Prosiding Seminar Nasional Ilmu Teknik
Volume 2 Nomor 2 Desember 2025
e-ISSN: 3063-5713; p-ISSN: 3063-4709, Hal 596-609
DOI: https://doi.org/10.61132/prosemnasproit.v2i2.66
Tersedia: https://prosiding.aritekin.or.id/index.php/PROSEMNASPROIT/

Implementasi *Data Mining* dengan Teknik *Smote* dan Fitur *Gain Ratio* Untuk Klasifikasi Kelayakan Siswa Penerima PIP di Kota Jambi

Dea Sabrina Candra[1*], Jasmir Jasmir[2], Elvi Yanti[3]
[1,2,3] Sistem Informasi, Ilmu Komputer, Universitas Dinamika Bangsa, Indonesia
Email: ayaacandra@gmail.com[1], mailto:ijay_jasmir@yahoo.com [2], elvote92@gmail.com[3]*

Alamat: Jl. Jend. Sudirman, The Hok, Kec. Jambi Sel., Kota Jambi, Jambi 36138
[*]Penulis Korespondensi: ayaacandra@gmail.com

Abstract*. The Indonesia Pintar Program (PIP) is an educational assistance program for students from underprivileged families, but determining the eligibility of recipients still faces obstacles in the form of subjectivity and data imbalance. This study aims to classify the eligibility of high school students receiving PIP in Jambi City using data mining methods. The SMOTE technique was applied to overcome class imbalance, and Gain Ratio feature selection was used to determine important attributes. The dataset used consisted of 19,596 student data with a training data distribution of 70% and testing data of 30%. The classification process used the Naïve Bayes, Decision Tree (J48), and Random Forest algorithms with the Use Training Set, 5-Fold, and 10-Fold Cross Validation testing schemes. The results show that SMOTE improves model performance, but feature selection in some cases reduces accuracy. Overall, Random Forest without feature selection provides the best results with an accuracy of 93.33% and is recommended as the most effective model for objectively determining PIP recipient eligibility.*

Keywords*: Data Mining; Classification; SMOTE; Gain Ratio; PIP.*

Abstrak. Program Indonesia Pintar (PIP) merupakan bantuan pendidikan bagi siswa dari keluarga kurang mampu, namun penentuan kelayakan penerimanya masih menghadapi kendala berupa subjektivitas dan ketidakseimbangan data. Penelitian ini bertujuan mengklasifikasikan kelayakan siswa SMA penerima PIP di Kota Jambi menggunakan metode data mining. Teknik SMOTE diterapkan untuk mengatasi ketidakseimbangan kelas dan seleksi fitur *Gain Ratio* digunakan untuk menentukan atribut penting. Dataset yang digunakan berjumlah 19.596 data siswa dengan pembagian *data training* 70% dan *data testing* 30%. Proses klasifikasi menggunakan algoritma Naïve Bayes, Decision Tree (J48), dan *Random Forest* dengan skema pengujian *Use Training Set, 5-Fold, dan 10-Fold Cross Validation*. Hasil penelitian menunjukkan bahwa SMOTE meningkatkan kinerja model, namun seleksi fitur pada beberapa kasus menurunkan akurasi. Secara keseluruhan, *Random Forest* tanpa seleksi fitur memberikan hasil terbaik dengan akurasi 93,33% dan direkomendasikan sebagai model paling efektif dalam menentukan kelayakan penerima PIP secara objektif.

Kata kunci: *Data Mining*; Klasifikasi; SMOTE; *Gain Ratio*; PIP.

1. LATAR BELAKANG

Pendidikan memegang peranan strategis dalam meningkatkan kualitas sumber daya manusia serta menjadi indikator kemajuan suatu bangsa. Namun demikian, kualitas pendidikan di Indonesia hingga kini masih menghadapi berbagai permasalahan, terutama yang berkaitan dengan sistem pendidikan dan pemerataan akses pendidikan. Kesenjangan sosial dan ekonomi menjadi faktor utama yang menyebabkan sebagian masyarakat belum mampu memperoleh pendidikan secara optimal akibat keterbatasan dalam memenuhi biaya pendidikan. Kondisi ini menuntut keterlibatan aktif pemerintah dalam menyediakan program bantuan pendidikan yang berfokus pada pemerataan akses dan peningkatan mutu pendidikan (Torang Siregar et al., 2024).

Naskah Masuk: 14 Oktober 2025; Revisi: 25 Oktober 2025; Diterima: 28 November 2025; Terbit: 30 Desember 2025;

Salah satu langkah yang diambil pemerintah untuk mengatasi permasalahan tersebut adalah melalui Program Indonesia Pintar (PIP). Program ini dirancang untuk meningkatkan mutu pendidikan bagi anak usia sekolah, menekan angka putus sekolah, serta membantu pemenuhan kebutuhan peserta didik dalam kegiatan pembelajaran (Purnama Oktavia & Lestari Anggreini, 2024; Rofiq et al., 2024). Namun, dalam implementasinya PIP masih menghadapi berbagai kendala, khususnya pada tahap penentuan kelayakan penerima bantuan. Proses seleksi yang masih mengandalkan pendekatan administratif dan manual menyebabkan penyaluran bantuan kerap tidak tepat sasaran dan berpotensi diterima oleh siswa yang secara ekonomi tergolong mampu (Edrial et al., 2022). Survei yang dilakukan oleh *Indonesia Corruption Watch* (ICW) pada tahun 2018 menunjukkan bahwa data penilaian kelayakan penerima PIP belum sepenuhnya akurat, di mana 41,9% warga miskin tercatat tidak terdaftar sebagai penerima PIP (Sari Oktapia Ningse et al., 2022). Ketidakakuratan data serta besarnya jumlah data siswa berdampak pada proses pengambilan keputusan yang cenderung subjektif, kurang transparan, dan berisiko menghasilkan keputusan yang tidak optimal.

