VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Perbandingan Algoritma *K-Nearest Neighbor (K-NN)* dan *Naive Bayes* dalam Klasifikasi Tingkat Kemiskinan di Indonesia

Juanuari[1], Maulana Ilyas[2], Rahmat Tri Widodo[3], Ilham Manzis[4], Yusnia Budiarti[5], Musriatun Napiah[6]
[123456] Universitas Bina Sarana Informatika, Jakarta
15230248@bsi.ac.id[1], 15230371@bsi.ac.id[2], 15345678@bsi.ac.id[3],
15230849@bsi.ac.id[4], Yusnia.ybi@bsi.ac.id[5], musriatun.mph@bsi.ac.id[6]

ABSTRACT

Poverty is a major issue in sustainable development in Indonesia that requires a data-driven analysis approach to produce more accurate identification. This study aims to compare the performance of the K-Nearest Neighbor (K-NN) and Naive Bayes algorithms in classifying poverty levels in Indonesia based on social and economic data. The dataset was obtained from the Kaggle platform with the title "Classification of Poverty Levels in Indonesia", which contains 514 district/city data with various poverty indicators. The data was divided with a ratio of 80% for training and 20% for testing, then classification was carried out using the K-NN algorithm with a value of K = 5 and Naive Bayes. Evaluation was carried out using a confusion matrix with metrics of accuracy, precision, recall, and F1-score. The results showed that K-NN provided the best results with an accuracy of 97.09%, precision of 100%, recall of 75.00%, and F1-score of 85.71%, while Naive Bayes achieved an accuracy of 95.15%, precision of 73.33%, recall of 91.67%, and F1-score of 81.48%. This study resulted in better performance of this model compared to the results of previous studies. Therefore, the K-NN algorithm with the right parameters can be used as an effective method to support the data-based poverty level classification process and assist the government in poverty alleviation management and planning policies.

Keywords: *K-Nearest Neighbor, Naive Bayes, poverty classification, machine learning*

ABSTRAK

Kemiskinan merupakan isu utama dalam pembangunan berkelanjutan di Indonesia yang memerlukan pendekatan analisis berbasis data untuk menghasilkan identifikasi yang lebih akurat. Penelitian ini bertujuan untuk membandingkan kinerja algoritma *K-Nearest Neighbor* (K-NN) dan *Naive Bayes* dalam mengklasifikasikan tingkat kemiskinan di Indonesia berdasarkan data sosial dan ekonomi. Dataset diperoleh dari platform Kaggle dengan judul "Klasifikasi Tingkat Kemiskinan di Indonesia", yang berisi 514 data kabupaten/kota dengan berbagai indikator kemiskinan. Data dibagi dengan rasio 80% untuk pelatihan dan 20% untuk pengujian, kemudian dilakukan klasifikasi menggunakan algoritma K-NN dengan nilai K = 5 dan *Naive Bayes*. Evaluasi dilakukan menggunakan *confusion matrix* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa K-NN memberikan hasil terbaik dengan *accuracy* 97,09%, *precision* 100%, *recall* 75,00%, dan *F1-score* 85,71%, sedangkan *Naive Bayes* mencapai *accuracy* 95,15%, *precision* 73,33%, *recall* 91,67%, dan *F1-score* 81,48%. Penelitian ini menghasilkan performa model ini lebih baik dibandingkan hasil penelitian sebelumnya. Oleh karena itu, algoritma K-NN dengan parameter yang tepat dapat digunakan sebagai metode yang efektif untuk mendukung proses klasifikasi tingkat kemiskinan berbasis data serta membantu pemerintah dalam pengelolaan dan perencanaan kebijakan pengentasan kemiskinan.

Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Kata Kunci: K-Nearest Neighbor, Naive Bayes, klasifikasi kemiskinan, *machine learning*

PENDAHULUAN

Kemiskinan masih menjadi permasalahan mendasar yang dihadapi Indonesia dalam mewujudkan pembangunan berkelanjutan. Berdasarkan laporan Badan Pusat Statistik (BPS) tahun 2023, persentase penduduk miskin tercatat sebesar 9,36% dari total populasi, dengan kesenjangan yang cukup mencolok antara wilayah perkotaan dan pedesaan (Anna, 2023). Kondisi ini menunjukkan bahwa kemiskinan tidak hanya dipengaruhi oleh tingkat pendapatan yang rendah, tetapi juga oleh keterbatasan akses terhadap pendidikan, layanan kesehatan, serta kesejahteraan sosial (Agus Triono & Sangaji, 2023).

Perkembangan teknologi kecerdasan buatan, khususnya di bidang machine learning, membuka peluang besar untuk mengolah data sosial-ekonomi guna menganalisis dan mengklasifikasikan tingkat kemiskinan berbasis data. Teknik pembelajaran mesin memungkinkan sistem mengenali pola kompleks dalam data dan menghasilkan prediksi yang akurat (Fauziah et al., 2022). Di antara berbagai algoritma yang sering digunakan dalam klasifikasi, *K-Nearest Neighbor* (K-NN) dan *Naive Bayes* merupakan dua metode populer. K-NN bekerja dengan mengukur jarak kemiripan antar data untuk menentukan kelas yang sesuai dengan kondisi aktual (Fitra & Rusdi, 2022), sedangkan *Naive Bayes* menggunakan pendekatan probabilistik yang sederhana namun efisien dalam menentukan label data (Rivaldo, Vito Junivan & Pranoto, 2024).

