



Classification of Poor Student Assistance Recipients Using Naive Bayes in the RI–RDTL Border Region

Klasifikasi Penerimaan Bantuan Siswa Miskin Menggunakan Naive Bayes di Wilayah Perbatasan RI-RDTL

**Madalena Bada Mau^{1*}, Krisantus J. Tey Seran²,
Dian Grace Ludji³, Regolinda Maneno⁴**

^{1,2,3,4}Program Studi Teknologi Informasi, Fakultas Pertanian, Sains dan Kesehatan,
Universitas Timor, Indonesia

E-Mail: ¹maulena025@gmail.com, ²krisantusteyseran@unimor.ac.id,
³dianludji@unimor.ac.id, ⁴regolinda@unimor.ac.id

Received Apr 16th 2026; Revised May 16th 2026; Accepted Jun 03rd 2026; Available Online Jun 20th 2026

Corresponding Author: Madalena Bada Mau

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The Poor Student Assistance (BSM) program aims to support students from low-income families in continuing their education. However, in practice, determining aid recipients still faces challenges related to subjectivity and the suboptimal use of real-world data. This study develops a classification model using the Naïve Bayes algorithm to predict the eligibility of junior high school students for financial assistance in the Indonesia–Timor Leste border region, specifically in Belu Regency. The data used in this study are secondary data obtained from the Department of Education for the period 2022–2024, comprising 2,607 students who received BSM assistance. The research stages include data preprocessing, categorical data transformation, normalization, data splitting, model development, and model evaluation using Holdout validation and K-fold cross-validation. The results show that the K-Fold cross-validation method achieves an average accuracy of 87.69%, which is higher than the Holdout validation accuracy of 87.16%. In addition, the highest accuracy in a given fold reaches 99.61%, indicating that the model can achieve optimal performance under certain data partitioning conditions. Overall, these findings confirm that the model has strong classification performance and the potential to serve as a decision-support system for determining education aid recipients in a more objective and targeted manner.

Keywords: BSM, Classification, Holdout Validation, K-Fold Cross-Validation, Naive Bayes

Abstrak

Program Bantuan Siswa Miskin (BSM) bertujuan membantu siswa dari keluarga kurang mampu agar tetap mengenyam pendidikan di bangku sekolah. Sampai saat ini, realita penentuan penerima bantuan masih menghadapi permasalahan subjektivitas dan kurang optimalnya pemanfaatan data riil di lapangan. Dalam penelitian ini membangun model klasifikasi menggunakan Algoritma Naïve Bayes untuk memprediksi kelayakan penerima bantuan siswa miskin tingkat SMP wilayah perbatasan Indonesia dan Timor Leste yakni Kabupaten Belu. Data yang digunakan merupakan data sekunder dari Dinas Pendidikan pada tahun 2022–2024 dan terdapat 2.607 siswa yang telah menerima bantuan BSM. Tahapan penelitian ini meliputi *preprocessing* data, transformasi data kategorik, normalisasi, *split* data, model metode serta evaluasi model menggunakan teknik *Holdout validation* dan *K-Fold cross-validation*. Hasil penelitian menunjukkan bahwa metode *K-Fold cross-validation* menghasilkan rata-rata akurasi sebesar 87,69%, lebih tinggi dibandingkan *Holdout validation* sebesar 87,16%. Selain itu, pada *fold* terbaik diperoleh akurasi sebesar 99,61%, yang menunjukkan bahwa model dapat mencapai performa optimal pada kondisi tertentu dan hasil ini menegaskan bahwa model memiliki kemampuan klasifikasi yang baik dan berpotensi digunakan sebagai sistem pendukung keputusan dalam menentukan penerima bantuan pendidikan secara lebih objektif dan tepat sasaran.

Kata Kunci: BSM, *Holdout validation*, *K-Fold cross-validation*, Klasifikasi, Naive Bayes

1. PENDAHULUAN

Pendidikan merupakan hak dasar yang dijamin oleh negara dan menjadi fondasi utama dalam pembangunan sumber daya manusia (SDM) [1][2]. Kondisi masyarakat yang menyenjam pendidikan di Indonesia saat ini mengalami peningkatan secara nasional, namun kesenjangan ekonomi masih menjadi tantangan utama dalam mewujudkan pemerataan pendidikan, khususnya bagi masyarakat yang berada di daerah terpencil. Data Badan Pusat Statistik (BPS), sekitar 9,36% penduduk Indonesia masih hidup di bawah garis kemiskinan, yang berdampak signifikan terhadap akses pendidikan anak-anak dari keluarga kurang mampu, khususnya pada jenjang Sekolah Menengah Pertama (SMP) [3][4]. Untuk mengatasi permasalahan tersebut, pemerintah mengembangkan berbagai program bantuan pendidikan, salah satunya program Bantuan Siswa Miskin (BSM) yang bertujuan membantu siswa dari keluarga kurang mampu agar tetap dapat melanjutkan pendidikan di sekolah.

Salah satu daerah yang mendapatkan BSM dari pemerintah adalah Kabupaten Belu yang terletak di Provinsi Nusa Tenggara Timur (NTT). Kabupaten ini berbatasan langsung dengan Negara Timor Leste dan masuk dalam kategori daerah Tertinggal, Terdepan, dan Terluar (3T). Kondisi ini ditandai dengan keterbatasan infrastruktur, akses teknologi, serta fasilitas pendidikan yang masih belum merata dan tentunya memerlukan perhatian khusus dari pemerintah daerah dan pusat. Pemerintah Kabupaten Belu telah meluncurkan berbagai program bantuan pendidikan guna mengurangi beban biaya pendidikan masyarakat, termasuk program BSM yang difokuskan pada siswa dari keluarga berpenghasilan rendah [5]. Namun demikian, proses penyaluran bantuan masih menghadapi permasalahan dalam ketepatan sasaran penerima, sehingga berpotensi menyebabkan ketidakefisienan dalam alokasi bantuan [6][7]. Beberapa kasus menunjukkan bahwa bantuan belum sepenuhnya diterima oleh siswa yang paling membutuhkan, khususnya dalam wilayah kabupaten ini.

Menanggapi masalah di atas, penelitian ini berfokus pada analisis penerima BSM sesuai dengan kriteria yang telah ditetapkan oleh pemerintah. Dengan jumlah penerima sebanyak 2.607 siswa/i, diperlukan model analisis data untuk skala yang besar. Salah satu pendekatan yang dapat diterapkan adalah teknik penggalian data (*data mining*). Dalam *data mining* ada sejumlah algoritma yang bisa digunakan dalam mengklasifikasikan data berdasarkan kategori tertentu. Algoritma Naive Bayes merupakan satu dari banyak algoritma klasifikasi pembelajaran mesin (*machine learning*), yang mampu melakukan analisis data secara otomatis berdasarkan pola tertentu. Penerapan Naive Bayes telah banyak digunakan karena terbukti dalam menangani data kategorikal serta efisiensi komputasi yang baik pada *dataset* berukuran besar [8][9]. Dalam evaluasi model, metode *Holdout validation* digunakan dengan membagi data menjadi data *training* dan data *testing*, sedangkan metode *K-Fold cross-validation* membagi data menjadi beberapa bagian (*fold*) untuk diuji secara bergantian sehingga menghasilkan evaluasi yang lebih stabil dan optimal.

