Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 4 October 2025, pp: 1450-1462
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Classification of Scholarship Eligibility Using Naïve Bayes with Attribute Optimization Based on K-Means Clustering

Klasifikasi Kelayakan Penerima Beasiswa Menggunakan Naive Bayes dengan Optimasi Atribut Berbasis K-Means *Clustering*

Azhiah Putri[1*], Jasmir[2], Benni Purnama[3]

[1,2,3]Magister of Information System, Universitas Dinamika Bangsa, Indonesia

E-Mail: [1]putriazhiah96@gmail.com, [2]ijay_jasmir@yahoo.com, [3]bennipurnama@unama.ac.id

Received Aug 20th 2025; Revised Oct 05th 2025; Accepted Oct 25th 2025; Available Online Oct 31th 2025
Corresponding Author: Azhiah Putri

Abstract

This study aims to classify the eligibility of scholarship recipients in Tabir District using the Naïve Bayes algorithm optimized through K-Means Clustering. The dataset, consisting of 4,155 student records, was processed through several preprocessing stages, including relevant attribute selection, data cleaning, and categorical-to-numerical transformation. The clustering process was carried out using K-Means with K = 2, 3, and 5, and evaluated using the Davies-Bouldin Index (DBI). The best result was obtained at K = 2 with a DBI value of 0.909, which was then used to group the data into two clusters, namely "Eligible" and "Not Eligible." The resulting clusters were used to optimize the attributes in the classification stage using the Naïve Bayes algorithm. Model performance was evaluated using the confusion matrix under three validation schemes: data split ratios of 70:30, 80:20, and 10-fold cross-validation. The results showed accuracies of 96.15%, 96.75%, and 97.91%; precisions of 99.90%, 99.85%, and 99.97%; recalls of 95.39%, 96.18%, and 97.47%; and F1-scores of 97.60%, 97.97%, and 98.71%, respectively. Based on these results, the 10-fold cross-validation method achieved the best performance by maintaining a balance between accuracy, precision, recall, and F1-score. Therefore, the integration of K-Means Clustering and Naïve Bayes proved to be effective in attribute optimization, resulting in an accurate, consistent, and reliable classification system to support scholarship recipient selection decisions.

Keyword: Confusion Matrix, Davies-Bouldin Index, K-Means, Naïve Bayes, Scholarship

Abstrak

Penelitian ini bertujuan untuk mengklasifikasikan kelayakan penerima beasiswa di Kecamatan Tabir dengan menggunakan algoritma Naïve Bayes yang dioptimasi melalui K-Means *Clustering*. *Dataset* berjumlah 4.155 siswa diproses melalui tahap pra-pemrosesan, mencakup seleksi atribut relevan, pembersihan data, serta transformasi kategori ke bentuk numerik. Proses *clustering* dilakukan dengan K-Means pada K = 2, 3, dan 5, lalu dievaluasi menggunakan *Davies-Bouldin Index* (DBI). Hasil terbaik diperoleh pada K = 2 dengan nilai DBI = 0,909, yang selanjutnya digunakan untuk mengelompokkan data menjadi dua klaster, yaitu "Layak" dan "Tidak Layak". Klaster yang dihasilkan kemudian digunakan untuk mengoptimasi atribut pada tahap klasifikasi menggunakan algoritma Naïve Bayes.. Evaluasi performa menggunakan *confusion matrix* dengan skema *split data* 70:30, 80:20, dan *10-fold cross validation*. Hasil menunjukkan akurasi masing-masing 96,15%, 96,75%, dan 97,91%; *precision* 99,90%, 99,85%, dan 99,97%; *recall* 95,39%, 96,18%, dan 97,47%; serta *F1-score* 97,60%, 97,97%, dan 98,71%. Berdasarkan hasil tersebut, metode *10-fold cross validation* memberikan performa terbaik karena mampu menjaga keseimbangan antara akurasi, *precision*, *recall*, dan *F1-score*. Dengan demikian, integrasi antara K-Means *Clustering* dan Naïve Bayes terbukti efektif dalam mengoptimasi atribut, serta menghasilkan sistem klasifikasi yang akurat, konsisten, dan andal untuk mendukung keputusan seleksi penerima beasiswa.

Kata Kunci: Beasiswa, *Confusion Matrix, Davies-Bouldin Index,* K-Means, Naïve Bayes

1. PENDAHULUAN

Pendidikan merupakan salah satu elemen penting dalam pembangunan bangsa, karena berperan dalam membentuk pengetahuan, keterampilan, dan karakter individu [1]. Pada jenjang pendidikan dasar, proses pembentukan karakter dan pondasi intelektual dimulai, sehingga akses pendidikan yang merata dan berkualitas menjadi sangat krusial [2]. Salah satu upaya pemerintah dalam meningkatkan akses dan kualitas pendidikan dasar adalah melalui program beasiswa. Beasiswa merupakan bantuan biaya pendidikan yang diberikan kepada peserta didik oleh lembaga pemerintah, swasta, maupun organisasi non-profit, dengan tujuan mendukung keberlanjutan proses pembelajaran. Beasiswa biasanya diberikan berdasarkan kriteria tertentu, seperti prestasi akademik, kondisi ekonomi, dan latar belakang sosial [3]

Salah satu program beasiswa yang cukup besar cakupannya di Indonesia adalah Program Indonesia Pintar (PIP). PIP memberikan bantuan tunai pendidikan kepada siswa jenjang SD hingga SMA/SMK yang memenuhi syarat ekonomi dan sosial tertentu. PIP merupakan bentuk komitmen pemerintah dalam menekan angka putus sekolah dan mendukung pemerataan pendidikan dasar [3]. Melalui program beasiswa tersebut, siswa dapat memenuhi kebutuhan pendidikan mereka, seperti buku, alat tulis, tas, sepatu, seragam sekolah, dan perlengkapan lainnya. Namun, dalam implementasinya, penentuan penerima PIP masih sering menemui permasalahan.

Namun dalam implementasinya, PIP masih menghadapi berbagai permasalahan di lapangan. Penelitian di empat daerah kunjungan kerja Presiden pada tahun 2017 mencatat adanya *inclusion error* dan *exclusion error*, di mana siswa yang tidak memenuhi kriteria tercatat sebagai penerima, sedangkan siswa yang layak justru tidak terakomodasi. Hal ini terkait dengan lemahnya sistem verifikasi data serta belum optimalnya integrasi data antar instansi [4]. Masalah lain yang muncul adalah proses seleksi penerima beasiswa di tingkat sekolah dasar masih banyak dilakukan secara manual, mengandalkan rekomendasi guru atau pengisian formulir tanpa analisis data secara mendalam. Hal ini berpotensi menimbulkan ketidaktepatan sasaran, yaitu siswa yang seharusnya layak tidak mendapat bantuan, sedangkan siswa yang kurang layak justru menerima beasiswa. Seleksi manual juga rawan bias, kurang efisien, dan sulit diterapkan pada jumlah siswa yang besar [5].