Perkembangan teknologi informasi, khususnya dalam bidang ilmu komputer, menawarkan alternatif solusi melalui penerapan metode data mining (Aprilyani et al., 2022). Data mining memungkinkan pengolahan data dalam skala besar untuk mengekstraksi pola dan informasi penting secara cepat, objektif, dan transparan, sehingga dapat mendukung proses pengambilan keputusan yang lebih akurat (Simanjuntak et al., 2022). Data mining merupakan proses eksplorasi data untuk menemukan informasi bernilai dan pengetahuan baru dari kumpulan data yang kompleks, yang selanjutnya dapat dimanfaatkan sebagai dasar pengambilan keputusan (Bachtiar & Mahradianur, 2023; Pebdika et al., 2023).

Berbagai algoritma klasifikasi dalam data mining telah banyak digunakan dan terbukti memiliki kinerja yang baik, di antaranya *Naïve Bayes*, *Decision Tree* (J48), dan *Random Forest*. Algoritma *Naïve Bayes* dikenal memiliki kemampuan analisis yang efektif dan efisien dalam mendukung pengambilan keputusan, sedangkan *Decision Tree* (J48) mampu menyajikan hasil klasifikasi dalam bentuk pohon keputusan yang mudah dipahami, meskipun memiliki kecenderungan mengalami overfitting pada data yang tidak seimbang. *Random Forest* sebagai pengembangan dari *Decision Tree* memanfaatkan teknik ensemble dan bagging untuk meningkatkan stabilitas dan akurasi model klasifikasi (Annisa et al., 2025).

Permasalahan yang sering dijumpai pada data bantuan sosial adalah ketidakseimbangan distribusi kelas antara kategori layak dan tidak layak, yang dapat menurunkan performa model klasifikasi apabila tidak ditangani dengan tepat. Oleh karena itu, penelitian ini menerapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi

ketidakseimbangan kelas dengan menghasilkan data sintetis pada kelas minoritas (Sutoyo & Fadlurrahman, 2020). Selain itu, untuk meningkatkan akurasi dan efisiensi model, digunakan teknik seleksi fitur *Gain Ratio Attribute Evaluation* guna mengurangi atribut yang kurang relevan serta meminimalkan bias pada data (Lutfia et al., 2024; Putra et al., 2024).

Sejumlah penelitian sebelumnya telah menerapkan metode klasifikasi data mining dalam penentuan penerima bantuan dengan hasil yang beragam. Penelitian yang mengombinasikan algoritma *Naïve Bayes* dan C4.5 dalam penentuan penerima PIP menghasilkan tingkat akurasi hingga 100% (Amalia et al., 2024), sementara penelitian lain yang menggunakan *Naïve Bayes* memperoleh akurasi sebesar 88,89% (Pebdika et al., 2023). Penelitian yang membandingkan beberapa algoritma klasifikasi juga menunjukkan perbedaan performa, di mana *Naïve Bayes, Decision Tree*, dan *Random Forest* menghasilkan tingkat akurasi yang bervariasi (Annisa et al., 2025). Selain itu, penerapan seleksi fitur *Gain Ratio* dan teknik SMOTE pada penelitian lain terbukti mampu meningkatkan kinerja model klasifikasi (Nugraha et al., 2023; Sandi et al., 2023; Wahab et al., 2019).

Berdasarkan permasalahan tersebut, diperlukan suatu pendekatan berbasis data mining yang mampu mengklasifikasikan kelayakan penerima Program Indonesia Pintar secara objektif dan akurat, khususnya pada data dengan kondisi ketidakseimbangan kelas. Oleh karena itu, penelitian ini mengombinasikan teknik SMOTE dan seleksi fitur Gain Ratio serta membandingkan beberapa algoritma klasifikasi untuk memperoleh model dengan tingkat akurasi terbaik.

2. KAJIAN TEORITIS

Data Mining

Data mining merupakan suatu proses penemuan pola serta informasi yang berharga dari suatu kumpulan data yang besar. Penerapan data mining memungkinkan pengekstrasian data dilakukan untuk mendapatkan pengetahuan yang berguna dari data yang awalnya tidak terstruktur. *Data mining* memiliki tujuan utama untuk mengubah data menjadi suatu informasi yang dapat digunakan sebagai acuan dalam proses pengambilan keputusan (Syahriani Syam, Yokelin Tokoro, Loso Judijanto, Melki Garonga, Frans Mikael Sinaga, Najirah Umar, I Putu Susila Handika, Johar Nur Iin, Apriyanto Apriyanto, 2024).

