaluasi per kelas menunjukkan bahwa kelas Seni memberikan performa terbaik dengan *presisi* dan *recall* sama tinggi yaitu 92.13%, kemudian kelas Olahraga memiliki nilai recall tertinggi sebesar 93.42%. dan kelas Akademik memiliki nilai *F1-score* yang juga cukup tinggi, yaitu 87.34%.Penggunaan metode bayes dalam klasifikasi multikelas pada dataset yang telah diseimbangkan dapat menghasilkan model yang cukup akurat dan tidak bisa terhadap satu kelas tertentu.

5.2 *Saran*

Penelitian ini masih menghadapi kendala dalam hal pengumpulan data, khususnya terkait dengan ketiadaan fitur-fitur spesifik yang mencerminkan aspek-aspek terperinci dari hobi masing-masing anak. Dalam dataset

p-ISSN : 2803-1507 e-ISSN : 2803-1531

yang digunakan, hobi anak-anak hanya diklasifikasikan secara umum ke dalam tiga kelompok utama, yaitu olahraga, seni, dan akademis, tanpa mempertimbangkan karakteristik unik dari setiap jenis hobi di dalam kelompok tersebut. Keterbatasan ini berpotensi menghambat kemampuan model dalam mengklasifikasikan potensi anak secara lebih akurat, karena tidak adanya variabel yang merepresentasikan elemen khas dari setiap hobi yang dapat menjadi indikator penting dalam klasifikasi minat anak SD.

DAFTAR PUSTAKA

[1] X. Xu and X. Zou, "Predicting Protein–Peptide Complex Structures by Accounting for Peptide Flexibility and the Physicochemical Environment," *J. Chem. Inf. Model.*, vol. 62, no. 1, pp. 27–39, 2022, doi: 10.1021/acs.jcim.1c00836.

[2] S. H. Husin and Y. Yaswinda, "Analisis Pembelajaran Sains Anak Usia Dini di Masa PANDEMI Covid-19," *J. Basicedu*, vol. 5, no. 2, pp. 581–595, 2021, doi: 10.31004/basicedu.v5i2.780.

[3] Y. Bustami, M. Mirnawati, and Y. E. Utami, "Model Pembelajaran Teams Games Tournament: Studi Meta¬-Analisis Berpikir Kritis dan Hasil Belajar Kognitif Siswa Sains," *Biosf. J. Biol. dan Pendidik. Biol.*, vol. 7, no. 1, pp. 30–40, 2022, doi: 10.23969/biosfer.v7i1.5454.

[4] A. S. Pulungan and E. Sulasmi, "Pengembangan Minat dan Bakat Siswa (Studi Kasus Pada Implementasi Kurikulum Merdeka Belajar di SD Negeri 023899 Binjai Timur)," *Jmp-Dmt*, vol. 5, no. 4, pp. 499–509, 2025, doi: 10.30596/jmp-dmt.v5i4.21597.

[5] E. Kustanti, U. Ba'diah, Sinawati, and F. Muhammad, "Klasifikasi Siswa berdasarkan Nilai pada Bidang Ekstrakurikuler Menggunakan Metode K-Nearest Neighbor," *J. Big Data Anal. Artif. Intell.*, vol. 4, no. 1, pp. 1–8, 2018, [Online]. Available: https://journal.ppkia.ac.id/index.php/JBIDAI/article/view/8

[6] H. Firdaus, A. Susilo, and Y. Irawan, "Sistem Pakar untuk Mendeteksi Minat dan Bakat: Sebuah Systematic Literature Review," *J. Ilm. Inform.*, vol. 7, no. 2, pp. 85–94, 2023, doi: 10.35316/jimi.v7i2.85-94.

[7] T. Jannah Cahyaning and H. Susanto Tri, "Mendiagnosa Siswa Yang Berbakat Olahraga Di Sdn Jenggolo Tuban Dengan Naive Bayes," *J. Ilm. Mat.*, vol. 1, no. 7, pp. 1–7, 2018, [Online]. Available: https://ejournal.unesa.ac.id/index.php/mathunesa/article/view/22130

[8] H. Naparin, "Klasifikasi Peminatan Siswa Sma Menggunakan," *Systemic*, vol. 2, no. 1, pp. 25–32, 2016, doi: 10.29080/systemic.v2i1.104.

[9] E. Tangkelobo, W. Mayaut, H. Listanto, and N. F. Sianipar, "Perbandingan Algoritma Klasifikasi Random Forest, Gaussian Naive Bayes, dan K-Nearest untuk Data Tidak Seimbang dan Data yang diseimbangkan dengan metode Random Undersampling pada dataset LCMS Tanaman Keladi Tikus," *Pros. Semin. Nas. Tek. Elektro, Inform. Sist. Inf.*, vol. 2, no. 1, pp. 11–16, 2023, doi: 10.1109/iccmso59960.2023.00033.