Penelitian sebelumnya [10] menunjukkan bahwa metode Naive Bayes mampu mengidentifikasi penerima bantuan pendidikan dengan tingkat akurasi lebih dari 80%, sehingga bisa digunakan sebagai sistem pendukung keputusan dalam sektor pendidikan. Namun demikian, penelitian tersebut masih memiliki keterbatasan karena hanya menggunakan metode *Holdout validation* dan berfokus pada akurasi tanpa mempertimbangkan kestabilan serta kemampuan generalisasi model. Selain itu, *dataset* yang digunakan relatif terbatas sehingga belum sepenuhnya merepresentasikan kondisi nyata di lapangan. Oleh karena itu, penelitian ini berupaya memperbaiki keterbatasan tersebut dengan menggunakan jumlah data yang lebih besar serta menerapkan perbandingan metode evaluasi, yaitu *Holdout validation* dan *K-Fold cross-validation*, untuk memperoleh hasil yang lebih akurat, stabil, dan representatif.

Penelitian ini menerapkan Algoritma Naive Bayes untuk mengklasifikasi penerimaan BSM tingkat SMP di Kabupaten Belu menggunakan empat atribut utama, yaitu pendapatan orang tua, pekerjaan orang tua, jumlah tanggungan keluarga, dan status rumah. Algoritma Naive Bayes dipilih karena mampu mengolah data kategorikal dengan baik, memiliki proses perhitungan yang sederhana, serta efisiensi komputasi yang tinggi. Selain itu, Naive Bayes tidak memerlukan proses pelatihan yang kompleks dan tetap mampu memberikan hasil klasifikasi yang baik meskipun digunakan pada *dataset* berukuran besar, sehingga cocok untuk permasalahan klasifikasi penerima bantuan pendidikan. Model dievaluasi menggunakan metode *Holdout validation* dengan pembagian data *training* sebesar 80% dan data *testing* sebesar 20%, serta metode *K-Fold cross-validation* dengan 10 *fold* untuk mengukur performa dan stabilitas model. Penelitian ini bertujuan untuk menganalisis dan mengklasifikasi penerima BSM secara lebih akurat dan objektif menggunakan algoritma Naive Bayes, serta membandingkan metode *Holdout validation* dan *K-Fold cross-validation* dalam mengevaluasi kinerja model guna memperoleh metode evaluasi yang paling optimal dan memiliki tingkat kestabilan yang baik.

2. METODOLOGI PENELITIAN

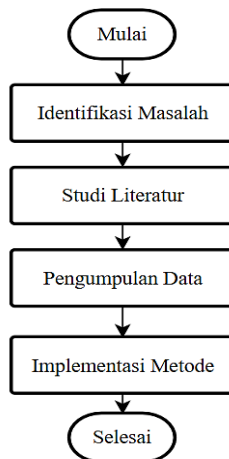
2.1. Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan data numerik yang dianalisis secara statistik dan komputasional untuk memperoleh hasil yang objektif dan sistematis [11]. Data yang digunakan merupakan data sekunder yang diperoleh dari Dinas Pendidikan Kabupaten Belu, Provinsi

NTT, yang berisi informasi penerimaan BSM di tingkat SMP. Kajian dalam penelitian kali ini difokuskan pada permasalahan ketidaktepatan sasaran penerima bantuan pendidikan di Kabupaten Belu.

2.2. Tahapan Penelitian

Tahapan penelitian dilakukan secara sistematis untuk membangun model klasifikasi kelayakan penerima Bantuan Siswa Miskin (BSM) menggunakan algoritma Naïve Bayes. Adapun tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Metodologi Penelitian.

2.2.1. Identifikasi Masalah

Pada tahap ini, peneliti mencari dan memahami masalah dalam proses penerimaan BSM di Kabupaten Belu, terutama terkait ketidaktepatan sasaran dan masih adanya subjektivitas. Selanjutnya, peneliti menentukan solusi dengan menggunakan Algoritma Naive Bayes untuk membantu proses klasifikasi penerima bantuan. Untuk mengetahui kinerja model, digunakan dua metode pengujian, yaitu *Holdout validation* dan *K-Fold cross-validation*. Tahap ini menjadi dasar dalam menentukan tujuan dan arah penelitian.

2.2.2. Studi Literatur

Pada tahap ini, peneliti mencari dan mempelajari berbagai referensi dari penelitian sebelumnya yang berkaitan dengan algoritma Naive Bayes, metode pembagian data seperti *Holdout validation* dan *K-Fold cross-validation*, serta evaluasi menggunakan *confusion matrix*.

2.2.3. Pengumpulan Data

Data dikumpulkan melalui pengajuan permohonan resmi kepada instansi terkait dengan melampirkan surat penelitian serta pernyataan kerahasiaan data untuk menjaga privasi siswa. Data yang diperoleh merupakan data sekunder yang diambil pada tahun 2024, dengan cakupan periode tiga tahun terakhir, yaitu 2022–2024. Data yang didapatkan berjumlah 2.607 siswa, yang mencakup informasi sosial ekonomi seperti pendapatan orang tua, pekerjaan orang tua, jumlah tanggungan keluarga, dan status rumah. Data tersebut digunakan sebagai dasar dalam proses klasifikasi untuk menentukan kelayakan penerima bantuan BSM.

2.2.4. Implementasi Metode

Pada tahap ini, proses klasifikasi dilakukan menggunakan algoritma Naive Bayes dengan bantuan bahasa pemrograman Python melalui Jupyter Notebook. *Dataset* yang telah dikumpulkan terlebih dahulu diimpor dan diproses, seperti pembersihan data, penghapusan duplikat, penanganan data kosong, serta transformasi data kategorikal menjadi data numerik. Proses *preprocessing* dilakukan sebelum pembagian data (*data splitting*), sehingga data yang digunakan pada tahap pelatihan dan pengujian telah berada dalam kondisi bersih dan siap digunakan. Setelah itu, data dibagi menjadi data *training* dan data *testing* sesuai dengan metode evaluasi yang digunakan. Selanjutnya, data dinormalisasi agar setiap atribut memiliki skala yang seimbang. Dalam proses pemodelan, Naive Bayes bekerja dengan menghitung probabilitas setiap kelas berdasarkan data yang tersedia. Algoritma ini menentukan kemungkinan suatu data termasuk ke dalam kelas tertentu dengan melihat nilai probabilitas tertinggi dari hasil perhitungan. Secara matematis, prinsip dasar Naïve Bayes dinyatakan sebagai berikut:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \quad (1)$$

di mana:

$P(H|X)$: probabilitas posterior,
 $P(X|H)$: *likelihood*,
 $P(H)$: probabilitas prior, dan
 $P(X)$: *evidence*.

Selanjutnya, *dataset* dibagi menggunakan dua metode, yaitu *Holdout validation* dan *K-Fold cross-validation*. Pada metode *Holdout validation*, data dibagi menjadi data *training* sebesar 80% dan data *testing* sebesar 20% menggunakan fungsi *train_test_split* pada Python. Sedangkan pada metode *K-Fold cross-validation*, data dibagi menjadi 10 bagian (*fold*) menggunakan *KFold*, di mana setiap bagian secara bergantian digunakan sebagai data uji dan sisanya sebagai data pelatihan. Pendekatan ini memungkinkan model diuji secara lebih menyeluruh karena semua data memiliki kesempatan untuk menjadi data latih dan data uji.

Terakhir, dilakukan evaluasi model untuk mengukur kinerja klasifikasi yang dihasilkan. Evaluasi dilakukan menggunakan *confusion matrix* dengan membandingkan hasil prediksi model dengan label aktual data uji. Akurasi dapat dihitung menggunakan tabel *Confusion matrix* dengan persamaan 2.

$$Akurasi = \left(\frac{TP+TN}{(TP+TN+FP+FN)} \right) \times 100 \quad (2)$$

Keterangan:

TP (*True Positive*) : Jumlah data yang layak dan diprediksi layak.
 TN (*True Negative*) : Jumlah data yang tidak layak dan diprediksi tidak layak.
 FP (*False Positive*) : Jumlah data yang tidak layak tetapi diprediksi sebagai layak.
 FN (*False Negative*) : Jumlah data yang layak tetapi diprediksi tidak layak.