Klasifikasi

Algoritma klasifikasi merupakan metode yang digunakan untuk mengelompokkan data ke dalam kelas-kelas berdasarkan kategori tertentu. Beberapa contoh algoritma klasifikasi antara lain *Decision Tree, Random Forest*, dan *Naïve Bayes* (Deny Jollyta, William Ramdhan, 2020).

Naïve Bayes

Naïve Bayes merupakan sebuah metode klasifikasi yang didasarkan pada *teorema bayes*, yang dimana konsep dasarnya adalah probabilitas yang memprediksi kemungkinan yang akan muncul di masa depan dengan pengalaman di masa lalu. *Naïve Bayes Classifier* merupakan algoritma yang menggunakan teknik probabilitas ataupun statistik untuk memprediksi (Harahap et al., 2023).

Decision Tree

Decision Tree merupakan metode klasifikasi yang bekerja dengan melakukan berbagai analisis data dengan memecah data dalam bentuk struktur pohon keputusan. *Decision Tree* yang dihasilkan berguna untuk melakukan ekplorasi data untuk mengidentifikasi hubungan tersembunyi antara variabel targe dan komponen variabel yang berbeda Metode ini menggunakan banyak algoritma. Algoritma J48 adalah salah satu metode klasifikasi yang menampilkan hasil pemodelan data dalam bentuk struktur pohon (*tree*) dengan nilai kelas di setiap bagian. Sebuah *roo*t adalah node bagian paling atas dari pohon keputusan (Pakpahan, 2021).

Random Forest

Random Forest merupakan algoritma yang dicetuskan oleh J.Ross Quinlan, *Random Forest* adalah turunan dari pendekatan ID3 untuk membangun pohon keputusan. *Random Forest* cocok digunakan untuk pengklasifikasian masalah pada *data mining* dan pembelajaran mesin. *Random Forest* menggabungkan atribut kelas untuk menemukan prediksi untuk data yang belum terlihat (Surono & Pusparini, 2020).

SMOTE (*Synthetic Minority Oversampling Technique)*

SMOTE *(Synthetic Minority Oversampling Technique)* merupakan metode *oversampling* yang digunakan untuk mengatasi ketidakseimbangan data. Penggunaan SMOTE bertujuan untuk menyeimbangkan kelas dengan menaikkan jumlah kelas minoritas secara acak dengan membuat data sintetis (Erlin et al., 2022).

Gain Ratio

Gain Ratio merupakan teknik seleksi fitur untuk memilih fitur terbaik diantara fitur-fitur lainnya. Penerapan seleksi fitur ini diharapkan dapat meningkatkan akurasi dari proses klasifikasi (Grandis et al. 2021).

K-Fold Cross Validation

Metode validasi *K-Fold* adalah cara untuk menguji akurasi. dari prediksi yang telah dilakukan berdasarkan atribut-atribut yang digunakan untuk mengevaluasi dan memastikan apakah model dapat melakukan pengukuran hasil yang stabil terhadap kinerjanya dengan tujuan untuk menghilangkan keraguan pada data. Metode ini bekerja dengan memecah atau membagi data ke dalam k bagian set yang sama dan melakukan instruksi dan pengujian terhadap data sebanyak k kali (Tempola et al., 2018).

Program Indonesia Pintar

Program Indonesia Pintar adalah bentuk inisiatif pemerintah Indonesia yang menawarkan bantuan keuangan yang diberikan kepada pelajar yang berasal dari latar belakang ekonomi keluarga kurang mampu. Bertujuan untuk meningkatkan serta memberikan akses pendidikan yang merata dan berkualitas. Program Indonesia Pintar ini diperkerkenalkan pertama kali pada tahun 2014 dengan dibuatnya Peraturan Presiden No. 7 Tahun 2014. Dalam Peraturan Kementerian Pendidikan dan Kebudayaan (Kemendikbud) ditugaskan untuk membuat PIP dengan memberikan Kartu Indonesia Pintar (KIP) untuk membantu siswa yang hidup dengan biaya rendah mendapatkan pendidikan yang layak (Tesa Vausia Sandiva et al., 2024).