Sejumlah penelitian terdahulu telah menunjukkan efektivitas kedua metode tersebut dalam menganalisis fenomena sosial-ekonomi. Fauziah et al. (2022) melaporkan bahwa dalam klasifikasi data kemiskinan di Papua, metode K-NN memperoleh akurasi sebesar 58,62%, sedangkan SVM mencapai 93,1%, yang menandakan bahwa performa K-NN sangat dipengaruhi oleh karakteristik data dan parameter yang digunakan (Fauziah et al., 2022). Sementara itu, penelitian oleh Mardiah et al. (2024) membuktikan bahwa K-NN efektif dalam mengklasifikasikan data ekonomi masyarakat di wilayah Nagari, Sumatera Barat, serta bermanfaat dalam mendukung kebijakan bantuan sosial di tingkat daerah (Mardiah et al., 2024).

Penelitian yang dilakukan oleh Djafar & Fauzan. (2024) juga menggunakan algoritma K-NN dengan teknik oversampling pada data campuran untuk mengklasifikasikan status kesejahteraan rumah tangga di Kabupaten Kulon Progo, Hasil penelitian tersebut menunjukkan peningkatan akurasi dan sensitivitas model hingga mendekati 79% (Djafar & Fauzan, 2024). Selanjutnya, Amalia et al. (2025) membandingkan algoritma K-NN dan C4.5 dalam menentukan kelayakan penerima bantuan langsung tunai (BLT) dan menemukan bahwa K-NN memberikan hasil yang stabil, terutama pada dataset berukuran kecil dengan variabel ekonomi dominan (Amalia et al., 2025). Studi lain oleh Duwo Jiwo Saputro et al. (2024) berfokus pada klasifikasi tingkat kemiskinan di Provinsi Jawa Barat dengan menerapkan algoritma K-NN, hasilnya menunjukkan bahwa pemilihan nilai k yang optimal mampu menghasilkan akurasi lebih dari 96% dengan nilai recall mencapai 100%, yang berarti seluruh data kategori miskin berhasil diidentifikasi dengan benar (Duwo Jiwo

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Saputro et al., 2024). Temuan ini mengindikasikan bahwa K-NN dapat memberikan hasil yang unggul apabila parameter dan proses *preprocessing* diterapkan secara tepat.

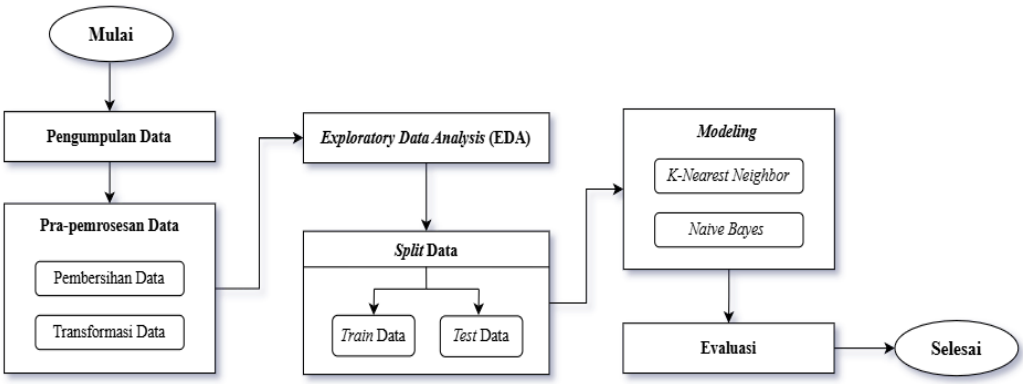
Sementara itu, Suci Mulyani et al. (2024) menerapkan algoritma *Naive Bayes* dalam klasifikasi tingkat kemiskinan di indonesia dan memperoleh akurasi sebesar 96,12% (Suci Mulyani et al., 2024). Hal ini menunjukkan bahwa metode probabilistik juga memiliki kinerja yang cukup baik, meskipun masih terdapat peluang peningkatan performa melalui pendekatan algoritma lain.

Berdasarkan berbagai hasil penelitian tersebut, penelitian ini bertujuan untuk melakukan analisis perbandingan antara algoritma *K-Nearest Neighbor* (K-NN) dan *Naive Bayes* dalam klasifikasi tingkat kemiskinan di Indonesia. Data yang digunakan mencakup indikator sosial-ekonomi dari seluruh kabupaten dan kota, meliputi tingkat pengangguran terbuka, indeks pembangunan manusia (IPM), pengeluaran per kapita, serta akses terhadap fasilitas dasar. Evaluasi kinerja dilakukan dengan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menentukan algoritma yang memiliki performa paling optimal.

METODE PENELITIAN

1. Alur Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang sistematis dan terstruktur untuk memastikan proses analisis berjalan secara ilmiah dan dapat dilakukan kembali dengan prosedur yang sama. Alur keseluruhan penelitian dapat dilihat pada Gambar 1, yang menunjukkan langkah-langkah mulai dari pengumpulan dataset hingga pembangunan model klasifikasi.



Gambar 1 Tahapan Penelitian
Sumber: Penulis, 20 Oktober 2025

Secara umum, tahapan penelitian ini terdiri atas enam langkah utama, yaitu:

1. Pengumpulan data,
2. Pra-pemrosesan data,
3. Analisis eksploratif (EDA),
4. Pembagian data (*Split* data),
5. Pembangunan model (*modeling*), dan

Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

6. Evaluasi.

Setiap tahap memiliki peran penting dalam menghasilkan model klasifikasi yang andal untuk mengidentifikasi tingkat kemiskinan di Indonesia secara akurat.

Seluruh proses analisis, pra-pemrosesan, pemodelan, dan evaluasi dilakukan menggunakan bahasa pemrograman Python pada platform Google Colab, yang memudahkan pengolahan data secara efisien dan memungkinkan eksekusi kode secara online tanpa konfigurasi lokal yang kompleks.