Untuk mengatasi hal tersebut, diperlukan pendekatan berbasis teknologi, salah satunya dengan *data mining*. *Data mining* merupakan proses mengekstraksi informasi yang berguna dari sejumlah besar data dengan memanfaatkan metode statistik, matematika, serta kecerdasan buatan [6]. Dalam penelitian ini digunakan kombinasi metode K-Means *Clustering* dan Naïve Bayes *Classification*. Algoritma K-Means *Clustering* adalah salah satu teknik pengelompokan data yang termasuk dalam kategori *unsupervised learning*. Algoritma ini bertujuan untuk membagi sekumpulan data ke dalam sejumlah klaster (*k*) berdasarkan kemiripan fitur antar data. Setiap data akan dimasukkan ke dalam satu klaster yang memiliki pusat (*centroid*) terdekat. Proses ini dilakukan secara iteratif untuk meminimalkan jarak antara data dan pusat klaster, sehingga menghasilkan pengelompokan yang optimal. Kelebihan dari K-Means adalah kesederhanaan dan efisiensinya dalam menangani *dataset* berukuran besar [7]. Dalam konteks ini, K-Means digunakan untuk mengelompokkan siswa ke dalam kategori "layak" dan "tidak layak" menerima beasiswa berdasarkan atribut-atribut seperti alat transportasi, penerima KPS, pekerjaan ayah penghasilan ayah, pekerjaan ibu, penghasilan ibu, layak PIP, kebutuhan khusus, jumlah saudara kandung, dan jarak rumah ke sekolah.

Sementara itu, Naïve Bayes *Classification* merupakan metode *supervised learning* yang didasarkan pada Teorema Bayes. Algoritma ini bekerja dengan menghitung probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan nilai atribut yang dimilikinya. Naïve Bayes dikenal memiliki kemampuan klasifikasi yang cepat dan akurat, terutama pada data berskala besar dan dengan fitur yang bersifat independen [8]. Selain itu Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya [9]. Keuntungan penggunaan Naive Bayes adalah metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan [10].

Untuk meningkatkan akurasi hasil klasifikasi, penelitian ini menggabungkan dua algoritma, yaitu K-Means *Clustering* dan Naïve Bayes. Dalam hal ini, K-Means *Clustering* tidak berperan sebagai algoritma utama klasifikasi, melainkan digunakan pada tahap awal untuk mengoptimasi atribut melalui proses pengelompokan data berdasarkan kemiripan karakteristik. Hasil pengelompokan tersebut kemudian dimanfaatkan oleh algoritma Naïve Bayes sebagai dasar untuk melakukan klasifikasi akhir terhadap kelayakan penerima beasiswa. Dengan demikian, penelitian ini berfokus pada penerapan kombinasi algoritma K-Means dan Naïve Bayes untuk mengklasifikasikan siswa yang layak atau tidak layak menerima beasiswa, dengan harapan dapat menghasilkan model prediksi yang lebih akurat dan efisien.

Penelitian terdahulu menunjukkan bahwa kombinasi beberapa algoritma *data mining* dapat meningkatkan efektivitas dalam proses analisis dan pengambilan keputusan. Usman dan Paraga [11] mengimplementasikan algoritma K-Means, Naïve Bayes, dan Decision Tree untuk memprediksi penjualan bahan bakar minyak (BBM), di mana K-Means digunakan untuk *clustering* data penjualan, sedangkan Naïve

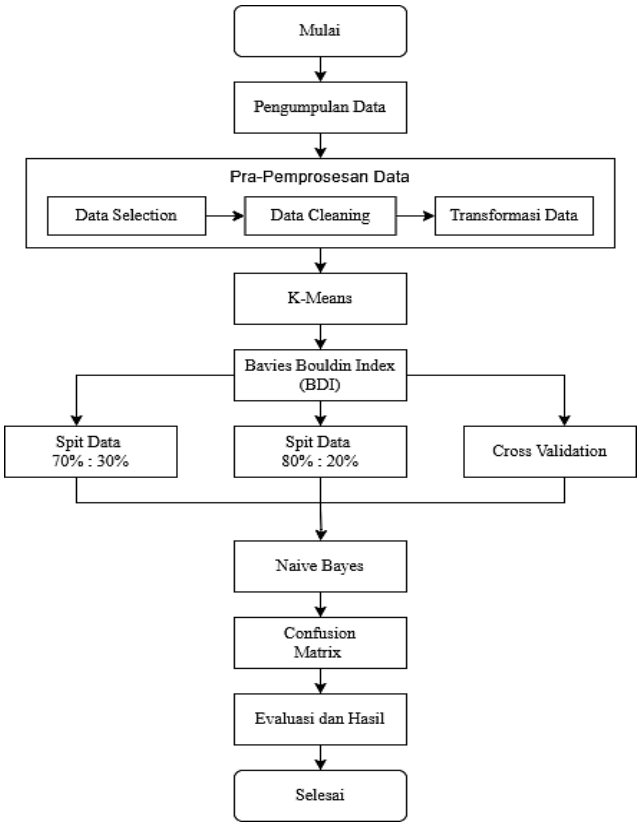
Bayes dan Decision Tree digunakan untuk prediksi. Hasilnya, kombinasi tersebut mampu mengoptimalkan distribusi BBM dan meningkatkan kepuasan pelanggan. Penelitian lain oleh Best, Foo, dan Tian [12] juga menerapkan pendekatan *hybrid* dengan menggabungkan K-Means dan Naïve Bayes untuk deteksi anomali pada sistem *Internet of Things* (IoT), dan berhasil meningkatkan akurasi deteksi hingga mencapai 90%–100%. Sementara itu, Kurniawan dan Susanto [13], menggunakan kombinasi kedua algoritma tersebut dalam analisis sentimen Pemilihan Presiden (Pilpres) 2019, di mana K-Means berfungsi untuk *clustering* data dan Naïve Bayes digunakan untuk *classification*. Hasil pengujian terhadap 100 dan 150 data uji menunjukkan akurasi rata-rata sebesar 93,35% dengan *error rate* sebesar 6,66%. Penelitian lain oleh Alviyah dkk. [14] mengimplementasikan algoritma Naïve Bayes untuk menentukan calon penerima beasiswa di SMK YPM 14 Sumobito Jombang, dengan hasil akurasi 90,48%, *precision* 96,88%, dan *recall* 83,33%. Selain itu, Rahayu dkk. [15] menerapkan algoritma K-Means untuk mengelompokkan calon penerima beasiswa Bidikmisi ke dalam empat klaster, yaitu sangat layak, kurang layak, dipertimbangkan, dan tidak layak menerima beasiswa.

Berdasarkan hal tersebut, penelitian ini dilakukan untuk mengklasifikasikan kelayakan penerima beasiswa menggunakan naïve bayes dengan optimasi atribut berbasis K-Means *clustering* dengan evaluasi menggunakan *confusion matrix* melalui parameter akurasi, presisi, *recall*, dan *F1-score*.