Dari hasil perbandingan tersebut diperoleh nilai TP, TN, FP, dan FN, yang kemudian digunakan untuk menghitung metrik evaluasi berupa *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil evaluasi ditampilkan dalam bentuk tabel dan visualisasi untuk mempermudah analisis performa model serta memastikan model memiliki kemampuan klasifikasi yang baik.

3. HASIL DAN PEMBAHASAN

3.1. Hasil

Data siswa tingkat SMP di Kabupaten Belu yang berkaitan dengan penerimaan Bantuan Siswa Miskin (BSM) periode 2022–2024 dengan jumlah total sebanyak 2.607 data masih berupa data mentah (*raw data*). Oleh karena itu, data tersebut perlu diolah terlebih dahulu sebelum digunakan dalam proses klasifikasi pada tahap selanjutnya. Tabel 1 merupakan sampel dari data yang telah diperoleh.

Tabel 1. Data siswa BSM tahun 2022 - 2024

No	Kode Peserta	Pendapatan Orang Tua	Pekerjaan Orang Tua	Jumlah Tanggungan	Status Rumah	Label
1	A1	Rp500.000	Petani	3	Milik Sendiri	Ya
2	A2	Rp500.000	Petani	3	Milik Sendiri	Ya
3	A3	Rp500.000	Petani	4	Milik Sendiri	Ya
4	A4	Rp500.000	Wiraswasta	4	Milik Sendiri	Tidak
...	...	...	...	...	...	...
2605	A2605	Rp500.000	Petani	2	Kontrak/sewa	Ya
2606	A606	Rp500.000	Petani	2	Kontrak/sewa	Ya
2607	A2607	Rp500.000	Petani	2	Milik Sendiri	Ya

Data mentah (*raw data*) diimpor ke dalam *Jupyter Notebook* menggunakan format Excel dengan bantuan *pandas*. Proses klasifikasi Naïve Bayes didukung oleh beberapa *library* seperti *numpy* untuk perhitungan numerik, serta *matplotlib.pyplot* dan *seaborn* untuk visualisasi data. Pengolahan data dilakukan dengan mengubah data kategorikal menjadi numerik menggunakan *LabelEncoder* dan melakukan normalisasi menggunakan *StandardScaler*.

Pembagian data dilakukan menggunakan *train_test_split* untuk metode *Holdout validation*, serta *KFold*, *StratifiedKFold*, dan *cross_val_score* untuk validasi silang. Model klasifikasi dibangun menggunakan Algoritma *GaussianNB*, kemudian dievaluasi menggunakan *accuracy_score*, *confusion_matrix*, *classification_report*, dan *ConfusionMatrixDisplay*. Selain itu, digunakan metode SMOTE untuk menangani ketidakseimbangan data, serta *library time* untuk mengukur waktu proses pelatihan dan pengujian model.