3. METODE PENELITIAN

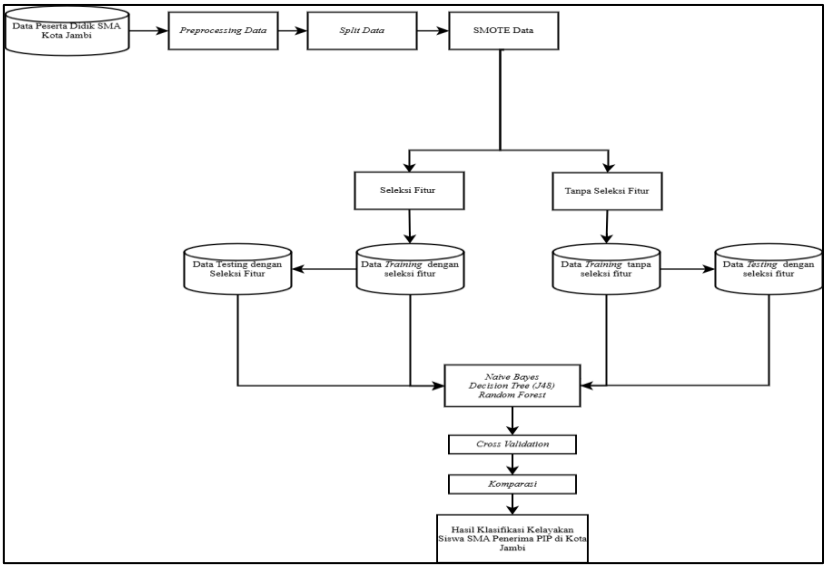
Penelitian ini menggunakan desain penelitian kuantitatif dan pendekatan eksperimen. Penelitian kuantitatif merupakan metodologi penelitian yang menggunakan teknik ilmiah untuk pengumpulan data, yang dilanjutkan dengan analisis statistik serta pembuatan kesimpulan dari hasil yang ditemukan (Susanto et al., 2024). Penelitian ini mengevaluasi kinerja model klasifikasi dalam menentukan apakah siswa SMA yang menerima Program Indonesia Pintar (PIP) layak untuk berpartisipasi berdasarkan data historis yang tersedia. Populasi penelitian mencakup seluruh data siswa SMA calon penerima PIP di Kota Jambi, dengan sampel berupa data siswa yang diperoleh dari BPMP Provinsi Jambi dan disesuaikan dengan kebutuhan penelitian. Data dikumpulkan melalui pengumpulan dokumen berupa dataset digital yang memuat atribut penentu kelayakan penerima PIP. Analisis data dilakukan

menggunakan perangkat lunak WEKA melalui tahapan *preprocessing*, pembagian data menjadi data pelatihan dan pengujian, penerapan teknik SMOTE untuk mengatasi ketidakseimbangan kelas, serta seleksi fitur. Dataset hasil pengolahan selanjutnya dianalisis menggunakan algoritma *Naïve Bayes, Decision Tree* (J48), dan *Random Forest*, dengan evaluasi kinerja model dilakukan menggunakan metode *K-Fold Cross Validation* dan indikator akurasi.

Proses klasifikasi kelayakan siswa penerima PIP digambarkan dalam kerangka kerja penelitian ini dapat dilihat pada gambar 1:

Gambar 1. Kerangka Kerja Penelitian

Selanjutnya, rancangan eksperimen yang digunakan untuk mengevaluasi kinerja model klasifikasi pada berbagai skenario pengujian ditunjukkan pada Gambar 2:



Gambar 2. Rancangan Eksperimen

4. HASIL DAN PEMBAHASAN

Pada tahapan proses ini dilakukan melalui tahap *pre-processing* data untuk memastikan data siap diolah menggunakan teknik data mining dengan menerapkan metode klasifikasi *Naïve Bayes*, *Decision Tree* (J48), dan *Random Forest*. Metode tersebut selanjutnya dikombinasikan dengan teknik SMOTE untuk penyeimbangan kelas serta seleksi fitur *Gain Ratio* guna meningkatkan akurasi model. Hasil analisis yang diperoleh kemudian diinterpretasikan

menjadi informasi yang menggambarkan tingkat kelayakan siswa dalam menerima Program Indonesia Pintar (PIP).

Analisis Data Penelitian.

Representasi Data

Daftar atribut yang digunakan dalam penelitian ini mengacu pada kriteria data siswa jenjang SMA yang terdiri dari 16 atribut. Rincian atribut tersebut dapat dilihat pada Tabel 1:

Tabel 1. Daftar Atribut

No	Nama	Tipe Data	Keterangan
1.	Jenis Kelamin	*Binomial*	Jenis Kelamin Siswa Dibagi Menjadi 2 : P Dan L
2.	Kebutuhan Khusus	*Polynominal*	Kebutuhan Khusus Yang Dimiliki Oleh Siswa
3.	Jenis Tinggal	*Polynominal*	Jenis Tinggal Atau Menetap Siswa
4.	Alat Transportasi	*Polynominal*	Kendaraan Atau Transportasi Yang Digunakan.
5.	Anak Keberapa	*Polynominal*	Siswa Merupakan Anak Keberapa
6.	Jenjang Pendidikan Ayah	*Polynominal*	Tingkat Pendidikan Orang Tua (Ayah)
7.	Pekerjaan Ayah	*Polynominal*	Jenis Pekerjaan Orang Tua Laki-Laki Dari Peserta Didik
8.	Penghasilan Ayah	*Polynominal*	Jumlah Penghasilan Orang Tua Siswa Per Bulan
9.	Kebutuhan Khusus Ayah	*Polynominal*	Kondisi Khusus Yang Diidap Orang Tua Peserta Didik
10.	Jenjang Pendidikan Ibu	*Polynominal*	Tingkat Pendidikan Orang Tua (Ibu)
11.	Pekerjaan Ibu	*Polynominal*	Jenis Pekerjaan Orang Tua Laki-Laki Dari Peserta Didik
12.	Penghasilan Ibu	*Polynominal*	Jumlah Penghasilan Orang Tua Siswa Per Bulan
13.	Kebutuhan Khusus Ibu	*Polynominal*	Kondisi Khusus Yang Diidap Orang Tua Peserta Didik
14.	Alasan Layak PIP	*Polynominal*	Faktor Penyebab Seorang Siswa Dinyatakan Layak Menerima Program Indonesia Pintar (PIP)
15.	Layak PIP	*Binomial*	Status dibagi menjadi 2 : Ya dan Tidak