2. Pengumpulan Data

Pada penelitian ini, data yang digunakan berasal dari dataset berjudul "Klasifikasi Tingkat Kemiskinan di Indonesia" yang tersedia di platform Kaggle, yaitu sebuah situs penyedia dan berbagi dataset publik yang banyak dimanfaatkan dalam pengembangan serta pengujian model machine learning. Dataset tersebut dapat diakses melalui tautan https://www.kaggle.com/datasets/ermila/ klasifikasi-tingkat-kemiskinan-di-indonesia.

Dataset ini terdiri dari 514 data kabupaten/kota di seluruh Indonesia, yang mencakup berbagai indikator sosial dan ekonomi sebagai variabel prediktor tingkat kemiskinan. Beberapa indikator utama di dalamnya meliputi persentase penduduk miskin, rata-rata lama sekolah, pengeluaran per kapita, indeks pembangunan manusia (IPM), usia harapan hidup, persentase akses sanitasi layak, persentase akses air minum layak, tingkat pengangguran terbuka, tingkat partisipasi angkatan kerja, serta nilai produk domestik regional bruto (PDRB) atas dasar harga konstan. Selain variabel-variabel tersebut, dataset ini juga memiliki satu atribut label yang berfungsi sebagai variabel target, yaitu klasifikasi tingkat kemiskinan, yang membagi setiap kabupaten atau kota ke dalam dua kategori, yakni "Miskin" dan "Tidak Miskin".

3. Pra-Pemrosesan Data (*Data Preprocessing*)

Sebelum data digunakan untuk membangun model klasifikasi, dilakukan tahap pra-pemrosesan data guna memastikan bahwa data berada dalam kondisi bersih, lengkap, dan konsisten. Tahapan ini sangat penting karena kualitas hasil model sangat dipengaruhi oleh kualitas data yang digunakan.

Tahapan pra-pemrosesan meliputi pembersihan data (*data cleaning*) dan transformasi data (*data transformation*). Pada tahap pembersihan data, dilakukan penghapusan baris yang memiliki nilai kosong pada kolom target, penggantian nilai hilang pada kolom numerik menggunakan nilai rata-rata (*mean*), serta penyeragaman format teks pada kolom "Provinsi" agar konsisten. Langkah ini dilakukan untuk menghindari kesalahan analisis dan mencegah bias model akibat data yang tidak lengkap.

Selanjutnya dilakukan tahap transformasi data, di mana variabel target "Klasifikasi Kemiskinan" diubah menjadi bentuk kategorikal dengan memberikan kode numerik 0 untuk kategori "Tidak Miskin" dan 1 untuk kategori "Miskin". Transformasi ini bertujuan agar data dapat diproses oleh algoritma pembelajaran mesin yang membutuhkan representasi numerik.

Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

4. *Exploratory Data Analysis* (EDA)

Tahapan berikutnya adalah *Exploratory Data Analysis* (EDA), yaitu proses eksplorasi awal terhadap data untuk memahami karakteristik, distribusi, dan struktur data sebelum dilakukan pemodelan. Tahapan ini bertujuan untuk memberikan gambaran menyeluruh mengenai data yang digunakan, termasuk sebaran nilai, pola umum, serta potensi ketidakseimbangan antar kelas yang dapat memengaruhi kinerja model pembelajaran mesin.

Proses EDA pada penelitian ini dilakukan menggunakan pustaka Matplotlib dan Seaborn dalam bahasa pemrograman Python, yang mendukung pembuatan visualisasi data secara informatif dan interaktif. Melalui tahap ini, peneliti dapat memeriksa komposisi data berdasarkan wilayah administratif, seperti distribusi jumlah kabupaten/kota per provinsi, serta membandingkan proporsi antara kategori miskin dan tidak miskin untuk memastikan keseimbangan data yang digunakan dalam pemodelan. Selain itu, EDA juga membantu dalam mengenali pola umum yang muncul pada variabel-variabel sosial dan ekonomi yang menjadi indikator kemiskinan.

Tahap ini sangat penting karena membantu peneliti mengidentifikasi potensi permasalahan pada data, seperti adanya *missing values*, *outlier*, atau kesalahan input yang mungkin belum terdeteksi pada tahap pra-pemrosesan. Dengan demikian, hasil EDA dapat digunakan sebagai dasar untuk menentukan langkah lanjutan dalam penyesuaian data sebelum masuk ke proses pelatihan model.

Secara keseluruhan, *Exploratory Data Analysis* berfungsi untuk memastikan bahwa data yang akan digunakan benar-benar representatif, konsisten, dan sesuai dengan tujuan penelitian. Melalui pemahaman mendalam terhadap struktur dan karakteristik data pada tahap ini, proses pemodelan yang dilakukan di tahap berikutnya dapat berlangsung dengan lebih efektif dan menghasilkan prediksi yang lebih akurat.

5. Pembagian Data (*Split* Data)

Setelah tahap EDA selesai, dataset dibagi menjadi dua bagian, yaitu 80% untuk data latih (*training data*) dan 20% untuk data uji (*testing data*). Pembagian ini dilakukan menggunakan fungsi ‘train_test_split’ dari pustaka Scikit-Learn.

Tujuan dari pembagian ini adalah agar model dapat dilatih pada sebagian besar data dan diuji menggunakan data yang belum pernah digunakan sebelumnya, sehingga evaluasi kinerja model dapat dilakukan secara objektif dan tidak bias.

6. Pembangunan Model (*Modeling*)

Setelah data melalui tahap pembagian menjadi data latih dan data uji, langkah berikutnya adalah pembangunan model klasifikasi (modeling). Tahap ini dilakukan untuk membandingkan dua algoritma pembelajaran mesin, yaitu *K-Nearest Neighbor* (K-NN) dan *Naive Bayes*, dalam mengklasifikasikan tingkat kemiskinan kabupaten/kota di Indonesia.

Model *K-Nearest Neighbor* (K-NN) bekerja dengan cara mengukur jarak antara data uji dan data latih menggunakan metrik jarak, seperti *Euclidean distance*.