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split, KFold, cross_val_score,
StratifiedKFold
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report, ConfusionMatrixDisplay
from imblearn.over_sampling import SMOTE
import time
# 1. Membaca Data Excel
df = pd.read_excel('Data Awal.xlsx')
df

```

Hasil *import* menampilkan *dataset* sebanyak 2607 baris dan 6 atribut, yaitu Nama Peserta Didik, Pendapatan Orang Tua, Pekerjaan Orang Tua, Jumlah Tanggungan, Status Rumah, dan Label, di mana atribut Label menunjukkan status kelayakan penerima bantuan.

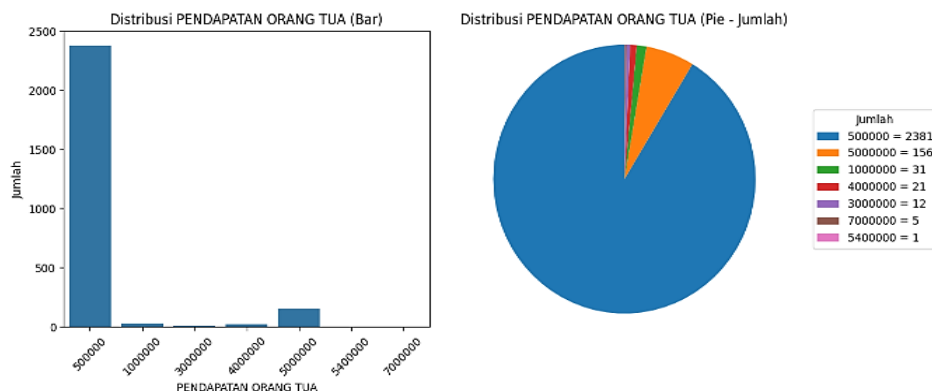
	NAMA PESERTA DIDIK	PENDAPATAN ORANG TUA	PEKERJAAN ORANG TUA	JUMLAH TANGGUNGAN	STATUS RUMAH	LABEL
0	Claudia Izabella Jessica Lay	500000	Petani	3	Milik Sendiri	Ya
1	Florentina Mendonca Araujo	500000	Petani	3	Milik Sendiri	Ya
2	Immanuel C.H Pius Antoin	500000	Petani	4	Milik Sendiri	Ya
3	Irvanto Nabu	500000	Wiraswasta	4	Milik Sendiri	Tidak
4	Joanico Amaral	500000	Petani	1	Milik Sendiri	Ya
...	...	...	...	...	...	...
2602	Cipriano Da Roza	500000	Wiraswasta	2	Kontrak/sewa	Ya
2603	Cristian Fandi Siri	500000	Wiraswasta	2	Kontrak/sewa	Ya
2604	Cristoforus Exposto	500000	Petani	2	Kontrak/sewa	Ya
2605	Dedyance Dasi	500000	Petani	2	Kontrak/sewa	Ya
2606	Faristo Aryanto Berek Wadan	500000	Petani	2	Milik Sendiri	Ya

2607 rows x 6 columns

Gambar 2. Hasil *Import dataset*

Data yang telah berhasil diimpor pada Gambar 2 kemudian dilakukan *preprocessing* untuk memastikan kualitas data sebelum digunakan dalam proses pemodelan. Tahap awal dilakukan pengecekan *missing value* untuk mengetahui apakah terdapat data kosong pada setiap atribut, dan hasilnya menunjukkan tidak ditemukan nilai kosong. Selanjutnya dilakukan pengecekan data duplikat dan tidak ditemukan data yang terduplikasi, sehingga seluruh data dapat digunakan tanpa perlu dilakukan penghapusan.

Setelah *dataset* dinyatakan bersih, dilakukan visualisasi data untuk melihat pola distribusi pada setiap atribut yang digunakan. Visualisasi ini bertujuan untuk memahami sebaran data serta melihat kecenderungan nilai pada masing-masing variabel. Hasil visualisasi ditampilkan dalam bentuk grafik sehingga memudahkan dalam interpretasi data.

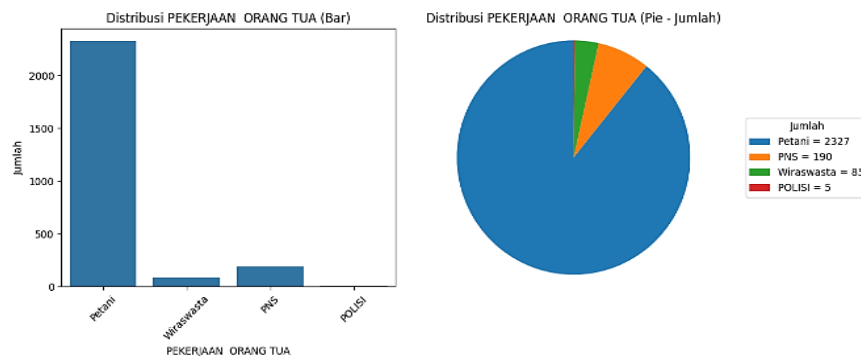


Gambar 3. Visualisasi Distribusi Variabel Pendapatan Orang Tua

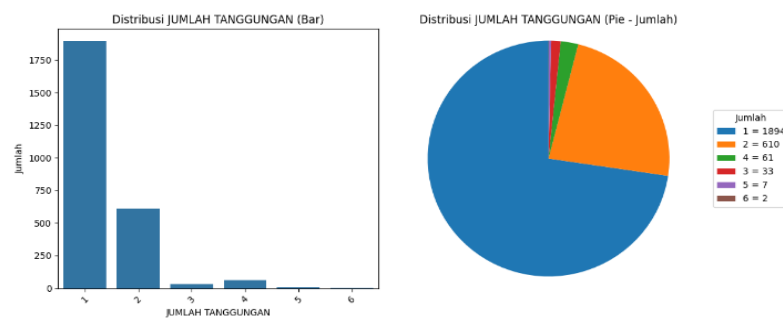
Gambar 3 merupakan visualisasi distribusi frekuensi variabel pendapatan orang tua berdasarkan jumlah siswa pada setiap kategori pendapatan. Parameter yang dianalisis adalah jumlah data siswa pada

masing-masing kategori pendapatan orang tua. Hasil visualisasi menunjukkan bahwa kategori pendapatan Rp500.000 memiliki jumlah data paling dominan dibandingkan kategori lainnya, yaitu sebanyak 2.381 data. Hal ini menunjukkan bahwa sebagian besar siswa berasal dari keluarga dengan tingkat pendapatan rendah.

Gambar 4 merupakan visualisasi distribusi variabel pekerjaan orang tua berdasarkan jumlah siswa pada setiap kategori pekerjaan. Parameter yang dianalisis adalah frekuensi atau jumlah data siswa pada masing-masing jenis pekerjaan orang tua. Hasil visualisasi menunjukkan bahwa kategori pekerjaan petani mendominasi *dataset* dengan jumlah terbesar dibandingkan kategori lainnya. Hal ini menggambarkan karakteristik sosial ekonomi masyarakat pada wilayah penelitian yang sebagian besar bekerja pada sektor pertanian, yang umumnya memiliki pendapatan tidak tetap.

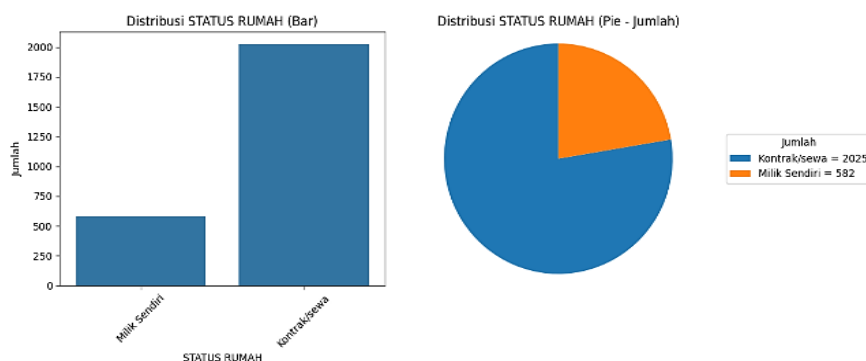


Gambar 4. Visualisasi Distribusi Variabel Pekerjaan Orang Tua



Gambar 5. Visualisasi Distribusi Variabel Jumlah Tanggungan

Pada Gambar 5 disajikan hasil analisis distribusi frekuensi variabel jumlah tanggungan keluarga. Berdasarkan visualisasi tersebut, terlihat bahwa mayoritas keluarga memiliki 1 hingga 2 tanggungan dengan jumlah paling dominan pada kategori 1 tanggungan. Distribusi ini menunjukkan bahwa karakteristik data didominasi oleh keluarga dengan jumlah tanggungan kecil hingga sedang. Kondisi ini menggambarkan tingkat beban ekonomi rumah tangga yang relatif beragam, yang dapat memengaruhi penilaian kelayakan dalam penerimaan bantuan.”

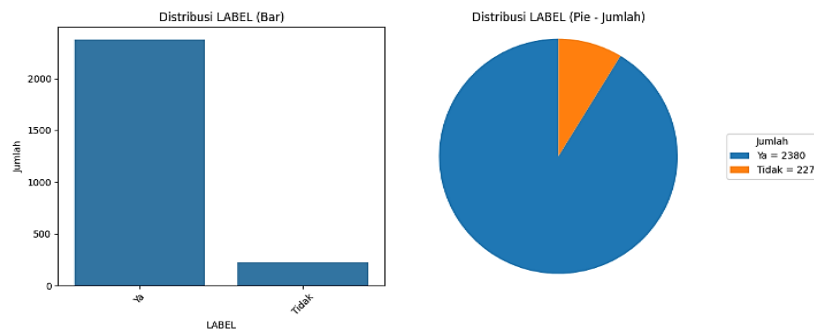


Gambar 6. Visualisasi Distribusi Status Rumah

Visualisasi pada Gambar 6 menunjukkan distribusi status kepemilikan rumah berdasarkan dua kategori, yaitu kontrak/sewa dan milik sendiri. Parameter yang dianalisis dalam visualisasi ini adalah jumlah (frekuensi) dan proporsi data pada masing-masing kategori status rumah. Hasil menunjukkan bahwa kategori

kontrak/sewa mendominasi dengan jumlah 2.025 rumah, sedangkan rumah milik sendiri sebanyak 582 rumah. Kondisi ini mengindikasikan bahwa sebagian besar keluarga peserta didik masih berada pada status tempat tinggal tidak permanen, yang dapat mencerminkan kondisi sosial ekonomi pada kelompok responden.

Sementara itu, distribusi label penerima bantuan pada Gambar 7 menunjukkan bahwa kategori “Ya” lebih dominan dibandingkan kategori “Tidak”. Parameter yang dianalisis pada visualisasi ini adalah variabel label penerima bantuan, yang merupakan variabel target dalam proses klasifikasi. Hasil visualisasi memperlihatkan bahwa jumlah data dengan label “Ya” sebanyak 2.380, sedangkan label “Tidak” sebanyak 227. Temuan ini menunjukkan bahwa mayoritas data dalam *dataset* termasuk dalam kategori penerima bantuan. Kondisi ini juga mengindikasikan adanya ketidakseimbangan data (*imbalanced data*), di mana jumlah kelas “Ya” jauh lebih besar dibandingkan kelas “Tidak”, yang berpotensi mempengaruhi kinerja model klasifikasi jika tidak ditangani dengan baik.



Gambar 7. Visualisasi Distribusi Variabel Label Penerima

Tahap berikutnya adalah transformasi data, yaitu mengubah data kategorikal menjadi numerik menggunakan *LabelEncoder*. Proses ini dilakukan agar data dapat diproses oleh Algoritma Naïve Bayes yang membutuhkan input berupa angka. Selanjutnya dilakukan seleksi fitur dengan menentukan atribut yang relevan sebagai variabel input, yaitu pendapatan orang tua, pekerjaan orang tua, jumlah tanggungan, dan status rumah, serta label sebagai variabel target. Setelah tahap *preprocessing* selesai, *dataset* kemudian dibagi menjadi data latih dan data uji untuk keperluan pemodelan dan evaluasi. Pada penelitian ini digunakan dua metode pembagian data, yaitu *Hold-out validation* dan *K-Fold cross-validation*.

Pada hasil pembagian data menggunakan metode *Hold-out validation*, *dataset* berhasil dipisahkan menjadi data latih dan data uji dengan komposisi 80% dan 20%. Hasil proses menunjukkan bahwa sebanyak 2.085 data digunakan sebagai data latih awal, sedangkan 522 data digunakan sebagai data uji. Proporsi kelas pada kedua kelompok data tetap terjaga sehingga distribusi label penerima dan bukan penerima bantuan tetap merepresentasikan kondisi *dataset* asli. Selanjutnya dilakukan proses penyeimbangan data pada data latih menggunakan metode SMOTE. Hasil penerapan SMOTE menunjukkan adanya peningkatan jumlah data latih menjadi 3.806 data, yang menandakan bahwa kelas minoritas berhasil diseimbangkan dengan kelas mayoritas. Proses ini menghasilkan distribusi kelas yang lebih proporsional sehingga model memiliki kesempatan belajar yang sama terhadap setiap kategori kelas. Hasil *split* data teknik *holdout validation* dapat dilihat pada Gambar 8.