Pre-Processing Data

Tahap pre-processing dilakukan sebelum proses data mining dengan tujuan mengatasi permasalahan pada data yang dapat memengaruhi kinerja model klasifikasi. Pada tahap ini, data awal diperiksa untuk memastikan kelengkapan dan konsistensi atribut, mengingat masih terdapat data dengan format tidak seragam serta nilai noise pada beberapa atribut utama. Proses pre-processing meliputi pengecekan kelengkapan data, penanganan nilai hilang menggunakan

metode *Replace Missing Value*, serta pembersihan data (*data cleaning*) untuk menghasilkan dataset yang siap digunakan pada tahap mining. Pre-processing dilakukan menggunakan perangkat lunak WEKA, di mana ditemukan nilai hilang pada beberapa atribut, antara lain alasan kelayakan PIP, pekerjaan dan penghasilan orang tua, serta jenjang pendidikan ibu. Dataset hasil pre-processing selanjutnya digunakan dalam proses klasifikasi menggunakan algoritma Naïve Bayes, Decision Tree (J48), dan *Random Forest*.

Proses *Split* Data

Data splitting atau pemisahan data merupakan teknik untuk membagi data ke dalam dua atau lebih bagian yang membentuk subhimpunan data. Umumnya, proses ini membagi data menjadi dua bagian, yaitu satu bagian digunakan untuk pelatihan model dan bagian lainnya digunakan untuk pengujian atau evaluasi. Alur proses split data dapat dilihat pada tahapan berikut.

a. Data *Training*

Setelah melalui tahapan proses *pre-processing* data, data akan Data dibagi menjadi dua bagian dengan perbandingan 70% sebagai *data Training* dan 30% sebagai *data Testing*. Selanjutnya adalah tahap pengambilan 70% data *Training* dari data Siswa SMA Kota Jambi yang sudah melalui tahapan *cleaning* data.

Tabel 2. Proses *Split Data Training 70%* dengan *Tools* WEKA sebelum SMOTE

***Remove Percentage*-P 70.0 (*InvertSelection 'True'*)**		
Attributes	15	
Instances	13717	
Sum of Weight	13717	
Nama Atribut	**Kelas**	
	Ya	**Tidak**
Score	5496	8221

Data training yang telah melalui tahap *data cleaning* atau *pre-processing* disusun sesuai kebutuhan penelitian dan disimpan dalam format Excel sebelum dikonversi ke format CSV. Proses pre-processing meliputi pembersihan data dan penghilangan noise akibat data ganda agar dataset siap digunakan pada tahap analisis. Selanjutnya, dilakukan pembagian data dengan proporsi 70% sebagai data training, sehingga diperoleh sebanyak 13.717 data yang digunakan untuk pelatihan model klasifikasi.

b. Data *Testing*

Pada proses perhitungan dengan menggunakan *Naïve Bayes*, *Decision Tree (J48)*, *Randim Forest* menggunakan data *Testing.* Data *Testing* yang akan digunakan pada penelitian ini diambil dari 30% jumlah data Siswa SM Kota Jambi yang sudah melalui tahapan *cleaning* yaitu sebanyak 5879 data. Berikut Visualisasi Hasil dari 30% jumlah data:

Tabel 3. Proses *Split Data Testing 70%* dengan *Tools* WEKA sebelum SMOTE

***Remove Percentage*-P 70.0 (*InvertSelection 'False'*)**		
Attributes	15	
Instances	5879	
Sum of Weight	5879	
Nama Atribut	**Kelas**	
	Ya	**Tidak**
Score	3526	2353

Dari data yang sudah melalui tahapan *cleaning* hasil dari 30% pengambilan untuk data *Testing* adalah 5879 data.

Penerapan SMOTE untuk Balancing Data

Pada tahap ini dilakukan penyeimbangan data menggunakan teknik Synthetic Minority Oversampling Technique (SMOTE) pada data training. Penerapan SMOTE didasarkan pada hasil analisis awal yang menunjukkan ketidakseimbangan distribusi kelas, di mana jumlah data siswa dengan label "Ya" (layak) lebih sedikit dibandingkan kelas "Tidak" (tidak layak). Ketidakseimbangan tersebut berpotensi menurunkan kinerja model klasifikasi karena dominasi kelas mayoritas, sehingga SMOTE digunakan untuk meningkatkan representasi kelas minoritas dan menghasilkan model yang lebih optimal.

Tabel 3. Hasil *Data Training* Setelah SMOTE

***Remove Percentage*-P 70.0 (*InvertSelection 'True'*)**		
Attributes	15	
Instances	19213	
Sum of Weight	19213	
Nama Atribut	**Kelas**	
	Ya	**Tidak**
Score	8221	10992

Penyeimbangan data *training* dilakukan menggunakan SMOTE. Sebelum penyeimbangan, jumlah data sebanyak 13.717 instance dengan distribusi kelas "Ya" 5.496 dan "Tidak" 8.221. Setelah SMOTE, jumlah *instance* meningkat menjadi 19.213, dengan kelas "Ya" bertambah menjadi 10.992 sementara kelas "Tidak" tetap 8.221, sehingga distribusi kelas menjadi lebih seimbang.