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Setiap data uji akan diklasifikasikan berdasarkan mayoritas kelas dari K tetangga terdekatnya. Dalam penelitian ini, nilai K = 5 digunakan karena memberikan keseimbangan optimal antara kompleksitas model dan tingkat akurasi. Semakin besar nilai K, semakin halus keputusan klasifikasi, namun dapat mengurangi sensitivitas terhadap pola minoritas.

Sementara itu, Model *Naive Bayes* digunakan sebagai pembanding karena sifatnya yang cepat, sederhana, dan efisien dalam menangani data kategorikal. Algoritma ini bekerja dengan menghitung probabilitas setiap kelas berdasarkan distribusi fitur, menggunakan asumsi independensi antar variabel. Meskipun sederhana, *Naive Bayes* sering kali memberikan hasil yang kompetitif pada dataset sosial-ekonomi dengan jumlah variabel terbatas.

7. Evaluasi

Tahap terakhir adalah evaluasi model, yang dilakukan menggunakan confusion *matrix* serta beberapa metrik kinerja seperti *accuracy*, *precision*, *recall*, dan *F1-score*. *Confusion Matrix* merupakan sebuah matrik dua dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kelas data sebenarnya, kemudian menghitung jumlah prediksi yang benar dan yang salah untuk setiap kategori kelas (Alfitriana Riska et al., 2023). *Confusion Matrix* dapat dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix*

	Prediksi Negatif	**Prediksi Positif**
Actual Negatif	TN	FP
Actual Positif	FN	TP

Keterangan:
True Positive (TP) adalah nilai prediksi benar dan nilai sebenarnya benar.
True Negative (TN) adalah nilai prediksi salah dan nilai sebenarnya salah.
False Positive (FP) adalah nilai prediksi benar dan nilai sebenarnya salah.
False Negative (FN) adalah nilai prediksi salah dan nilai sebenarnya benar.
Model prediksi yang dibangun akan diuji dengan perhitungan *Accuracy, Recall, Precission,* dan *F1-score* (Adane et al., 2023). Berikut adalah persamaan yang digunakan:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\% \quad (1)$$

$$Precission = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

$$F1\text{-}Score = 2 \times \frac{(persisi \times recall)}{(persisi + recall)} \times 100\% \quad (4)$$

Evaluasi ini bertujuan untuk menilai sejauh mana model mampu melakukan klasifikasi secara akurat dan konsisten. Apabila hasil pengujian menunjukkan performa yang baik, maka penelitian dianggap selesai dan siap untuk diimplementasikan atau dipublikasikan.

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Pada penelitian ini, data yang digunakan berasal dari dataset "Klasifikasi Tingkat Kemiskinan di Indonesia" yang tersedia di platform Kaggle, sebuah situs penyedia data yang sering digunakan untuk pengembangan dan pengujian model machine learning. Dataset tersebut dapat diakses melalui tautan https://www.kaggle.com/datasets/ermila/klasifikasi-tingkat-kemiskinan-di-indonesia.

Dataset ini terdiri dari 514 data kabupaten/kota di Indonesia dan mencakup berbagai indikator sosial serta ekonomi, antara lain persentase penduduk miskin, rata-rata lama sekolah, pengeluaran per kapita, indeks pembangunan manusia (IPM), usia harapan hidup, persentase akses sanitasi layak, persentase akses air minum layak, tingkat pengangguran terbuka, tingkat partisipasi angkatan kerja, serta nilai PDRB atas harga konstan dari pengeluaran. Selain atribut-atribut tersebut, dataset ini juga memiliki satu atribut label yang digunakan sebagai variabel target, yaitu klasifikasi tingkat kemiskinan.

	Provinsi	Kab/Kota	Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)	Rata-rata Lama Sekolah Penduduk 15+ (Tahun)	Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)	Indeks Pembangunan Manusia	Umur Harapan Hidup (Tahun)	Persentase rumah tangga yang memiliki akses terhadap sanitasi layak	Persentase rumah tangga yang memiliki akses terhadap air minum layak	Tingkat Pengangguran Terbuka	Tingkat Partisipasi Angkatan Kerja	PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)	Klasifikasi Kemiskinan
0	ACEH	Simeulue	18,98	9,48	7148.0	66,41	65,28	71,56	87,45	5,71	71,15	1648096.0	0.0
1	ACEH	Aceh Singkil	20,36	8,68	8776.0	69,22	67,43	69,56	78,58	8,36	62,85	1780419.0	1.0
2	ACEH	Aceh Selatan	13,18	8,88	8180.0	67,44	64,4	62,55	79,65	6,46	60,85	4345784.0	0.0
3	ACEH	Aceh Tenggara	13,41	9,67	8030.0	69,44	68,22	62,71	86,71	6,43	69,62	3487157.0	0.0
4	ACEH	Aceh Timur	14,45	8,21	8577.0	67,83	68,74	66,75	83,16	7,13	59,48	8433526.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
509	PAPUA	Puncak	36,26	2,16	5412.0	43,17	65,86	11,43	85,03	0,94	89,43	831070.0	1.0
510	PAPUA	Dogiyai	28,81	4,94	5415.0	55	65,85	12,11	71,24	5,68	78,20	906904.0	1.0
511	PAPUA	Intan Jaya	41,66	3,09	5328.0	48,34	65,69	0,36	35,01	1,43	75,75	767101.0	1.0
512	PAPUA	Deiyai	40,59	3,25	4673.0	49,96	65,36	0,00	85,23	0,79	85,01	841296.0	1.0
513	PAPUA	Kota Jayapura	11,39	11,57	14937.0	80,11	70,52	85,31	97,10	11,67	63,75	22852202.0	0.0

Gambar 2 Tampilan Dataset Tingkat Kemiskinan di Wilayah Indonesia
Sumber: Penulis, 20 Oktober 2025