Tujuan penelitian ini adalah: (1) menerapkan algoritma K-Means untuk pengelompokan kelayakan siswa, (2) memanfaatkan hasil *clustering* sebagai fitur tambahan dalam model Naïve Bayes, serta (3) mengevaluasi tingkat akurasi model gabungan tersebut. Manfaat yang diharapkan yaitu memberikan solusi seleksi beasiswa yang lebih objektif bagi sekolah, memperkaya kajian ilmiah dalam penerapan *data mining*, serta meningkatkan transparansi dan keadilan bagi siswa penerima bantuan pendidikan.

2. METODOLOGI PENELITIAN

Untuk menghasilkan penelitian yang berkualitas dengan tujuan yang terarah, diperlukan langkah-langkah penelitian yang tersusun secara sistematis. Tahapan-tahapan yang perlu dilakukan dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Metode Penelitian

2.1. Pengumpulan Data

Data penelitian diperoleh dari Dapodik (Data Pokok Pendidikan) yang dikelola oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Kemendikbudristek). Data yang digunakan dalam penelitian ini Adalah data tahun ajaran 2024/2025 yang berjumlah 4.155 siswa dari 23 sekolah dasar di Kecamatan Tabir, Kabupaten Merangin. Atribut data meliputi no, sekolah, nama siswa, alat transportasi, status kepemilikan KPS, pekerjaan ayah, penghasilan ayah, pekerjaan ibu, penghasilan ibu, rombel saat ini, status

penerimaan PIP, kebutuhan khusus, anak keberapa, jumlah saudara kandung, serta jarak rumah ke sekolah. Data diperoleh melalui kerja sama dengan koordinator wilayah pendidikan dan operator sekolah.

2.2. Pra-Pemrosesan Data

Pada tahap Pra-Pemrosesan Data, , data yang telah dikumpulkan dari aplikasi Dapodik akan diproses untuk memastikan kualitas dan kesiapan data sebelum dianalisis lebih lanjut. Tahapan ini mencakup beberapa langkah penting [16], yaitu:

1. *Data Selection* - salah satu tahap awal dalam proses *data preprocessing* yang bertujuan untuk memilih subset data yang relevan dari *dataset* yang tersedia, sesuai dengan tujuan analisis atau pemodelan.
2. *Data Cleaning* – Mengidentifikasi dan mengatasi data yang tidak lengkap, duplikat, atau inkonsisten, seperti data siswa yang kosong atau terdapat kesalahan dalam pengisian atribut
3. Transformasi Data – Mengonversi format data agar sesuai dengan kebutuhan analisis, seperti standarisasi format pekerjaan orang tua, pengkodean kategori untuk atribut tertentu

Tahap *pre-processing* ini bertujuan untuk meningkatkan akurasi dan efektivitas analisis data yang akan dilakukan pada tahap berikutnya

2.3. Algoritma K-Means

Algoritma K-Means merupakan salah satu metode klasterisasi yang populer dalam *unsupervised learning* dengan tujuan mengelompokkan data ke dalam sejumlah klaster berdasarkan tingkat kemiripan. Setiap klaster diwakili oleh sebuah pusat (*centroid*), dan data ditempatkan pada klaster dengan jarak terdekat, umumnya menggunakan *Euclidean Distance* [17].

Kelebihan algoritma ini adalah sederhana, cepat, efisien pada *dataset* berukuran besar, serta mudah diimplementasikan. Namun, kelemahannya antara lain harus menentukan jumlah klaster (K) di awal, hasil sangat bergantung pada pemilihan *centroid* awal, serta kurang efektif untuk data dengan bentuk klaster tidak bulat, ukuran bervariasi, atau mengandung *outlier*.

Langkah-langkah K-Means meliputi:

1. Menentukan jumlah klaster (K).
2. Menentukan *centroid* awal secara acak.
3. Menghitung jarak tiap data ke *centroid* dengan rumus *Euclidean Distance*:

$$d(x,c) = \sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \quad (1)$$

Keterangan:
$x = Data,\ c = Centroid,\ n = Jumlah\ atribut$

4. Menempatkan data ke klaster dengan jarak terdekat.
5. Menghitung ulang *centroid* menggunakan rata-rata data dalam klister
6. Mengulangi proses hingga *centroid* stabil atau mencapai iterasi maksimum.

2.4. *Davies-Boulding Index* (DBI)

DBI adalah salah satu indeks validitas *cluster* yang digunakan untuk mengukur kualitas hasil pengelompokan dengan memperhatikan rasio antara jarak intra-*cluster* dan inter-*cluster*. Semakin kecil nilai DBI maka semakin baik kualitas *cluster* yang terbentuk, karena data dalam satu *cluster* lebih homogen dan antar *cluster* semakin terpisah [17]. DBI termasuk dalam kategori evaluasi internal yang tidak memerlukan label kelas sehingga cocok digunakan dalam unsupervised learning seperti K-Means [18]. Selain itu, penelitian terbaru juga menegaskan bahwa DBI masih menjadi salah satu metode evaluasi *clustering* yang paling banyak digunakan karena sifatnya yang sederhana dan konsisten dalam membandingkan hasil *clustering* [19]. Secara matematis, DBI dirumuskan sebagai:

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} \underset{j \neq i}{max}\left(\frac{S_i + S_j}{M_{1j}}\right) \quad (2)$$

Keterangan:
$k = Jumlah\ cluster,\quad M_{ij} = Jarak\ antara\ centroid\ cluster\ ke\ i\ dan\ ke\ j$
$S_i = Rata - rata\ jarak\ antar\ titik\ data\ dengan\ centroid\ cluster\ ke - 1$

2.5 *Split Data* dan *Cross Validation*

Split data atau *hold-out validation* merupakan teknik evaluasi sederhana yang digunakan dalam proses pembelajaran mesin dengan cara membagi *dataset* menjadi dua bagian, yaitu data latih (*training set*) dan data uji (*testing set*). Data latih digunakan untuk membangun dan melatih model, sedangkan data uji berfungsi

untuk menilai kinerja model terhadap data yang belum pernah dilihat sebelumnya. Pembagian data ini bertujuan untuk mengukur kemampuan generalisasi model dalam menghadapi data baru. Rasio pembagian yang umum digunakan adalah 70:30 atau 80:20, di mana sebagian besar data digunakan untuk pelatihan dan sisanya untuk pengujian. Menurut Widodo dkk. [20], metode pembagian data seperti ini efektif untuk mengevaluasi performa model pembelajaran mesin pada *dataset* yang memiliki ukuran besar.

Cross Validation merupakan metode evaluasi yang dianggap lebih andal dibandingkan teknik *split data* karena dapat meminimalkan bias yang timbul akibat pembagian *dataset* yang tidak seimbang. Prinsip dasar dari metode ini adalah membagi *dataset* menjadi beberapa bagian yang disebut *folds*. Salah satu metode yang paling umum digunakan adalah *k-Fold Cross Validation*, di mana *dataset* dibagi menjadi k lipatan (*fold*) dengan ukuran yang relatif sama. Proses pelatihan dan pengujian dilakukan sebanyak k kali secara bergantian, di mana pada setiap iterasi satu *fold* digunakan sebagai data uji, sedangkan k–1 *fold* lainnya digunakan sebagai data latih. Nilai performa akhir model diperoleh dari rata-rata hasil evaluasi seluruh iterasi, sehingga memberikan estimasi kinerja model yang lebih stabil dan representatif. Menurut Tembusai dkk. [21], penggunaan *k-Fold Cross Validation* terbukti mampu meningkatkan akurasi model klasifikasi karena setiap data memiliki kesempatan untuk menjadi data uji dan data latih secara bergantian.