Tabel 4. Hasil *Data Testing* Setelah Data Training di SMOTE

***Remove Percentage*-P 70.0 (*InvertSelection 'False'*)**		
Attributes	15	
Instances	5879	
Sum of Weight	5879	
Nama Atribut	**Kelas**	
	Ya	**Tidak**
Score	3526	2353

Pada hasil data *Testing* tidak terjadi perubahan jumlah *Instance* dan jumlah kelas "Ya" dan kelas "Tidak", karena SMOTE hanya dilakukan pada data *Training* saja.

Seleksi Fitur *Gain Ratio*

Berdasarkan hasil yang didapatkan menggunakan seleksi fitur *Gain Ratio* maka pada penelitian ini peneliti mengambil atribut dengan bobot >0.005. Atribut atau fitur yang diambil didasarkan juga pada faktor kelayakan pemberian bantuan PIP. Adapun atribut yang terpilih sebagai berikut:

Penghasilan Ayah 0.10814, Penghasilan Ibu 0.073112, Pekerjaan Ayah 0.072077, Jenjang Pendidikan Ayah 0.061124, Pekerjaan Ibu 0.057714, Jenjang Pendidikan Ibu 0.056253, Alat Transportasi 0.041883, Kebutuhan Khusus 0.038784, Kebutuhan Khusus Ayah 0.028264, Alasan layak PIP 0.026455, Kebutuhan Khusus Ibu 0.020721, Jenis Tinggal 0.01470.

Hasil Evaluasi Algoritma

Hasil Evaluasi Algoritma Tanpa Seleksi Atribut

Pengujian klasifikasi menggunakan algoritma Naïve Bayes, Decision Tree (J48), dan *Random Forest* dilakukan dengan bantuan perangkat lunak WEKA melalui skema *Use Training Set*, *5-Fold Cross-Validation*, dan *10-Fold Cross-Validation*. Hasil pengujian tersebut menghasilkan perbandingan performa ketiga algoritma tanpa penerapan seleksi atribut pada data *training* dan data *testing*, yang selanjutnya disajikan pada Gambar 3 berikut:

	NAÏVE BAYES						*DECISION TREE (J48)*						*RANDOM FOREST*					
	70% Training Set			30% Testing Set			70% Training Set			30% Testing Set			70% Training Set			30% Testing Set		
Performance	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	
Accuracy %	83.636	83.537	83.542	81.918	82.122	81.952	82.454	79.701	79.732	81.952	77.104	90.834	80.325	80.596	93.332	75.948	75.401	
Precision %	0.892	0.889	0.889	0.714	0.717	0.715	0.798	0.798	0.798	0.844	0.769	0.910	0.803	0.805	0.933	0.757	0.753	
Recall %	0.704	0.703	0.703	0.915	0.915	0.914	0.797	0.797	0.797	0.820	0.771	0.908	0.803	0.806	0.933	0.759	0.755	
F-Measure	0.786	0.703	0.785	0.802	0.804	0.802	0.795	0.794	0.795	0.821	0.769	0.908	0.803	0.806	0.933	0.756	0.752	

Gambar 3. Hasil Performa *Naïve Bayes Decision Tree* (J48), dan *Random Forest* Data *Training* dan testing tanpa seleksi atribut

Berdasarkan hasil pengujian akurasi pada ketiga algoritma, untuk data *Training*, nilai akurasi tertinggi diperoleh dengan metode *Use Training Set***,** yaitu 83,64% untuk *Naïve Bayes*, 82,45% untuk *Decision Tree* (J48), dan 90,83% untuk *Random Forest*. Sedangkan pada data *Testing*, akurasi terbaik bervariasi tergantung metode yang digunakan, dengan *Random Forest* menggunakan *Use Training Set* mencapai akurasi tertinggi sebesar 93,33%, diikuti *Naïve Bayes* dengan 82,12% menggunakan *Use Training Set 5-Fold Cross Validation*, dan *Decision Tree* sebesar 81,95% dengan. Hasil ini menunjukkan bahwa *Random Forest* memiliki performa akurasi terbaik secara keseluruhan, baik pada data *Training* maupun *Testing*.

Hasil Evaluasi Algoritma Tanpa Seleksi Atribut

Pengujian klasifikasi menggunakan algoritma *Naïve Bayes, Decision Tree* (J48), dan *Random Forest* dilakukan dengan bantuan perangkat lunak WEKA melalui skema *Use Training Set*, *5-Fold Cross-Validation*, dan *10-Fold Cross-Validation*. Hasil pengujian tersebut menghasilkan perbandingan performa ketiga algoritma dengan penerapan seleksi atribut pada data *training* dan data *testing*, yang selanjutnya disajikan pada Gambar 4 berikut:

	NAÏVE BAYES						*DECISION TREE (J48)*						*RANDOM FOREST*					
	70% Training Set			30% Testing Set			70% Training Set			30% Testing Set			70% Training Set			30% Testing Set		
Performance	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation	Use Training Set	5 *Fold* Cross Validation	10 *Fold* Cross Validation
Accuracy %	83.682	83.610	83.630	81.969	81.987	81.987	80.237	79.316	79.503	79.588	77.360	77.734	87.341	80.200	80.143	89.556	77.122	77.173
Precision %	0.844	0.843	0.844	0.844	0.843	0.843	0.804	0.794	0.796	0.794	0.772	0.776	0.875	0.802	0.801	0.895	0.769	0.770
Recall %	0.837	0.836	0.836	0.820	0.819	0.819	0.802	0.793	0.795	0.796	0.774	0.777	0.873	0.803	0.801	0.896	0.771	0.772
F-Measure	0.833	0.832	0.833	0.821	0.821	0.821	0.800	0.790	0.792	0.794	0.772	0.776	0.872	0.801	0.800	0.895	0.769	0.770

Gambar 4. Hasil Performa *Naïve Bayes Decision Tree (J48)*, dan *Random Forest* Data Training dan testing dengan seleksi atribut

Hasil pengujian algoritma *Naïve Bayes, Decision Tree* (J48), dan *Random Forest* menunjukkan *Random Forest* sebagai yang terbaik. Akurasi tertinggi pada data *Training* 70% diperoleh *Random Forest* dengan *Use Training Set*, yaitu 87,34%. *Naïve Bayes* mencapai 83%, sedangkan *Decision Tree* 80%. Pola ini menunjukkan *Use Training Set* umumnya lebih akurat daripada *Cross Validation*, khususnya untuk *Random Forest*. Pada data *Testing*, *Random Forest* dengan *5 Fold Cross Validation* memperoleh akurasi terbaik, 77,12%, meski lebih rendah dibanding data *Training*. Kesimpulannya, *Random Forest* unggul dalam prediksi dibanding kedua algoritma lainnya.

5. KESIMPULAN DAN SARAN

Hasil penelitian menunjukkan bahwa metode *Naïve Bayes, Decision Tree* (J48), dan *Random Forest* dengan SMOTE dan seleksi fitur *Gain Ratio* dapat digunakan untuk mengklasifikasikan kelayakan siswa SMA penerima PIP di Kota Jambi, dengan *Random Forest* sebagai algoritma paling akurat dan stabil. Seleksi fitur *Gain Ratio* pada penelitian ini justru menurunkan akurasi karena menghilangkan informasi penting, sehingga *Random Forest* tanpa seleksi fitur menjadi model terbaik dan direkomendasikan untuk penentuan kelayakan secara objektif. Penelitian selanjutnya disarankan menguji algoritma dan model lain, menambah atribut sosial ekonomi, memperbarui data secara berkala, serta mengembangkan sistem menjadi Sistem Pendukung Keputusan berbasis aplikasi atau web.

DAFTAR REFERENSI

Amalia, A., Irma Purnamasari, A., & Ali, I. (2024). Implementasi Algoritma C4.5 Dan Naïve Bayes Dalam Pengambilan Keputusan Untuk Program Indonesia Pintar (Pip) Di Sekolah Dasar Negeri 04 Majalangu. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1889–1896. https://doi.org/10.36040/jati.v8i2.8311

Annisa, R., Aprizal, D., Dewi, R. K., Ayuandita, D. S., Oktaviani, A., & Azzahra, M. (2025). Perbandingan Algoritma Naïve Bayes, Decision Tree, KNN, dan Random Forest Untuk Memprediksi Data Penduduk Penerima BPJS Di Lampung Timur. *Jurnal of Data Science Methods and Applications*, *01*(01), 35132. https://doi.org/10.30873/jodmapps.v1i1.pp42-49

Aprilyani, N., Zulfa, I., & Syahputra, H. (2022). Penerapan Algoritma Decision Tree C4.5 Untuk Model Penentuan Penerima Beasiswa Program Indonesia Pintar (Pip) Studi Kasus Sma Negeri 3 Timang Gajah. *Jurnal Teknik Informatika Dan Elektro*, *5*(1), 96–109. https://doi.org/10.55542/jurtie.v5i1.452

Bachtiar, L., & Mahradianur, M. (2023). Analisis Data Mining Menggunakan Metode Algoritma C4.5 Menentukan Penerima Bantuan Langsung Tunai. *Jurnal Informatika*, *10*(1), 28–36. https://doi.org/10.31294/inf.v10i1.15115

Deny Jollyta, William Ramdhan, M. Z. (2020). *Konsep Data Mining Dan Penerapan*. Deepublish.

Edrial, Putrama rangga, & Sujastiawan Ade. (2022). *Evaluasi Kebijakan Program Indonesia Pintar (Pip) Di Sma*. *3*(1), 109–118.

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 677–690. https://doi.org/10.30812/matrik.v21i3.1726

Grandis, G. F., Arumsari, Y., & Indriati. (2021). Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *5*(8), 3507–3514.