```

=== DATA LATIH SEBELUM SMOTE ===
PENDAPATAN ORANG TUA  PEKERJAAN  ORANG TUA  JUMLAH TANGGUNGAN  \
0      -0.290857      0.205752      0.990018
1      -0.290857      0.205752      0.990018
2      -0.290857      0.205752     -0.519325
3      -0.290857      0.205752     -0.519325
4      -0.290857      0.205752     -0.519325

STATUS RUMAH  LABEL  TIPE_DATA
0      -0.536104      1  Data Latih (Sebelum SMOTE)
1      -0.536104      1  Data Latih (Sebelum SMOTE)
2      -0.536104      1  Data Latih (Sebelum SMOTE)
3      1.865310      1  Data Latih (Sebelum SMOTE)
4      -0.536104      1  Data Latih (Sebelum SMOTE)
Jumlah data latih sebelum SMOTE : 2085

=== DATA LATIH SETELAH SMOTE ===
PENDAPATAN ORANG TUA  PEKERJAAN  ORANG TUA  JUMLAH TANGGUNGAN  \
0      -0.290857      0.205752      0.990018
1      -0.290857      0.205752      0.990018
2      -0.290857      0.205752     -0.519325
3      -0.290857      0.205752     -0.519325
4      -0.290857      0.205752     -0.519325
...
4      1.865310      1  Data Uji
Jumlah data uji : 522

File berhasil disimpan: hasil_zscore_smote_split.xlsx

```

Gambar 8. Hasil *Split* Data Teknik *Holdout validation*

Sedangkan Hasil pembagian pada teknik *K-Fold cross-validation* dengan $k = 10$ pada Gambar 9 menunjukkan bahwa *dataset* berhasil dibagi menjadi sepuluh bagian dengan ukuran yang relatif seimbang, di mana setiap bagian secara bergantian digunakan sebagai data uji sementara bagian lainnya berperan sebagai data latih. Pada setiap *fold*, proporsi kelas tetap terjaga sehingga distribusi label penerima bantuan tetap konsisten di seluruh proses validasi. Setiap data memperoleh kesempatan yang sama untuk menjadi data latih maupun data uji, sehingga hasil evaluasi tidak bergantung pada satu pembagian data tertentu saja.

```

Fold 1:
Indeks Train: [0, 1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41,
Indeks Test : [3, 10, 17, 33, 74, 78, 80, 83, 99, 100, 108, 113, 118, 142, 150, 159, 180, 186, 207, 209, 210, 217, 247, 269, 274, 277, 279, 282, 286, 288, 320]

Fold 2:
Indeks Train: [0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40,
Indeks Test : [9, 13, 39, 67, 94, 98, 103, 106, 109, 110, 121, 125, 128, 137, 147, 153, 155, 156, 161, 173, 175, 176, 178, 188, 190, 192, 205, 206, 220, 230,

Fold 3:
Indeks Train: [0, 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39,
Indeks Test : [6, 16, 55, 56, 59, 71, 92, 104, 144, 149, 151, 163, 167, 174, 181, 196, 214, 218, 226, 232, 242, 257, 270, 273, 284, 296, 298, 299, 300, 302, 3

Fold 4:
Indeks Train: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42,
Indeks Test : [14, 19, 24, 31, 38, 44, 52, 60, 63, 66, 68, 70, 88, 123, 146, 166, 183, 184, 203, 234, 240, 243, 244, 246, 253, 262, 266, 280, 287, 293, 295,

Fold 5:
Indeks Train: [0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 24, 27, 28, 29, 31, 32, 33, 34, 35, 37, 38, 39, 41, 43, 44, 45, 47,
Indeks Test : [5, 18, 23, 25, 26, 30, 36, 40, 42, 46, 53, 65, 72, 134, 140, 160, 162, 164, 199, 204, 227, 231, 245, 251, 256, 258, 260, 271, 305, 307, 312, 3

Fold 6:
Indeks Train: [1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 42, 43, 44,
Indeks Test : [0, 8, 11, 22, 27, 34, 41, 47, 49, 64, 73, 85, 95, 117, 126, 129, 131, 132, 141, 171, 185, 191, 194, 197, 201, 208, 212, 221, 225, 236, 238, 268
...