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

2. Pra-Pemrosesan Data (*Data Preprocessing*)

Tahap pra-pemrosesan data dilakukan untuk memastikan data dalam kondisi bersih, konsisten, dan siap digunakan pada proses analisis. Tahapan ini terdiri dari dua proses utama, yaitu pembersihan data (*data cleaning*) dan transformasi data (*data transformation*). Pada tahap pembersihan data (*data cleaning*), dilakukan beberapa langkah untuk mengatasi data yang tidak lengkap atau tidak sesuai, yaitu menghapus baris dengan nilai *null* pada kolom Klasifikasi Kemiskinan, mengisi nilai kosong pada kolom numerik menggunakan rata-rata, serta mengisi nilai kosong pada kolom kategorikal dengan nilai yang paling sering muncul. Langkah-langkah tersebut dilakukan untuk memastikan tidak ada data yang hilang atau salah format yang dapat memengaruhi hasil analisis.

```
# 1. Drop baris yang targetnya null (Klasifikasi Kemiskinan)
df = df.dropna(subset=["Klasifikasi Kemiskinan"])

# 2. Mengisi nilai kosong di kolom numerik dengan rata-rata
num_cols = df.select_dtypes(include=[np.number]).columns
imputer_mean = SimpleImputer(strategy="mean")
df[num_cols] = imputer_mean.fit_transform(df[num_cols])

# 3. Mengisi nilai kosong di kolom kategorikal dengan nilai paling sering muncul
cat_cols = df.select_dtypes(include=["object"]).columns
imputer_mode = SimpleImputer(strategy="most_frequent")
df[cat_cols] = imputer_mode.fit_transform(df[cat_cols])
```

Gambar 3 Pembersihan data
Sumber: Penulis, 20 Oktober 2025

```
================================================================================
                Missing values setelah data di bersihkan
================================================================================
Provinsi                                                                    0
Kab/Kota                                                                    0
Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)             0
Rata-rata Lama Sekolah Penduduk 15+ (Tahun)                                 0
Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)                0
Indeks Pembangunan Manusia                                                  0
Umur Harapan Hidup (Tahun)                                                  0
Persentase rumah tangga yang memiliki akses terhadap sanitasi layak         0
Persentase rumah tangga yang memiliki akses terhadap air minum layak        0
Tingkat Pengangguran Terbuka                                                0
Tingkat Partisipasi Angkatan Kerja                                          0
PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)                  0
Klasifikasi Kemiskinan                                                      0
```

Gambar 4 Hasil pembersihan data
Sumber: Penulis, 20 Oktober 2025

Selanjutnya, dilakukan transformasi data dengan mengubah nilai numerik menjadi nominal pada kolom Klasifikasi Kemiskinan dengan ketentuan 0 sebagai “Tidak Miskin” dan 1 sebagai “Miskin”. Proses ini membantu membuat data lebih mudah dipahami.

```
# 4. Mengubah numerik menjadi nominal pada kolom Label/Target
df["Klasifikasi Kemiskinan"] = df["Klasifikasi Kemiskinan"].map({
    0.0: "Tidak Miskin",
    1.0: "Miskin"
})
```

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

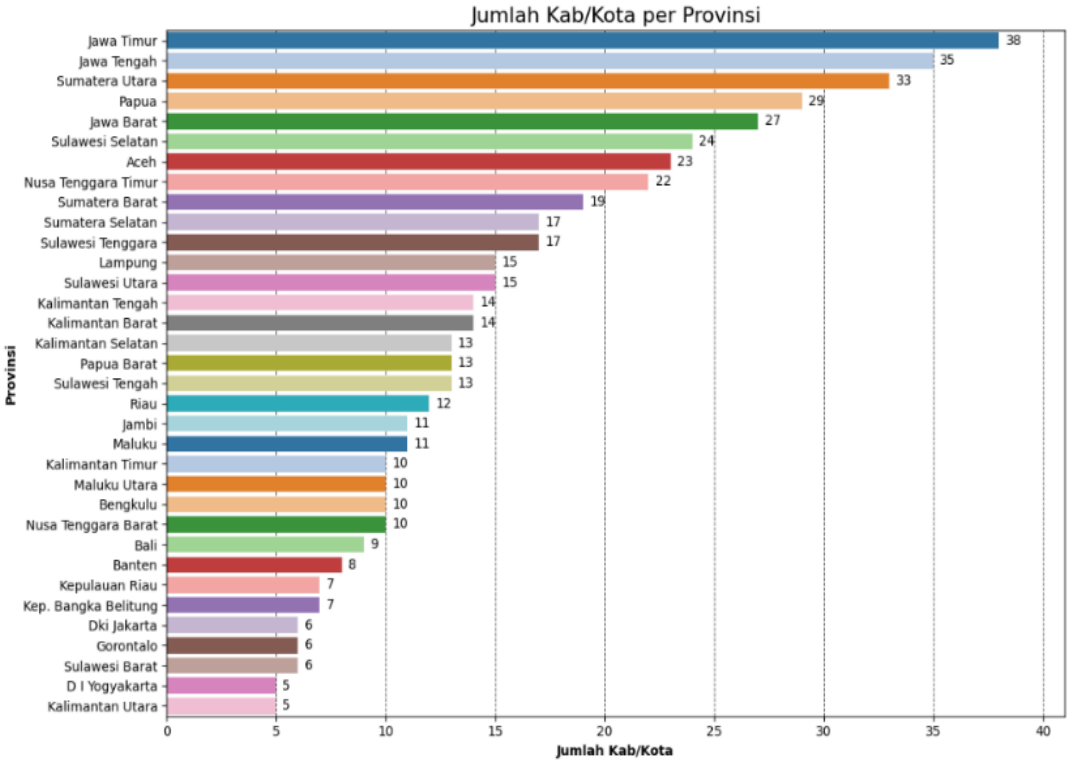
Gambar 5 Transformasi nilai numerik menjadi nominal pada kolom Target
Sumber: Penulis, 20 Oktober 2025

3. *Exploratory Data Analysis* (EDA)

Tahap ini bertujuan untuk memahami karakteristik data serta menemukan pola atau hubungan yang dapat membantu proses klasifikasi. Melalui *Exploratory Data Analysis* (EDA), dilakukan analisis visual dan deskriptif untuk memperoleh wawasan awal terhadap persebaran tingkat kemiskinan di Indonesia.