2.6 Algoritma Naive Bayes

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan "naive" dimana diasumsikan kondisi antar atribut saling bebas. Algoritma Naïve Bayes memiliki keunggulan. dimana keunggulannya antara lain cepat dan efisien, mudah diimplementasikan, dan dapat digunakan untuk data diskrit dan kontinu [22], probabilitas posterior dihitung dengan:

$$P(C|X) = \frac{P(X|C)\ .\ P(C)}{P(X)}. \tag{3}$$

2.7 Evaluasi Dengan *Confusion Matrix*

Confusion Matrix merupakan salah satu metode evaluasi yang digunakan untuk mengukur performa algoritma klasifikasi. Matriks ini membandingkan hasil prediksi model dengan data aktual untuk memberikan informasi detail mengenai tingkat kesalahan dan akurasi prediksi. Menurut Helmud dkk. [23]. *Confusion Matrix* berbentuk tabel (klasifikasi biner) dengan elemen pada Tabel 1 [24].

Tabel 1. Elemen *Confusion Matrix*

Prediksi/*Actual*	Positif (Aktual)	Negatif (Aktual)
Positif (Prediksi)	*True Positive* (TP)	*False Positive* (FP)
Negatif (Prediksi)	*False Negative* (FN)	*True Negative* (TN)

Dari *Confusion Matrix* ini, beberapa metrik evaluasi dapat dihitung dengan persamaan 4 –7 [25] :

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1-Score = 2\ .\ \frac{Precision\ .\ Recall}{Precision+Recall} \tag{7}$$

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data tahun ajaran 2024/2025 yang berjumlah 4.155 siswa dari 23 dengan atribut data meliputi no, sekolah, nama siswa, alat transportasi, status kepemilikan KPS, pekerjaan ayah, penghasilan ayah, pekerjaan ibu, penghasilan ibu, rombel saat ini, status penerimaan PIP, kebutuhan khusus, anak keberapa, jumlah saudara kandung, serta jarak rumah ke sekolah.

3.2. Pra-Pemrosesan Data

3.2.1 *Data Selection*

Pada tahap ini dilakukan proses pemilihan data dengan cara menghapus atribut yang tidak relevan terhadap penelitian. Atribut yang dihapus antara lain sekolah, nama, rombel, dan anak ke berapa, karena

hanya berfungsi sebagai identitas dan tidak berpengaruh dalam proses penelitian kelayakan siswa penerima beasiswa. Dengan demikian, data yang tersisa adalah atribut-atribut yang benar-benar berhubungan dengan penelitian. Atribut-atribut yang digunakan dalam penelitian dapat dilihat pada Tabel 2.

Tabel 2. Atribut-atribut yang digunakan

Kode	Atribut	Keterangan	Nilai
A1	Alat Transportasi	Alat yang digunakan siswa untuk pergi ke sekolah	1. Sepeda motor, 2. Jalan kaki, 3. Sepeda, 4. Mobil, 5. Ojek, 6. Lainnya
A2	Penerima KPS	Status kepemilikan KPS (Kartu Perlindungan Sosial)	1. Tidak, 2. Iya
A3	Pekerjaan Ayah	Jenis pekerjaan ayah kandung	1. Buruh, 2. Karyawan BUMN, 3. Wiraswasta, 4. Wirausaha, 5. Peternakan, 6. Petani, 7. Pedagang, 8. Pensiunan, 9. PNS/TNI/Polri, 10. Sudah Meninggal, 11. TKI, 12. Tidak bekerja, 13. Lainnya
A4	Penghasilan Ayah	Jumlah penghasilan bulanan ayah	1. Tidak Berpenghasilan, 2. < Rp500.000, 3. Rp500.000–Rp999.999, 4. Rp1.000.000–Rp1.999.999, 5. Rp2.000.000–Rp4.999.999, 6. Rp5.000.000–Rp20.000.000
A5	Pekerjaan Ibu	Jenis pekerjaan ibu kandung	1. Buruh, 2. Karyawan BUMN, 3. Wiraswasta, 4. Wirausaha, 5. Peternakan, 6. Petani, 7. Pedagang, 8. Pensiunan, 9. PNS/TNI/Polri, 10. Sudah Meninggal, 11. TKI, 12. Tidak bekerja, 13. Lainnya
A6	Penghasilan Ibu	Jumlah penghasilan bulanan ibu	1. Tidak Berpenghasilan, 2. < Rp500.000, 3. Rp500.000–Rp999.999, 4. Rp1.000.000–Rp1.999.999, 5. Rp2.000.000–Rp4.999.999, 6. Rp5.000.000–Rp20.000.000
A7	Layak PIP	Status usulan siswa sebagai penerima PIP dari sekolah	1. Tidak, 2. Iya
A8	Kebutuhan Khusus	Status kebutuhan khusus	1. Tidak, 2. Iya
A9	Jumlah saudara	Jumlah Saudara Kandung	Angka diskrit (misal: 0, 1, 2, 3, 4, dst.)
A10	Jarak Rumah ke Sekolah	Estimasi jarak dari rumah ke sekolah	1,2,3,4,5,6

3.2.2 *Data Cleaning*

Hasil deteksi *missing value* yang diperoleh melalui proses analisis menggunakan aplikasi *Altair RapidMiner* 2024 1.0 tidak ditemukan adanya data yang hilang (*missing value*) pada seluruh atribut yang digunakan. Dengan demikian, *dataset* yang digunakan dalam penelitian ini dinyatakan lengkap dan layak untuk digunakan pada tahap *data preprocessing* lanjutan tanpa memerlukan proses imputasi atau penghapusan data.

3.2.3 Transformasi Data

Beberapa atribut bersifat kategorikal diubah menjadi bentuk numerik menggunakan teknik *label encoding* agar dapat diolah oleh algoritma *machine learning*. Maka dari 10 atribut yang digunakan 8 atribut yang dilakukan transformasi data (alat transportasi, penerima KPS, pekerjaan ayah,penghasilan ayah, pekerjaan ibu, penghasilan ibu,layak PIP, dan kebutuhan khusus). Hasil transformasi data dapat dilihat pada Tabel 3.