[10] D. Trisanto, N. Rismawati, M. Izzatillah, and M. F. Mulya, "Sistem Pakar Identifikasi Bakat Anak Usia Dini Menggunakan Metode Hill Climbing," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 6, no. 4, pp. 750–760, 2022, doi: 10.52362/jisamar.v6i4.904.

[11] D. A. Putri and D. H. Hutasuhut, "Hubungan Pola Asuh Otoriter Orang Tua Dengan Minat Belajar Siswa Kelas Viii Mts. Darul Ilmi Tahun Ajaran 2021/2022," *J. Inov. Penelit.*, vol. 3, no. 3, pp. 5343–5350, 2022, doi: 10.47492/jip.v3i3.1864.

[12] A. Cetinkaya, "Prediction of Middle School Students' Programming Talent Using Artificial Neural Networks," *Eng. Sci. Technol. an Int. J.*, vol. 23, no. 6, pp. 1301–1307, 2020, doi: 10.1016/j.jestch.2020.07.005.

[13] M. Siener, I. Faber, and A. Hohmann, "Prognostic Validity of Statistical Prediction Methods Used for Talent Identification in Youth Tennis Players Based on Motor Abilities," *Appl. Sci.*, vol. 11, no. 15, p. 7051, 2021, doi: 10.3390/app11157051.

[14] S. Jauhiainen and S. Äyrämö, "Talent identification in soccer using a one-class support vector machine," *Int. J. Comput. Sci. Sport*, vol. 18, no. 3, pp. 125–136, 2019, doi: 10.2478/ijcss-2019-0021.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[15] V. Nasir, "Penerapan Algoritma Naive Bayes Classifier Untuk Evaluasi Kinerja Akademik Mahasiswa Universitas Bina Darma," *J. Inform. dan Komput.*, vol. 6, no. 2, pp. 38–39, 2021, doi: 10.26798/jiko.v5i2.227.

[16] E. Poerwandono and J. Perwitosari, "Penerapan Data Mining Untuk Penilaian Kinerja Karyawan Di PT. Riksa Dinar DJaya Menggunakan Metode Naïve Bayes Classification," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 336–340, 2023, doi: 10.30865/mib.v6i1.3331.

[17] I. D. Ratih, S. M. Retnaningsih, and V. M. Dewi, "Klasifikasi Kualitas Tanah Menggunakan Metode Naive Bayes Classifier," *J. Apl. Mat. dan Stat.*, vol. 1, no. 1, pp. 11–20, 2022, doi: 10.53625/jams.v1i1.4427.

[18] S. Choirunnisa, "Metode hibrida oversampling dan undersampling untuk menangani ketidakseimbangan data kegagalan akademik Universitas XYZ," Institut Teknologi Sepuluh Nopember (ITS), 2019. [Online]. Available: https://repository.its.ac.id/60454/

[19] E. Saputro and D. Rosiyadi, "Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes," *Bianglala Inform.*, vol. 10, no. 1, pp. 42–47, 2022, doi: 10.31294/bi.v10i1.11739.

[20] I. S. Ramadhan and A. Salam, "Teknik Random Undersampling untuk Mengatasi Ketidakseimbangan Kelas pada CT Scan Kista Ginjal," *Techno.Com*, vol. 23, no. 1, pp. 20–28, 2024, doi: 10.62411/tc.v23i1.9738.

[21] F. Diapoldo Silalahi, T. Wijanarko, A. Putra, and E. Siswanto, "Machine Learning Technique for Credit Card Scam Detection," *J. Technol. Informatics Eng.*, vol. 1, no. 1, pp. 50–79, Apr. 2022, doi: 10.51903/jtie.v1i1.143.

[22] R. Peranginangin, E. J. G. Harianja, I. K. Jaya, and B. Rumahorbo, "Penerapan Algoritma Safe-Level-Smote Untuk Peningkatan Nilai G-Mean Dalam Klasifikasi Data Tidak Seimbang," *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 4, no. 1, pp. 67–72, 2020, doi: 10.46880/jmika.vol4no1.pp67-72.

[23] R. S. Nurhalizah, R. Ardianto, and P. Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, pp. 61–72, 2024, doi: 10.54082/jiki.168.

[24] F. R. Aftha Harianto, Z. Alawi, and I. A. Sa'ida, "Pengaruh Komposisi Split Data Pada Akurasi Klasifikasi Penderita Diabetes Menggunakan Algoritma Machine Learning," *J. Sist. Inf. dan Inform.*, vol. 8, no. 1, pp. 36–44, 2025, doi: 10.47080/simika.v8i1.3663.

[25] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African J. Biomed. Res.*, vol. 27, no. 4, pp. 4023–4031, 2024, doi: 10.53555/ajbr.v27i4s.4345.