Harahap, F., Fahrozi, W., Adawiyah, R., Siregar, E. T., & Harahap, A. Y. N. (2023). Implementasi Data Mining dalam Memprediksi Produk AC Terlaris untuk Meningkatkan Penjualan Menggunakan Metode Naive Bayes. *Jurnal Unitek*, *16*(1), 41–51. https://doi.org/10.52072/unitek.v16i1.541

Lutfia, A., Gunawan, G., Rohman, R. S., & Gunawan, A. (2024). Penerapan Seleksi Fitur Gain Ratio Pada Prediksi Penyakit Jantung Berbasis Naïve Bayes. *Jurnal Responsif : Riset Sains Dan Informatika*, *6*(1), 1–10. https://doi.org/10.51977/jti.v6i1.1396

Nugraha, M. A., Mazdadi, M. I., Farmadi, A., Muliadi, & Saragih, T. H. (2023). Penyeimbangan Kelas SMOTE dan Seleksi Fitur Ensemble Filter pada Support Vector Machine untuk Klasifikasi Penyakit Liver. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(6), 1273–1284. https://doi.org/10.25126/jtiik.2023107234

Pakpahan, N. S. (2021). Implementasi Data Mining Menggunakan Algoritma J48 Dalam Menentukan Pola Itemset Belanja Pembeli (Study Kasus: Swalayan Brastagi Medan). *Journal of Computing and Informatics Research*, *1*(1), 7–13.

Pebdika, A., Herdiana, R., & Solihudin, D. (2023). Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima Pip. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 452–458. https://doi.org/10.36040/jati.v7i1.6303

Purnama Oktavia, I., & Lestari Anggreini, N. (2024). Implementasi Algoritma Naive Bayes Dengan Metode Klasifikasi Dalam Menentukan Siswa Penerima Bantuan Program Indonesia Pintar. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(6), 11152–11158. https://doi.org/10.36040/jati.v8i6.11241

Putra, I. M. A. A. D., Sunarya, I. M. G., & Gunadi, I. G. A. (2024). Perbandingan Algoritma Naive Bayes Berbasis Feature Selection Gain Ratio dengan Naive Bayes Kovensional dalam Prediksi Komplikasi Hipertensi. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *6*(1), 37–49. https://doi.org/10.35746/jtim.v6i1.488

Rofiq, M. A., Kurniati, N., & Surya Editya, A. (2024). Klasifikasi Kelayakan Data Beasiswa PIP Pada MINU Sumokali Menggunakan Metode Decision Tree. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *11 No 1*(x), 1–5.

Sandi, S. A., Novianto, Y., & Sandi, S. A. (2023). Klasifikasi Kelayakan Keluarga Penerima Bantuan Langsung Tunai Menggunakan Gain Ratio Dan Naïve Bayes Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM). *Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)*, *3*(April), 433–442.

Sari Oktapia Ningse, W. R., Sumarno, S., & Nasution, Z. M. (2022). C4.5 Algorithm Classification for Determining Smart Indonesia Program Recipients at MIS Al-Khoirot. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(1), 65–76. https://doi.org/10.55123/jomlai.v1i1.165

Simanjuntak, A. Y., Simatupang, I. S. S., & Anita. (2022). Implementasi Data Mining Menggunakan Metode Naïve Bayes Classifier Untuk Data Kenaikan Pangkat Dinas. *Journal of Science and Social Research*, *4307*(1), 85–91.

Surono, G., & Pusparini, N. N. (2020). Journal of technology information. *Jurnal Of Technology Information*, *5*(2), 99–104.

Susanto, P. C., Arini, D. U., Yuntina, L., & Panatap, J. (2024). *Konsep Penelitian Kuantitatif : Populasi , Sampel , dan Analisis Data (Sebuah Tinjauan Pustaka)*. *3*(1), 1–12.

Sutoyo, E., & Fadlurrahman, M. A. (2020). Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(3), 379. https://doi.org/10.26418/jp.v6i3.42896

Syahriani Syam, Yokelin Tokoro, Loso Judijanto, Melki Garonga, Frans Mikael Sinaga, Najirah Umar, I Putu Susila Handika, Johar Nur Iin, Apriyanto Apriyanto, A. T. S. (2024). *Data Mining : Teori dan Penerapannya dalam Berbagai Bidang* (S. S. Efitra Efitra , Elok Pamela (ed.)). PT. Sonpedia Publishing Indonesia.

Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(5), 577–584. https://doi.org/10.25126/jtiik.201855983

Tesa Vausia Sandiva, Defit, S., & Nurcahyo, G. W. (2024). Implementasi Algoritma C4.5 Untuk Prediksi Penerima Beasiswa Program Indonesia Pintar. *Jurnal KomtekInfo*, *11*(4), 354–362. https://doi.org/10.35134/komtekinfo.v11i4.582

Torang Siregar, Riski Ardian, Ahmad Arisman, & Iskandarsyah. (2024). Studi Kasus Sma N 1 Sinunukan : Implementasi Algoritma K-Nearest Neighbor Untuk Klasifikasi Penerima Beasiswa Program Indonesia Pintar (Pip). *Jurnal Cermatika*, *4*(1), 9–25. https://doi.org/10.64168/cermatika.v4i1.1324

Wahab, A., Samarinda, S., Lishania, I., Goejantoro, R., & Nasution, Y. N. (2019). Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Comparison of the Classification for Naive Bayes Method and the Decision Tree Algorithm (J48) for Stroke Patients in Abdul W. *Jurnal EKSPONENSIAL*, *10*(2), 135–142.