Tabel 3. Hasil transformasi data

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1	1	6	4	12	1	1	1	2	1
1	1	6	4	12	1	1	1	2	1
1	1	6	3	12	1	1	1	1	4
1	1	3	4	12	1	1	1	2	3
2	1	6	4	12	1	1	1	1	1
1	1	6	3	12	1	2	1	0	1
1	1	6	3	12	1	2	1	0	1
2	1	6	3	12	1	2	1	2	1
2	1	6	4	12	1	2	1	0	1
2	1	6	4	12	1	2	1	0	1

3.3. Algoritma K-Means

Setelah dilakukan proses perhitungan *clustering* baik secara manual menggunakan Microsoft Excel maupun dengan bantuan perangkat lunak Altair RapidMiner 2024 1.0, diperoleh hasil yang konsisten, yaitu menghasilkan komposisi anggota *cluster* yang sama. Hal ini menunjukkan bahwa implementasi algoritma K-Means pada kedua pendekatan memberikan keluaran yang identik, sehingga dapat dipastikan bahwa proses perhitungan yang dilakukan telah sesuai. Jumlah anggota pada setiap *cluster* dapat dilihat bahwa hasil pengelompokan data dengan algoritma K-Means dengan K=2 menghasilkan jumlah anggota *cluster* sebanyak 3399 data pada *Cluster* 0 dan 756 data pada *Cluster* 1, dapat dilihat pada Tabel 4.

Tabel 4. Hasil *Clustering*

Cluster	Jumlah Data	Persentase
C0	3399	81,8 %
C1	756	18,2 %
Total	4155	100 %

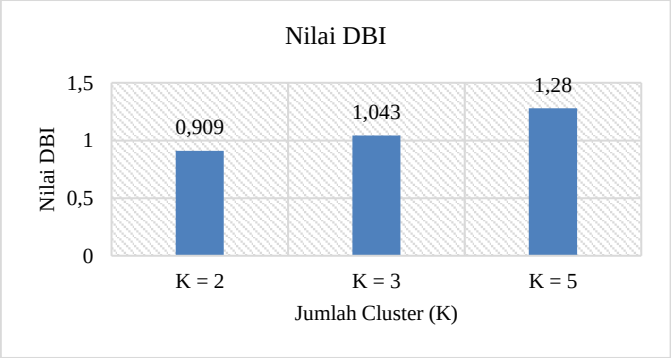
3.4. *Davies-Bouldin Index* (DBI)

Setelah diperoleh hasil *clustering*, langkah selanjutnya adalah melakukan evaluasi performa untuk mengetahui kualitas pengelompokan yang dihasilkan, proses evaluasi DBI dilakukan dengan bantuan aplikasi *RapidMiner* menggunakan parameter jumlah *cluster* K = 2, K = 3, dan K = 5. Setelah rancangan dijalankan lalu menghasilkan nilai DBI untuk setiap jumlah *cluster* yang diuji. Hasil perhitungan DBI dapat dilihat pada Tabel 5.

Tabel 5. Hasil perhitungan DBI

Jumlah *Cluster* (K)	Nilai DBI
K = 2	0.909
K = 3	1.043
K = 5	1.280

Tabel 5 menunjukkan hasil evaluasi kualitas *cluster* menggunakan DBI. Nilai DBI digunakan untuk menilai seberapa baik *cluster* yang terbentuk, dengan prinsip bahwa semakin kecil nilai DBI, maka kualitas *cluster* semakin baik karena data dalam *cluster* menjadi lebih kompak dan antar *cluster* semakin terpisah. Untuk memperjelas perbandingan, hasil DBI divisualisasikan dalam bentuk grafik pada Gambar 2.



Gambar 2. Grafik nilai DBI berdasarkan jumlah *cluster*

Dari Gambar 2 terlihat bahwa nilai DBI terkecil diperoleh pada K = 2, sehingga jumlah *cluster* terbaik dalam penelitian ini adalah 2 *cluster*. Secara praktis, pembentukan dua *cluster* ini dapat diinterpretasikan sebagai kelompok siswa yang layak menerima beasiswa dan kelompok siswa yang tidak layak menerima beasiswa. Hasil pengelompokan ini selanjutnya menjadi dasar dalam analisis klasifikasi menggunakan algoritma Naïve Bayes,

3.5. Naïve Bayes

Tahap berikutnya adalah melakukan transformasi label agar lebih bermakna sesuai dengan tujuan penelitian. Setelah proses pengelompokan data menggunakan algoritma K-Means menghasilkan dua *cluster*, yaitu *cluster*_0 dan *cluster*_1, maka nilai *cluster*_0 dipetakan menjadi kategori “Layak”, sedangkan nilai *cluster*_1 dipetakan menjadi kategori “Tidak Layak”. Dengan demikian, atribut hasil *clustering* yang semula berbentuk kode teknis dapat ditransformasi menjadi label kategorikal yang merepresentasikan kondisi sebenarnya terkait kelayakan siswa penerima beasiswa. Transformasi *cluster* dapat dilihat pada Tabel 6.

Tabel 6. Transformasi *Cluster* ke Klasifikasi

Jumlah Data	Hasil *Cluster* K-Means	Label Baru
3399	C0	Layak
756	C1	Tidak Layak

3.6. Evaluasi Model Dengan *Split Data* dan *Cross Validation*

3.6.1 *Split Data* 70% : 30%

Tahap pertama dimulai dengan *operator read excel*, selanjutnya, data yang sudah dibaca dibagi menjadi dua bagian menggunakan operator *split data* perbandingan data yang digunakan dalam penelitian ini adalah 70% *training* dan 30% *testing*, berikutnya adalah naïve bayes yang digunakan untuk membangun model klasifikasi berbasis probabilitas dari data *training*, setelah model terbentuk, *operator apply model* digunakan untuk mengaplikasikan model naïve bayes pada data *testing*, tahap terakhir adalah evaluasi model dengan *operator performance*. Hasil dari proses dapat dilihat pada Tabel 7.

Tabel 7. Hasil Spit Data 70% : 30%

Accuracy = 96.15%

Prediksi / Aktual	*True* Layak	*True* Tidak Layak	*Class Precision*
Pred. Layak	973	1	99.90%
Pred. Tidak Layak	47	226	82.78%
Class recall	95.39%	99.56%	

Hasil evaluasi model klasifikasi menggunakan algoritma Naïve Bayes dengan pembagian data 70% sebagai data *training* dan 30% sebagai data *testing* ditunjukkan pada Tabel 7. Dari hasil pengujian diperoleh tingkat akurasi sebesar 96,15%.

3.6.2 *Split Data* 80% : 20%

Proses yang dilakukan sama seperti diatas hanya menggantikan *split data* perbandingan data yang digunakan dalam penelitian ini adalah 70% *training* dan 30% *testing*. Menjadi *split data* perbandingan data yang digunakan dalam penelitian ini adalah 80% *training* dan 20% *testing*. Hasil dari proses dapat dilihat pada Tabel 8.

Tabel 8. Hasil Spit Data 80% : 20%

Accuracy = 96.15%

Prediksi / Aktual	*True* Layak	*True* Tidak Layak	*Class Precision*
Pred. Layak	654	1	99.85%
Pred. Tidak Layak	26	150	85.23%
Class recall	96.18%	99.34%	

Hasil evaluasi model klasifikasi menggunakan algoritma Naïve Bayes dengan pembagian data 80% sebagai data *training* dan 20% sebagai data *testing* ditunjukkan pada Tabel 8. Dari hasil pengujian diperoleh tingkat akurasi sebesar 96,75%.