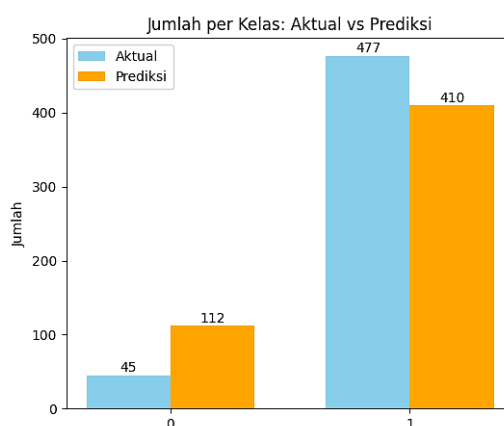
Waktu total K-Fold 10 + SMOTE + simpan indeks: 20.5846 detik
File berhasil disimpan: hasil_kfold_10_smote_index.xlsx

```

Gambar 9. Hasil split data teknik 10-fold cross-validation

Setelah proses pembagian data selesai, tahap selanjutnya adalah membangun model klasifikasi menggunakan Algoritma Naïve Bayes. Evaluasi performa model dilakukan menggunakan dua skenario pengujian, yaitu *Hold-out validation* dan *K-Fold cross-validation*, sehingga kemampuan model dapat dianalisis secara menyeluruh. Pada skenario *Hold-out validation*, *dataset* dibagi menjadi data latih dan data uji. Model dilatih menggunakan data latih yang telah melalui proses penyeimbangan kelas dengan metode SMOTE, kemudian diuji menggunakan data yang tidak terlibat pada proses pelatihan.

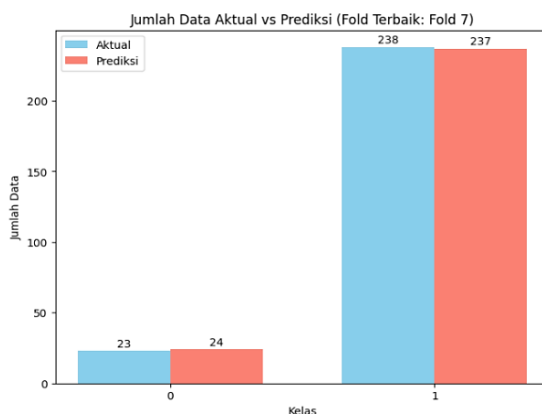
Pada penelitian ini, label kelas dibagi menjadi dua kategori, yaitu kelas 0 yang menunjukkan siswa tidak layak menerima bantuan, dan kelas 1 yang menunjukkan siswa layak menerima bantuan. Penentuan label ini digunakan sebagai acuan dalam proses klasifikasi untuk membedakan tingkat kelayakan penerima BSM. Pada Gambar 10 hasil klasifikasi menunjukkan terdapat 45 data kelas aktual 0 yang seluruhnya berhasil diklasifikasikan dengan benar sebagai kelas 0. Sementara itu, pada kelas aktual 1 sebanyak 477 data, model berhasil memprediksi 410 data secara benar, sedangkan 67 data lainnya mengalami kesalahan klasifikasi menjadi kelas 0. Secara keseluruhan, model menghasilkan 112 prediksi kelas 0 dan 410 prediksi kelas 1. Secara keseluruhan, distribusi hasil prediksi mendekati distribusi data aktual, yang menunjukkan bahwa model telah mampu mempelajari pola data dengan baik. Visualisasi rekapitulasi hasil klasifikasi dalam bentuk diagram batang memperlihatkan kesesuaian jumlah antara kelas aktual dan kelas prediksi. Perbedaan jumlah yang relatif kecil menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data baru.



Gambar 10. Hasil klasifikasi teknik *Holdout validation*

Sedangkan hasil klasifikasi dengan skenario 10-fold cross validation pada Gambar 11 menunjukkan bahwa rata-rata akurasi keseluruhan mencapai 0,8769 (87,69%), menandakan bahwa model memiliki

kemampuan generalisasi yang baik terhadap *dataset* dan *fold* terbaik diperoleh pada *Fold* ke-7 dengan nilai akurasi tertinggi sebesar 0,9962, yang menunjukkan bahwa pada pembagian data tertentu model mampu mengklasifikasikan hampir seluruh data dengan benar. Stabilitas nilai akurasi antar *fold* mengindikasikan bahwa model tidak bergantung pada satu pola pembagian data saja. Berdasarkan *fold* terbaik pada kelas 0, jumlah data aktual sebanyak 23 dan hasil prediksi sebanyak 24 data, sedangkan pada kelas 1, jumlah data aktual sebanyak 238 dan hasil prediksi sebanyak 237 data. Perbedaan yang sangat kecil ini menunjukkan tingkat kesalahan klasifikasi yang rendah. Visualisasi diagram batang pada *fold* terbaik semakin memperjelas bahwa distribusi prediksi hampir identik dengan distribusi data aktual. Hal ini menegaskan bahwa model Naïve Bayes mampu melakukan klasifikasi secara konsisten dan stabil ketika diuji menggunakan validasi silang.



Gambar 11. Hasil Klasifikasi Teknik 10-Fold cross-validation

Berdasarkan hasil klasifikasi menggunakan metode *Holdout validation* dan 10-Fold cross-validation, selanjutnya dilakukan perbandingan untuk melihat metode evaluasi mana yang memberikan performa terbaik. Perbandingan ini dilakukan dengan melihat nilai akurasi, tingkat kesalahan klasifikasi, serta kestabilan model dalam memprediksi data (lihat Tabel 2).

Tabel 2. Perbandingan hasil evaluasi model

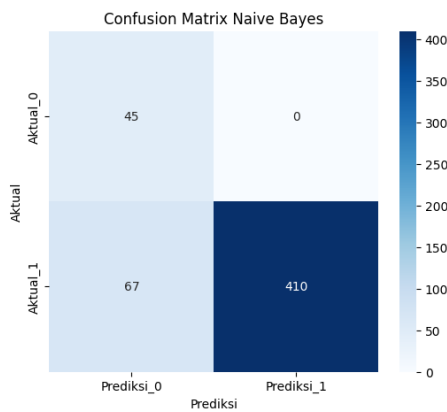
No	Metode Evaluasi	Kelas 0 Aktual	Kelas 0 Prediksi	Kelas 1 Aktual	Kelas 1 Prediksi
1	<i>Holdout validation</i> (80:20)	45	45	477	410
3	<i>K-Fold</i> (Fold Terbaik)	23	24	238	237

Analisis lebih lanjut setelah dilakukan evaluasi model, yakni mengukur kinerja dari kedua teknik *Holdout validation* dan *k-fold cross* dengan memanfaatkan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *f1-score* untuk mengetahui kemampuan klasifikasi model Naive Bayes secara menyeluruh.

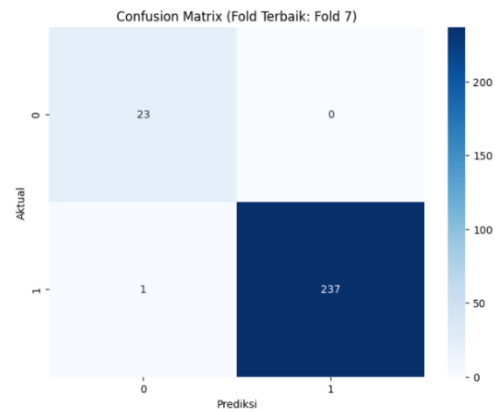
Pada *Hold-out validation* Gambar 12, *dataset* dibagi menjadi data *training* dan data *testing*. Hasil evaluasi menunjukkan model berhasil mengklasifikasikan 455 dari 522 data dengan benar dengan tingkat akurasi sekitar 87%. *Confusion matrix* menunjukkan 45 *True Negative* (TN) dan 410 *True Positive* (TP), tanpa *False Positive* (FP), serta 67 *False Negative* (FN). Hasil ini menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik, meskipun masih terdapat kesalahan prediksi pada sebagian data kelas positif.

Selanjutnya, evaluasi menggunakan *K-Fold cross-validation* (Gambar 13) dilakukan dengan membagi data ke dalam beberapa *fold* sehingga setiap data memiliki kesempatan menjadi data uji. Hasil pengujian menunjukkan bahwa *fold* ke-7 menghasilkan performa terbaik dengan akurasi tertinggi. *Confusion matrix* pada *fold* terbaik menghasilkan 23 TN, 237 TP, 0 FP, dan 1 FN, yang menunjukkan tingkat kesalahan prediksi yang sangat rendah. Nilai *precision*, *recall*, dan *F1-score* pada kedua kelas berada pada kategori sangat tinggi, sehingga performa model dapat dikatakan optimal dan stabil.

Terlihat dari hasil visualisasi pada Gambar 13 bahwa metode *K-Fold cross-validation* memiliki jumlah kesalahan klasifikasi yang lebih rendah dibandingkan dengan metode *Holdout validation*. Hal ini ditunjukkan dari nilai FN yang jauh lebih kecil, yaitu hanya 1 data pada *K-Fold* dibandingkan dengan 67 data pada *Holdout*. Selain itu, kedua metode tidak menghasilkan FP, yang menunjukkan bahwa model tidak melakukan kesalahan dalam memprediksi kelas negatif. Dengan demikian, metode *K-Fold cross-validation* menunjukkan performa yang lebih baik dalam hal akurasi dan ketepatan klasifikasi, karena mampu meminimalkan kesalahan prediksi serta menghasilkan model yang lebih stabil dibandingkan metode *Holdout validation*.



Gambar 12. Confusion matrix pada Holdout validation



Gambar 13. Confusion matrix pada 10-fold cross-validation

Untuk mengetahui kemampuan hasil model pada data yang belum diketahui kelasnya, dilakukan prediksi terhadap 150 data baru dengan dua teknik evaluasi, yaitu *Holdout validation* dan *K-Fold cross-validation*. Berdasarkan hasil prediksi data baru, *Holdout validation* memberikan hasil pada kelas 0 berjumlah 31 data dan kelas 1 berjumlah 119 data. Hal ini menunjukkan bahwa mayoritas data baru diprediksi masuk ke kelas 1 atau kategori layak, yaitu sekitar 79,3% dari total data, sedangkan kelas 0 atau kategori tidak layak hanya sekitar 20,7%.