Dalam penelitian ini, terdapat 3 (tiga) jenis analisis yang dilakukan pada tahap EDA, yaitu:

1. EDA Jumlah Kabupaten/Kota per Provinsi, yang digunakan untuk melihat distribusi data berdasarkan wilayah administrasi di seluruh Indonesia. Analisis ini memberikan gambaran mengenai persebaran jumlah kabupaten/kota di setiap provinsi.

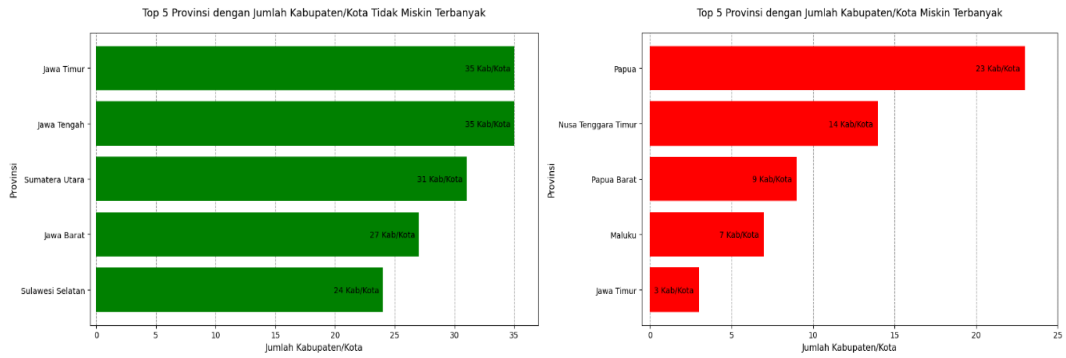


Gambar 6 Hasil EDA Jumlah Kabupaten/Kota per Provinsi
Sumber: Penulis, 20 Oktober 2025

Hasil EDA menunjukkan bahwa jumlah kabupaten/kota di Indonesia bervariasi di setiap provinsi. Provinsi seperti Jawa Timur, Jawa Tengah, dan Sumatera Utara memiliki jumlah kabupaten/kota terbanyak, sedangkan DI Yogyakarta dan Kalimantan Utara memiliki jumlah paling sedikit. Variasi ini menggambarkan perbedaan luas wilayah dan kompleksitas administratif di tiap daerah. Semakin besar jumlah kabupaten/kota, semakin beragam pula karakteristik sosial dan ekonomi yang dimiliki, yang dapat memengaruhi analisis tingkat kemiskinan di wilayah tersebut.

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

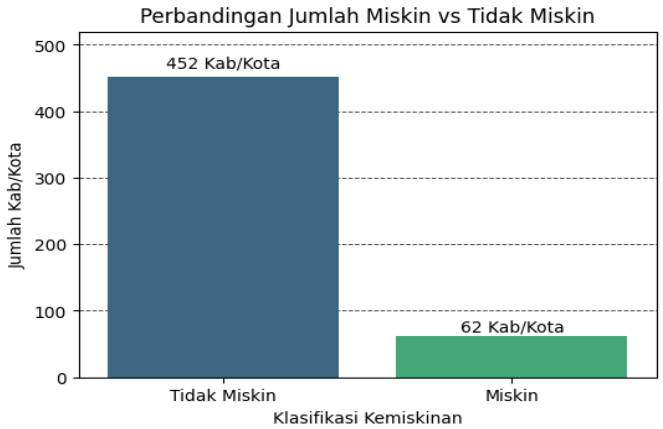
2. EDA Top 5 Provinsi dengan Jumlah Kabupaten/Kota Miskin dan Tidak Miskin Terbanyak, yang bertujuan untuk mengidentifikasi provinsi dengan jumlah wilayah miskin maupun tidak miskin paling banyak, sehingga dapat diketahui daerah dengan ketimpangan ekonomi terbesar.



Gambar 7 Hasil EDA Top 5 Provinsi dengan Jumlah Kabupaten/Kota Miskin dan Tidak Miskin Terbanyak
Sumber: Penulis, 20 Oktober 2025

Hasil EDA menunjukkan bahwa provinsi seperti Jawa Timur dan Jawa Tengah memiliki jumlah kabupaten/kota dengan kategori Tidak Miskin paling banyak, sedangkan provinsi seperti Papua dan Nusa Tenggara Timur mendominasi kategori Miskin. Temuan ini menggambarkan adanya ketimpangan sosial-ekonomi antar wilayah di Indonesia, di mana daerah bagian barat umumnya memiliki tingkat kesejahteraan yang lebih baik dibandingkan bagian timur. Analisis ini memberikan gambaran penting mengenai konsentrasi wilayah miskin yang dapat menjadi fokus utama dalam perumusan kebijakan pemerataan pembangunan.

3. EDA Perbandingan Jumlah Miskin vs Tidak Miskin, yang dilakukan untuk melihat proporsi keseluruhan antara kelompok masyarakat miskin dan tidak miskin dalam dataset. Analisis ini membantu memahami keseimbangan data yang akan memengaruhi performa model klasifikasi.