3.6.3 *Cross Validation*

Pada penelitian ini digunakan teknik *k-fold cross validation* dengan nilai k = 10, yang artinya *dataset* dibagi menjadi sepuluh lipatan (*fold*) dengan ukuran yang relatif sama. Setiap lipatan secara bergantian digunakan sebagai data uji (*testing*), sementara sembilan lipatan lainnya digunakan sebagai data latih (*training*). Proses ini diulang sebanyak sepuluh kali sehingga setiap lipatan berperan sebagai data uji tepat satu kali. Setelah dijalankan maka setiap *fold* menghasilkan nilai akurasi yang berbeda sesuai dengan variasi data latih dan data uji yang digunakan. Hasil akurasi pada setiap *fold* dapat dilihat pada Tabel 9.

Tabel 9. Hasil Akurasi pada Setiap *Fold* (Naïve Bayes, k=10)

Fold ke-	Akurasi (%)
1	99.04
2	97.35
3	98.56
4	99.52
5	97.11
6	98.08
7	98.07
8	97.59
9	95.91
10	97.83

Fold ke-	Akurasi (%)
Rata-rata	97.91

Berdasarkan Tabel 9 terlihat bahwa nilai akurasi pada setiap *fold* berada pada kisaran 95,91% hingga 99,52% dengan rata-rata akurasi sebesar 97,91%. Hal ini menunjukkan bahwa algoritma Naïve Bayes memiliki performa yang stabil pada setiap proses validasi, serta mampu menghasilkan tingkat akurasi yang tinggi dalam mengklasifikasikan data kelayakan siswa penerima beasiswa. Hasil dari proses dapat dilihat pada Tabel 10.

Tabel 10. Hasil *Cross Validation*

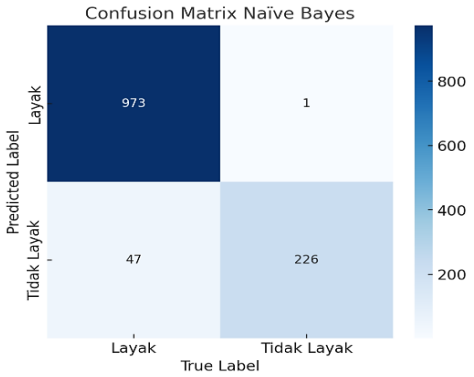
Accuracy = 96.15%

Prediksi / Aktual	*True* Layak	*True* Tidak Layak	*Class Precision*
Pred. Layak	3313	1	99.97%
Pred. Tidak Layak	86	755	89.77%
Class recall	97.47%	99.87%	

3.7 *Confusion Matrix*

3.7.1 *Confusion Matrix Split Data* 70% : 30%

Setelah dilakukan proses *split data* dengan perbandingan 70% :30% untuk data uji, langkah berikutnya adalah menghitung nilai *confusion matrix* untuk mengevaluasi kinerja algoritma Naïve Bayes dalam proses klasifikasi. *Confusion matrix* digunakan untuk menggambarkan perbandingan antara hasil prediksi sistem dengan data aktual, sehingga dapat diketahui tingkat ketepatan model dalam mengklasifikasikan data.



Gambar 3. *Confusion Matrix* Naïve Bayes (*Split Data* 70% : 30%)

Dari Gambar 3 dapat dijelaskan bahwa model berhasil mengklasifikasikan 973 data Layak secara benar (*True Positive*) dan 226 data Tidak Layak secara benar (*True Negative*). Namun masih terdapat 1 data Tidak Layak yang salah diklasifikasikan sebagai Layak (*False Positive*) serta 47 data Layak yang salah diklasifikasikan sebagai Tidak Layak (*False Negative*). Selanjutnya, hasil perhitungan metrik evaluasi berupa *precision*, *recall*, dan *F1-score* dapat dilihat pada Tabel 11 berikut:

Tabel 11. Hasil Evaluasi Kinerja Naïve Bayes (*Split Data* 70% : 30%)

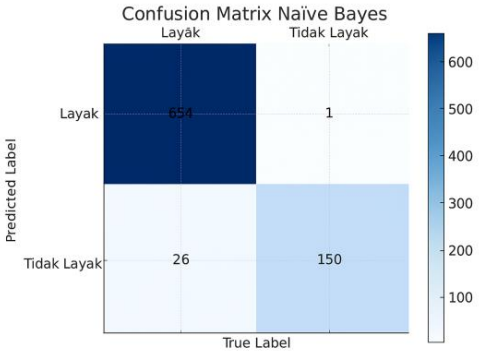
Kelas	*Precision*	*Recall*	*F1-Score*
Layak	99.90 %	95.39 %	97.59 %
Tidak Layak	82.78 %	99.56 %	90.40 %

3.7.2 *Confusion Matrix Split Data* 80% : 20%

Setelah dilakukan proses *split data* dengan perbandingan 80% :20% untuk data uji, langkah berikutnya adalah menghitung nilai *confusion matrix* untuk mengevaluasi kinerja algoritma Naïve Bayes dalam proses klasifikasi.

Dari Gambar 4 dapat dijelaskan bahwa model berhasil mengklasifikasikan 654 data Layak secara benar (*True Positive*) dan 150 data Tidak Layak secara benar (*True Negative*). Namun masih terdapat 1 data Tidak Layak yang salah diklasifikasikan sebagai Layak (*False Positive*) serta 26 data Layak yang salah diklasifikasikan sebagai Tidak Layak (*False Negative*).

Selanjutnya, hasil perhitungan metrik evaluasi berupa *precision*, *recall*, dan *F1-score* dapat dilihat pada Tabel 12.



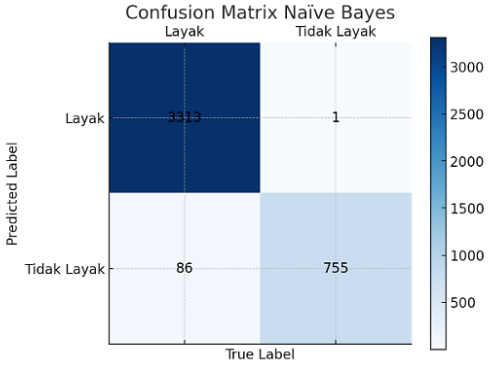
Gambar 4. *Confusion Matrix* Naïve Bayes (*Split Data* 80% : 20%)

Tabel 12. Hasil Evaluasi Kinerja Naïve Bayes (*Split Data* 80% : 20%)

Kelas	*Precision*	*Recall*	*F1-Score*
Layak	99.85 %	96.18 %	97.98 %
Tidak Layak	85.23 %	99.34 %	91.74 %

3.7.3 *Confusion Matrix Cross Validation*

Setelah dilakukan proses *Cross validation*, langkah berikutnya adalah menghitung nilai *confusion matrix* untuk mengevaluasi kinerja algoritma Naïve Bayes dalam proses klasifikasi.



Gambar 5. *Confusion Matrix* Naïve Bayes (*Cross Validation*)

Dari Gambar 5 dapat dijelaskan bahwa model berhasil mengklasifikasikan 3313 data Layak secara benar (*True Positive*) dan 755 data Tidak Layak secara benar (*True Negative*). Namun masih terdapat 1 data Tidak Layak yang salah diklasifikasikan sebagai Layak (*False Positive*) serta 86 data Layak yang salah diklasifikasikan sebagai Tidak Layak (*False Negative*). Selanjutnya, hasil perhitungan metrik evaluasi berupa *precision*, *recall*, dan *F1-score* dapat dilihat pada Tabel 10.