Sementara itu, pada metode *K-Fold cross-validation*, visualisasi juga menunjukkan distribusi hasil prediksi data baru ke dalam dua kelas. Hasilnya memperlihatkan bahwa sebagian besar data kembali diprediksi masuk ke kelas 1 atau layak diterima, sedangkan kelas 0 atau tidak layak diterima memiliki jumlah yang sangat sedikit, bahkan hampir tidak ada data yang terklasifikasi. Secara keseluruhan, kedua metode menunjukkan pola hasil yang serupa, yaitu dominasi prediksi pada kelas 1 atau kelas yang layak. Hal ini mengindikasikan bahwa sebagian besar data baru memiliki karakteristik yang cenderung masuk ke dalam kategori penerima bantuan, sesuai dengan pola yang telah dipelajari oleh model sebelumnya.

3.2. Pembahasan

Dari hasil penelitian di atas yang menggunakan Algoritma Naïve Bayes untuk mengklasifikasikan kelayakan penerima BSM tingkat SMP di Kabupaten Belu dengan data tahun 2022–2024 sebanyak 2.607 siswa, *dataset* yang digunakan terdiri dari beberapa atribut sosial ekonomi seperti pendapatan orang tua, pekerjaan orang tua, jumlah tanggungan keluarga, serta status kepemilikan rumah yang dijadikan dasar dalam proses klasifikasi. Tahapan penelitian dimulai dari proses prapengolahan data yang mencakup pembersihan data, pengubahan data kategorik menggunakan label *encoding*, normalisasi data, serta penanganan ketidakseimbangan kelas dengan metode SMOTE yang diterapkan setelah proses pembagian data (*train-test split*) pada data latih untuk mengurangi dominasi kelas mayoritas, sehingga model tidak cenderung bias terhadap kelas tertentu, meskipun demikian masih terindikasi adanya ketidakseimbangan data awal yang berpotensi menyebabkan model lebih cenderung memprediksi kelas mayoritas serta indikasi *overfitting* pada hasil *K-Fold* tertentu yang menunjukkan akurasi sangat tinggi dibandingkan *hold-out validation*, dimana sebelum penerapan SMOTE performa model cenderung lebih rendah dan kurang seimbang terutama pada kelas minoritas, sedangkan setelah SMOTE terjadi peningkatan kemampuan model dalam mengenali kedua kelas secara lebih proporsional dengan peningkatan yang lebih terlihat pada *recall* dan *f1-score* kelas minoritas dibandingkan kondisi sebelum dilakukan *balancing data*, namun tetap perlu diwaspadai bahwa peningkatan signifikan pada *Fold* tertentu dapat mengindikasikan model terlalu menyesuaikan pola data latih. Langkah ini dilakukan agar kualitas data menjadi lebih baik sehingga model dapat menangkap pola dengan lebih maksimal. Setelah itu, proses pemodelan dilakukan menggunakan Algoritma Gaussian Naïve Bayes, sedangkan evaluasi kinerja model menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *f1-score*.

Hasil pengujian menunjukkan bahwa model memiliki performa yang cukup baik pada kedua metode evaluasi. Pada *Hold-out validation* diperoleh akurasi sebesar 87,16%, sedangkan pada *K-Fold cross-validation*, khususnya pada *Fold* ke-7, akurasi mencapai 99,61%, dimana perbedaan ini juga memperkuat indikasi adanya variasi performa yang cukup besar antar skenario pengujian yang dapat dipengaruhi oleh distribusi data dan efek SMOTE pada data latih. Secara umum, hasil ini menunjukkan bahwa Algoritma Naïve Bayes mampu memberikan kinerja klasifikasi yang tinggi dan cukup konsisten, sehingga dapat dijadikan metode yang efektif dalam membantu proses penentuan penerima BSM secara lebih objektif dan berbasis data.

Selain itu, pengujian pada data baru menunjukkan bahwa model mampu melakukan generalisasi dengan baik, di mana sebagian besar data diprediksi masuk ke kelas layak menerima bantuan. Hasil ini

diperoleh melalui perhitungan probabilitas Naïve Bayes yang menggabungkan nilai prior dan *likelihood* dari setiap atribut, kemudian memilih kelas dengan probabilitas tertinggi sebagai hasil akhir. Hasil tersebut menunjukkan bahwa model bekerja sesuai prinsip Naïve Bayes, yaitu menentukan kelas berdasarkan peluang terbesar, meskipun dominasi prediksi pada kelas layak juga tetap mengindikasikan adanya kecenderungan bias terhadap kelas mayoritas yang masih perlu dievaluasi lebih lanjut.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan Algoritma Naïve Bayes untuk memprediksi kelayakan penerima Bantuan Siswa Miskin (BSM) tingkat SMP wilayah perbatasan RI-RDTL di Kabupaten Belu menggunakan data tahun 2022–2024 sebanyak 2.607 siswa. Proses penelitian dilakukan melalui tahapan praproses data, penyeimbangan kelas menggunakan SMOTE, serta evaluasi model menggunakan metode *Holdout validation* dan *K-Fold cross-validation*. Hasil pengujian menunjukkan bahwa metode *K-Fold cross-validation* memberikan performa terbaik dengan tingkat akurasi sebesar 99,61%, lebih tinggi dibandingkan dengan *Holdout validation* yang memperoleh akurasi sebesar 87,16%. Hal ini menunjukkan bahwa penggunaan *K-Fold* mampu menghasilkan model yang lebih stabil karena data digunakan secara bergantian sebagai data pelatihan dan pengujian.

Model klasifikasi yang dihasilkan mampu mengidentifikasi kelayakan penerima bantuan secara efektif berdasarkan atribut sosial ekonomi siswa, sehingga dapat membantu proses pengambilan keputusan menjadi lebih objektif dan berbasis data. Implementasi metode Naïve Bayes pada penelitian ini menunjukkan bahwa pendekatan klasifikasi probabilistik dapat diterapkan sebagai sistem pendukung keputusan untuk menentukan penerima bantuan pendidikan.

Secara praktis, hasil penelitian ini berpotensi digunakan oleh pemerintah daerah maupun instansi pendidikan sebagai alat bantu untuk meningkatkan ketepatan sasaran program bantuan pendidikan. Untuk penelitian selanjutnya, disarankan melakukan perbandingan dengan algoritma klasifikasi lain atau menambahkan variabel pendukung yang lebih luas agar performa model dapat dianalisis secara lebih komprehensif serta meningkatkan kemampuan generalisasi model pada *dataset* yang lebih beragam.

REFERENSI

- [1] M. Ali, A. I. Ali, And A. R. Ali, “Transformasi Pendidikan Sebagai Pilar Pengembangan Sumber Daya Manusia Unggul Indonesia,” *Pros. Semin. Nas. Pembang. Ekon. Berkelanjutan Dan Ris. Ilmu Sos.*, Vol. 1, Pp. 1–12, 2025.
- [2] B. Sulfiani, Rahmatilla, A. T. R. Armita, A. Soraya, M. Edi, And B. D. Aksa, “Dampak Kualitas Pendidikan Terhadap Pengembangan Keterampilan Sumber Daya Manusia,” *J. Ilm. Pendidik. Dasar*, Vol. 10, No. September, 2025.