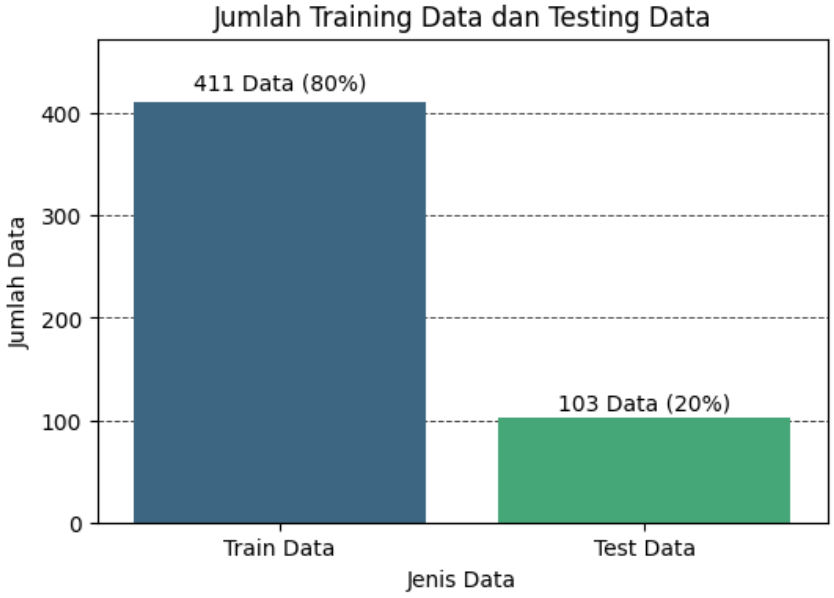
Gambar 8 Hasil EDA Perbandingan Jumlah Miskin vs Tidak Miskin
Sumber: Penulis, 20 Oktober 2025

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Hasil visualisasi pada EDA ini menunjukkan bahwa jumlah kabupaten/kota dengan kategori "Tidak Miskin" lebih banyak dibandingkan dengan kategori "Miskin". Kondisi ini menggambarkan bahwa secara umum, sebagian besar wilayah di Indonesia berada pada tingkat kesejahteraan yang relatif baik. Namun, perbedaan jumlah tersebut juga menandakan adanya ketidakseimbangan distribusi ekonomi di beberapa daerah, terutama di provinsi-provinsi dengan tingkat kemiskinan yang masih tinggi. Temuan ini penting sebagai dasar dalam memahami sebaran kondisi sosial-ekonomi masyarakat Indonesia serta menjadi acuan awal dalam analisis dan pemodelan klasifikasi tingkat kemiskinan.

4. *Split* Data (Pembagian Data)

Selanjutnya tahap *Split* Data pada penelitian ini, data dibagi menggunakan rasio 80% sebagai data latih dan 20% sebagai data uji. Rasio ini dipilih karena dinilai seimbang antara jumlah data untuk pembelajaran dan data untuk pengujian model. Hasil pembagian data dapat dilihat pada gambar 9 berikut.



Gambar 9 *Split* Data
Sumber: Penulis, 20 Oktober 2025

Berdasarkan Gambar 9, dari total 514 data yang digunakan, sebanyak 411 data (80%) digunakan untuk melatih model dan 103 data (20%) digunakan untuk pengujian. Pembagian ini bertujuan agar kedua algoritma memiliki kesempatan yang sama dalam mempelajari karakteristik data dan dapat dibandingkan secara objektif berdasarkan hasil evaluasi performa klasifikasinya.

5. *Modeling*

Selanjutnya tahapan *modeling* dilakukan untuk membangun model klasifikasi tingkat kemiskinan kabupaten/kota di Indonesia menggunakan dua algoritma, yaitu *K-Nearest Neighbor* (K-NN) dan *Naive Bayes*. Proses ini dilaksanakan menggunakan bahasa pemrograman *Python* pada platform Google Colab dengan memanfaatkan pustaka *Scikit-Learn*.

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Pada tahap pertama, dilakukan pemodelan menggunakan algoritma *K-Nearest Neighbor*. Algoritma ini bekerja dengan menentukan sejumlah tetangga terdekat dari data uji berdasarkan jarak terpendek terhadap data latih. Nilai K = 5 digunakan karena memberikan hasil *accuracy* terbaik dibandingkan dengan nilai K lainnya.

```
# Model KNN
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
```

Gambar 10 Pemodelan *K-Nearest Neighbor*
Sumber: Penulis, 20 Oktober 2025

Kemudian dilakukan pemodelan menggunakan algoritma *Naive Bayes* sebagai pembanding terhadap metode K-NN. Algoritma ini menggunakan pendekatan probabilistik untuk menentukan kelas berdasarkan distribusi data pada setiap fitur.

```
# Model Naive Bayes
nb = GaussianNB()
nb.fit(X_train, y_train)
y_pred_nb = nb.predict(X_test)
```

Gambar 11 Pemodelan *Naive Bayes*
Sumber: Penulis, 20 Oktober 2025

6. Evaluasi

Setelah menyelesaikan tahapan pemodelan dengan algoritma *K-Nearest Neighbor* (K-NN) dan *Naive Bayes*, dilakukan proses evaluasi untuk mengukur kinerja model klasifikasi tingkat kemiskinan kabupaten/kota di Indonesia. Evaluasi menggunakan *confusion matrix* untuk mengetahui jumlah data yang diklasifikasikan dengan benar maupun salah oleh masing-masing model. Penilaian akurasi didasarkan pada hasil *confusion matrix*. Proses pemanggilan *confusion matrix* dilakukan dengan mengimpor fungsi 'confusion_matrix' dari 'sklearn metrics', kemudian fungsi tersebut dipanggil menggunakan parameter 'y_test' dan hasil prediksi model (y_pred). Parameter 'y_test' merepresentasikan label sebenarnya dari data uji, sedangkan 'y_pred' adalah label hasil prediksi model. Hasil dari fungsi ini disimpan dalam variabel cm, yang kemudian divisualisasikan menggunakan *heatmap* untuk memudahkan interpretasi.