Tabel 10. Hasil Evaluasi Kinerja Naïve Bayes (*Cross Validation*)

Kelas	*Precision*	*Recall*	*F1-Score*
Layak	99.85 %	96.18 %	97.98 %
Tidak Layak	85.23 %	99.34 %	91.74 %

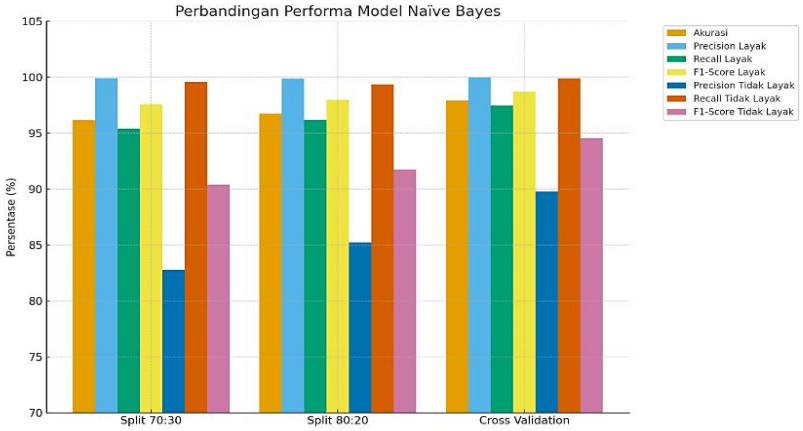
3.8 Evaluasi Dan Hasil

Tabel 10. Ringkasan Hasil Evaluasi Naïve Bayes

Metrik Evaluasi	*Split Data* 70:30	*Split Data* 80:20	*Cross Validation*
Akurasi	96,15 %	96,75 %	97,91 %
Precision (Layak)	99,90 %	99,85 %	99,97 %
Recall (Layak)	95,39 %	96,18 %	97,47 %
F1-Score (Layak)	97,59 %	97,98 %	98,70 %
Precision (Tidak Layak)	82,78 %	85,23 %	89,77 %
Recall (Tidak Layak)	99,56 %	99,34 %	99,87 %
F1-Score (Tidak Layak)	90,40 %	91,74 %	94,55 %

Dari Tabel 10 dan Gambar 6 menunjukkan bahwa penggunaan *cross validation* terbukti menghasilkan performa paling optimal dan stabil. Akurasi meningkat hingga 97,91%, dan metrik evaluasi pada kedua kelas menjadi lebih seimbang. Peningkatan *precision* pada kelas Tidak Layak hingga 89,77% menunjukkan bahwa kesalahan klasifikasi semakin berkurang, sementara *recall* yang tetap sangat tinggi menegaskan bahwa kemampuan model dalam mendeteksi data Tidak Layak hampir sempurna. Selain itu, *F1-Score* pada kedua kelas juga lebih tinggi dibandingkan metode *split data*, yang menandakan adanya keseimbangan antara *precision* dan *recall*.

Hasil evaluasi menunjukkan bahwa akurasi, *precision*, *recall*, dan *F1-Score* konsisten tinggi pada semua skema pengujian, dengan performa terbaik pada metode *cross validation* yang mencapai akurasi 97,91%. *Precision* pada kelas Layak mendekati sempurna, *recall* kelas Tidak Layak sangat tinggi, serta *F1-Score* meningkat konsisten, sehingga model terbukti andal untuk klasifikasi kelayakan beasiswa.



Gambar 6. Grafik Perbandingan Performa Model Naïve Bayes

3.9. Diskusi

Penelitian ini bertujuan untuk mengklasifikasikan kelayakan penerima beasiswa SD di Kecamatan Tabir dengan menggunakan kombinasi algoritma K-Means dan Naïve Bayes. *Dataset* yang digunakan berjumlah 4.155 siswa yang telah melalui proses pra-pemrosesan, meliputi pembersihan data, transformasi kategori ke bentuk numerik, serta seleksi atribut yang relevan. Setelah data dinyatakan bersih dan siap digunakan, dilakukan proses *clustering* menggunakan K-Means untuk mengelompokkan data berdasarkan kesamaan karakteristik, kemudian hasil klaster digunakan sebagai masukan dalam proses klasifikasi menggunakan algoritma Naïve Bayes.

Proses pengujian dilakukan menggunakan tiga skenario pembagian data, yaitu *split data* 70:30, *split data* 80:20, dan 10-*fold cross validation*. Setiap skenario menghasilkan *confusion matrix* yang digunakan untuk menghitung nilai akurasi, presisi, *recall*, dan *F1-score* sebagai indikator kinerja model. Hasil pengujian di atas menunjukkan bahwa kombinasi algoritma K-Means dan Naïve Bayes memberikan performa klasifikasi yang cukup tinggi dalam menentukan kelayakan penerima beasiswa. Penggunaan K-Means pada tahap awal berperan penting dalam mengelompokkan data ke dalam klaster homogen, sehingga ketika dilakukan klasifikasi dengan Naïve Bayes.

Jika dibandingkan dengan penelitian sebelumnya, hasil ini sejalan dengan studi yang dilakukan oleh Sanjaya (2024) dan Zafitri (2023) yang menyatakan bahwa kombinasi metode *unsupervised learning* dan *supervised learning* mampu meningkatkan kinerja model prediksi karena data yang dikelompokkan terlebih dahulu menjadi lebih teratur. Penelitian oleh D. B. Sanjaya dan I. N. Suastika (2025) juga menegaskan pentingnya integrasi metode analitik dalam mendukung pengambilan keputusan di bidang pendidikan. Temuan dalam penelitian ini menunjukkan kesesuaian dengan teori tersebut, di mana proses *clustering* dengan K-Means dapat memperkuat kemampuan Naïve Bayes dalam melakukan klasifikasi probabilistik yang lebih akurat.

Pengembangan algoritma, penelitian ini hanya menggunakan kombinasi K-Means dan Naïve Bayes. Untuk penelitian selanjutnya, dapat dilakukan perbandingan dengan algoritma lain seperti K-Medoids, DBSCAN, Decision Tree, Random Forest, Support Vector Machine (SVM), atau Neural Network, sehingga dapat diketahui metode terbaik untuk kasus serupa.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma K-Means berhasil mengelompokkan data siswa menjadi dua *cluster* (Layak dan Tidak Layak) dengan kualitas pengelompokan cukup baik (DBI = 0,909). Label hasil *clustering* yang digunakan sebagai fitur tambahan pada Naïve Bayes meningkatkan representasi data

sehingga mendukung proses klasifikasi secara lebih akurat. Hasil evaluasi model menunjukkan performa sangat baik, dengan akurasi 96,15% pada *split data* 70:30, 96,75% pada *split data* 80:20, dan mencapai nilai tertinggi 97,91% pada 10-*fold cross validation*. *Precision* kelas Layak berkisar antara 99,85–99,97%, *recall* 95,39–97,47%, dan *F1-Score* 97,59–98,70%, sedangkan untuk kelas Tidak Layak, *precision* 82,78–89,77%, *recall* 99,34–99,87%, dan *F1-Score* 90,40–94,55%. Hasil ini membuktikan bahwa kombinasi K-Means dan Naïve Bayes efektif dan andal untuk mendukung sistem pendukung keputusan (SPK) dalam penentuan kelayakan penerima beasiswa, mampu mengurangi bias, serta dapat diterapkan pada skala data yang lebih besar.