- [3] B. P. Statistik, *Statistik Indonesia 2023*. Jakarta: Badan Pusat Statistik, 2023.
- [4] I. M. G. Putri, H. Rahmayati, And G. S. Perdana, “Pengaruh Tingkat Literasi Dan Pendidikan Terhadap Tingkat Kemiskinan Di Indonesia , Serta Implikasi Peran Widyaiswara,” *Prosiding Pertem. Ilm. Tah. Nas. Widyaiswara*, Vol. 1, No. 122, Pp. 528–538, 2025.
- [5] R. Falah, “Implementasi Program Indonesia Pintar (Pip) Pada Tingkat Sma/Smk (Studi Kasus Pada Sma Negeri 1 Dan Smk Negeri 1 Di Kabupaten Majene Provinsi Sulawesi Barat),” Vol. 8, No. 2, 2024.
- [6] A. S. Ramadhani And Mansur, “Sistem Pendukung Keputusan Pemilihan Penerimaan Beasiswa Siswa Miskin (Bsm) Sekolah Menengah Atas Menggunakan Metode Multi Objective Optimization On The Basis Of Ratio Analysis (MOORA),” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 10, no. 2, pp. 2182–2188, 2026.
- [7] Mesran, S. Nastiti, K. Sussolaikah, I. Saputra, and D. P. Utomo, “Penerapan Multi Objective Optimization on the Basis of Ratio Analysis (MOORA) dalam Penentuan Bantuan Siswa Miskin (BSM) pada SMK,” *J. Decis. Support Syst. Res.*, vol. 2, no. 1, pp. 11–19, 2024.
- [8] Ulumuddin and P. Widodo, “Analisis Kinerja Algoritma Naive Bayes dalam Klasifikasi Data pada Pasien Tuberkulosis Berbasis Data Mining,” *J. Glob. Technol. Comput.*, vol. 5, no. 1, pp. 75–81, 2025, doi: 10.47065/jogtc.v5i1.8999.
- [9] M. A. Faradeya and E. R. Subhiyakto, “Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma Naive Bayes,” *J. Algoritm.*, vol. 22, no. 1, pp. 115–127, 2025, doi: 10.33364/algoritma/v.22-1.2178.
- [10] A. Nata and S. Suparmadi, “Analisis sistem pendukung keputusan dengan model klasifikasi berbasis machine learning dalam penentuan penerima Program Indonesia Pintar,” *Journal of Science and Social Research*, vol. 5, no. 3, p. 697, 2022, doi:10.54314/jssr.v5i3.1041.
- [11] M. Waruwu, S. Natijatul, P. R. Utami, and E. Yanti, “Metode Penelitian Kuantitatif: Konsep , Jenis , Tahapan dan Kelebihan,” *J. Ilm. profesi Pendidik.*, vol. 10, pp. 917–932, 2025.
- [12] N. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, “Implementasi algoritma Naive Bayes untuk analisis sentimen ulasan Shopee pada Google Play Store,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 1, pp. 47–54, 2022,

- doi:10.57152/malcom.v2i1.195.
- [13] T. Agustiranti *et al.*, “Penerapan Naive Bayes terhadap sentimen analisis media sosial Twitter pengguna Kereta Cepat Jakarta-Bandung (Whoosh),” *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 297–305, 2024, doi:10.55338/jikoms.v7i1.2946.
- [14] F. Alghifari and D. Juardi, “Penerapan data mining pada penjualan makanan dan minuman menggunakan algoritma Naïve Bayes,” *Jurnal Ilmiah Informatika*, vol. 9, no. 2, pp. 75–81, 2021.
- [15] A. Asri, A. Arifin, and W. Handoko, “Implementasi metode Naive Bayes untuk klasifikasi penerima program keluarga harapan,” *Journal of Computer*, vol. 2, no. 1, pp. 21–26, 2022.
- [16] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, “Implementasi data mining dengan algoritma Naïve Bayes untuk klasifikasi kelayakan penerima bantuan sembako,” *JURIKOM (Jurnal Riset Komputer)*, vol. 8, no. 6, p. 219, 2021, doi:10.30865/jurikom.v8i6.3655.
- [17] E. Etriyanti, D. Syamsuar, and N. Kunang, “Implementasi data mining menggunakan algoritme Naive Bayes classifier dan C4.5 untuk memprediksi kelulusan mahasiswa,” *Telematika*, vol. 13, no. 1, pp. 56–67, 2020.
- [18] [10] H. Hidayatullah, P. Purwanto, and Y. Umaidah, “Penerapan Naïve Bayes dengan optimasi Information Gain dan SMOTE untuk analisis sentimen pengguna aplikasi ChatGPT,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 3, pp. 1546–1553, 2023, doi:10.36040/jati.v7i3.6887.
- [19] A. M. Nur, W. Nursali, Nurhidayati, and I. Fathurrahman, “Penerapan metode Naïve Bayes untuk penentuan penerima beasiswa Program Indonesia Pintar,” *Infotek: Jurnal Informatika dan Teknologi*, vol. 7, no. 1, pp. 93–102, 2024.
- [20] J. Jamaludin, C. Rozikin, and A. S. Y. Irawan, “Klasifikasi jenis buah mangga dengan metode Backpropagation,” *Techné: Jurnal Ilmiah Elektroteknika*, vol. 20, no. 1, pp. 1–12, 2021, doi:10.31358/techne.v20i1.231.
- [21] Kementerian Pendidikan dan Kebudayaan, “Bantuan pendidikan PIP tahun 2023 mulai disalurkan.” [Online]. Available: <https://puslapdik.kemendikdasmen.go.id/bantuan-pendidikan-pip-tahun-2023-mulai-disalurkan/>
- [22] R. Nurhidayat and K. E. Dewi, “Penerapan algoritma K-Nearest Neighbor dan fitur ekstraksi N-Gram dalam analisis sentimen berbasis aspek,” *Komputa: Jurnal Ilmiah Komputer dan Informatika*, vol. 12, no. 1, pp. 91–100, 2023, doi:10.34010/komputa.v12i1.9458.
- [23] R. Oktafiani, A. Hermawan, and D. Avianto, “Pengaruh komposisi split data terhadap performa klasifikasi penyakit kanker payudara menggunakan algoritma machine learning,” *Jurnal Sains dan Informatika*, vol. 9, pp. 19–28, 2023, doi:10.34128/jsi.v9i1.622.
- [24] I. Priyanto *et al.*, “Penerapan algoritma metode Naïve Bayes untuk penentuan penerimaan bantuan Program Indonesia Pintar (PIP),” *Jurnal Manajemen Informatika Jayakarta*, vol. 4, no. 2, p. 162, 2024, doi:10.52362/jmijayakarta.v4i2.1355.
- [25] M. R. Qisthiano, T. B. Kurniawan, E. S. Negara, and M. Akbar, “Pengembangan model untuk prediksi tingkat kelulusan mahasiswa tepat waktu dengan metode Naïve Bayes,” *Jurnal Media Informatika Budidarma*, vol. 5, no. 3, p. 987, 2021, doi:10.30865/mib.v5i3.3030.
- [26] F. Reynaldy Valerian, M. Syarief, and D. Abdul Fatah, “Klasifikasi tingkat obesitas menggunakan metode GBM dan *confusion matrix*,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 2242–2249, 2025, doi:10.36040/jati.v9i2.13062.
- [27] Safitri, “Application of the Naïve Bayes algorithm for determining scholarship recipients at SD Negeri 6 Ketapang,” *Jurnal Informatika dan Sistem Informasi*, vol. 6, no. 1, pp. 43–52, 2022.
- [28] I. Safitri, A. Satria, and R. M. Badri, “Implementasi algoritma K-Nearest Neighbor dalam klasifikasi indeks pembangunan manusia Provinsi Sumatera Selatan tahun 2023,” *Digital Transformation Technology*, vol. 4, no. 2, pp. 768–775, 2024, doi:10.47709/digitech.v4i2.4614.
- [29] R. Yanuar and A. Andarsyah, “Sentimen analisis aplikasi Posaja pada Google Play Store menggunakan Support Vector Machine,” *Jurnal Teknik Informatika*, vol. 16, no. 2, pp. 1–7, 2024.
- [30] P. S. Zakaria, R. Julianto, and R. S. Bernada, “Implementasi Naive Bayes menggunakan Python dalam klasifikasi data,” *BIIKMA: Buletin Ilmiah Ilmu Komputer dan Multimedia*, vol. 1, no. 2, pp. 126–131, 2023.