REFERENSI

[1] D. B. Sanjaya and I. N. Suastika, "Memahami Hakikat Pembelajaran Pkn Sebagai Pendidikan Karakter Di Sekolah Dasar," *Pendas: Jurnal Ilmiah Pendidikan Dasar*, vol. 10, no. 03, pp. 494–504, 2025, doi: https://doi.org/10.23969/jp.v10i03.28868.

[2] E. Zafitri, M. Mutiara, W. Asni, and R. Ananda, "PENINGKATAN AKSES MUTU DAN PEMERATAAN PENDIDIKAN," *Pendas: Jurnal Ilmiah Pendidikan Dasar*, vol. 9, no. 2, pp. 4336–4346, 2024.

[3] J. Jumanah and H. Rosita, "Evaluasi program indonesia pintar dalam upaya pemerataan pendidikan," *The Indonesian Journal of Public Administration (IJPA)*, vol. 8, no. 2, pp. 72–84, 2022, doi: https://doi.org/10.52447/ijpa.v8i2.6042.

[4] I. Zamjani, "Pelaksanaan Program Indonesia Pintar bagi penerima Kartu Indonesia Pintar reguler: Studi di empat daerah kunjungan kerja presiden tahun 2017," *Jurnal Penelitian Kebijakan Pendidikan*, vol. 11, no. 2, pp. 64–82, 2018, doi: https://doi.org/10.24832/jpkp.v11i2.225.

[5] S. Sulistiyanto, E. Nadeak, N. Rahmi, and M. Malahayati, "Metode Data Mining dalam Kasus Seleksi Beasiswa: Literature Review," *Jurnal Penelitian Inovatif*, vol. 4, no. 3, pp. 1091–1100, 2024, doi: https://doi.org/10.54082/jupin.468.

[6] J. J. Xu, N. Yuruk, Z. Feng, and T. Schweiger, "Knowledge Discovery and Data Mining.," 2014.

[7] M. Suyal and S. Sharma, "A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning," *Journal of Artificial Intelligence and Systems*, vol. 6, no. 1, pp. 85–95, 2024, doi: https://doi.org/10.33969/AIS.2024060106.

[8] F. N. dan A. K. D. Wulandari, "Penerapan algoritma Naïve Bayes untuk klasifikasi kelayakan penerimaan beasiswa mahasiswa ," *Jurnal Teknologi dan Sistem Komputer (JUSTIN)*, vol. 8, no. 3, 2020.

[9] B. Bustami, "Penerapan algoritma Naive Bayes untuk mengklasifikasi data nasabah asuransi," *TECHSI-Jurnal Teknik Informatika*, vol. 5, no. 2, 2013, doi: https://doi.org/10.29103/techsi.v5i2.154.

[10] A. Saleh, "Implementasi metode klasifikasi naive bayes dalam memprediksi besarnya penggunaan listrik rumah tangga," *Creative Information Technology Journal*, vol. 2, no. 3, pp. 207–217, 2015, doi: https://doi.org/10.24076/citec.2015v2i3.49.

[11] U. Arfan and N. Paraga, "Perbandingan Algoritma K-Means, Naïve Bayes dan Decision Tree Dalam Memprediksi Penjualan Bahan Bakar Minyak: The Comparison of K-Means, Naïve Bayes and Decision Tree Algorithm in Predicting Fuel Oil Sales," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1379–1389, 2024, doi: https://doi.org/10.57152/malcom.v4i4.1566.

[12] L. Best, E. Foo, and H. Tian, "A Hybrid Approach: Utilising Kmeans Clustering and Naive Bayes for IoT Anomaly Detection," *arXiv preprint arXiv:2205.04005*, 2022, doi: https://doi.org/10.1007/978-3-031-08270-2_7.

[13] I. Kurniawan and A. Susanto, "Implementasi metode K-Means dan Naïve Bayes Classifier untuk analisis sentimen pemilihan presiden (pilpres) 2019," *Jurnal Eksplora Informatika Vol*, vol. 9, no. 1, 2019, doi: https://doi.org/10.30864/eksplora.v9i1.237.

[14] B. Budiman and I. Umami, "Implementasi algoritma Naïve Bayes untuk menentukan calon penerima beasiswa di SMK YPM 14 Sumobito Jombang," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 4, no. 2, pp. 446–454, 2022, doi: https://doi.org/10.47233/jteksis.v4i2.570.

[15] A. E. Rahayu, K. Hikmah, N. Yustia, and A. C. Fauzan, "Penerapan K-Means Clustering Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa," *ILKOMNIKA*, vol. 1, no. 2, pp. 82–86, 2019, doi: https://doi.org/10.28926/ilkomnika.v1i2.23.

[16] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and," *Techniques, Waltham: Morgan Kaufmann Publishers*, 2012.

[17] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit*, vol. 46, no. 1, pp. 243–256, 2013, doi: https://doi.org/10.1016/j.patcog.2012.07.021.

[18] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.

[19] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017, doi: https://doi.org/10.1016/j.neucom.2017.06.053.
[20] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimization on machine learning for prediction," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 6, no. 4, pp. 2407–2414, 2022, doi: https://doi.org/10.33395/sinkron.v7i4.11792.
[21] Z. R. Tembusai, H. Mawengkang, and M. Zarlis, "K-nearest neighbor with k-fold cross validation and analytic hierarchy process on data classification," *International Journal of Advances in Data and Information Systems*, vol. 2, no. 1, pp. 1–8, 2021, doi: https://doi.org/10.25008/ijadis.v2i1.1204.
[22] P. Domingos, "A few useful things to know about machine learning," *Commun ACM*, vol. 55, no. 10, pp. 78–87, 2012, doi: https://doi.org/10.1145/2347736.2347755.
[23] E. Helmud, F. Fitriyani, and P. Romadiana, "Classification comparison performance of supervised machine learning random forest and decision tree algorithms using confusion matrix," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 1, pp. 92–97, 2024, doi: https://doi.org/10.32736/sisfokom.v13i1.1985.
[24] M. Ahsan and W. Harianto, "Komparasi Tingkat Akurasi Information Gain Dan Gain Ratio Pada Metode K-Nearest Neighbor," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 1, pp. 386–391, 2022, doi: https://doi.org/10.36040/jati.v6i1.4694.
[25] M. S. A. D. A. S. A. S. Kurnia Muludi, "Sentiment Analysis Of Energy Independence Tweets Using Simple Recurrent Neural Network," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 4, pp. 2460–7258, 2021, doi: https://doi.org/10.22146/ijccs.66016.