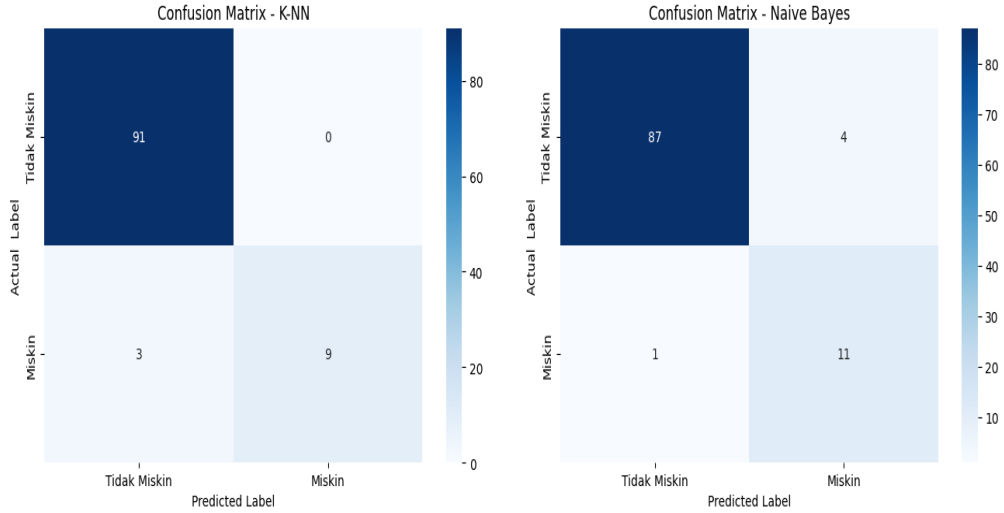
```
from sklearn.metrics import confusion_matrix
# === KNN ===
cm_knn = confusion_matrix(y_test, y_pred_knn)
sns.heatmap(cm_knn, annot=True, fmt="d", cmap="Blues", ax=axes[0,0],
            xticklabels=labels, yticklabels=labels)
axes[0,0].set_title("Confusion Matrix - K-NN")
axes[0,0].set_xlabel("Predicted Label")
axes[0,0].set_ylabel("Actual  Label")

# === Naive Bayes ===
cm_nb = confusion_matrix(y_test, y_pred_nb)
sns.heatmap(cm_nb, annot=True, fmt="d", cmap="Blues", ax=axes[0,1],
            xticklabels=labels, yticklabels=labels)
axes[0,1].set_title("Confusion Matrix - Naive Bayes")
axes[0,1].set_xlabel("Predicted Label")
axes[0,1].set_ylabel("Actual  Label")
```

Gambar 12 Pemanggilan Confusion Matrix
Sumber: Penulis, 20 Oktober 2025

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

Hasil *confusion matrix* dapat dilihat pada gambar 13 berikut.



Gambar 13 *Confusion Matrix* gabungan algoritma *K-Nearest Neighbor* (K-NN) dan *Naive Bayes*

Sumber: Penulis, 20 Oktober 2025

Berdasarkan Gambar 13 untuk model *K-Nearest Neighbor* dengan parameter K = 5, terdapat 9 kabupaten/kota yang diprediksi miskin dan memang benar miskin (*True Positive*/TP), serta 91 kabupaten/kota diprediksi tidak miskin dan memang tidak miskin (*True Negative*/TN). Namun, terdapat 3 kabupaten/kota yang diprediksi tidak miskin tetapi sebenarnya miskin (*False Negative*/FN), dan 0 atau tidak ditemukan kabupaten/kota yang diprediksi miskin padahal sebenarnya tidak miskin (*False Positive*/FP).

Sementara itu, Gambar 13 untuk model *Naive Bayes* menunjukkan 11 kabupaten/kota yang diprediksi miskin dan benar-benar miskin (*True Positive*/TP), serta 87 kabupaten/kota diprediksi tidak miskin dan memang tidak miskin *(True Negative*/TN). Namun, terdapat 1 kabupaten/kota diprediksi tidak miskin tetapi sebenarnya miskin (*False Negative*/FN), dan 4 kabupaten/kota diprediksi miskin padahal sebenarnya tidak miskin (*False Positive*/FP).

Kemudian, berdasarkan hasil *confusion matrix* pada Gambar 13, dilakukan perhitungan untuk algoritma K-NN guna memperoleh matriks evaluasi model. Nilai-nilai seperti *accuracy, precision, recall,* dan *F1-score* dihitung berdasarkan rumus evaluasi klasifikasi dengan dasar jumlah TP, TN, FP, dan FN. Berikut penghitungan masing-masing metrik tersebut:

1. *Accuracy*

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\%$$

$$Accuracy = \frac{(9+91)}{(9+91+0+3)} \times 100\%$$

$$Accuracy = \frac{100}{103} \times 100\%$$

$$Accuracy = 97{,}09\%$$

Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

2. *Precission*

$$Precission = \frac{TP}{TP+FP} \times 100\%$$

$$Precission = \frac{9}{9+0} \times 100\%$$

$$Precission = \frac{9}{9} \times 100\%$$

$$Precission = 100\%$$

3. *Recall*

$$Recall = \frac{TP}{TP+FN} \times 100\%$$

$$Recall = \frac{9}{9+3} \times 100\%$$

$$Recall = \frac{9}{12} \times 100\%$$

$$Recall = 75\%$$

4. *F1-Score*

$$F1\text{-}Score = 2 \times \frac{(persisi \times recall)}{(persisi + recall)} \times 100\%$$

$$F1\text{-}Score = 2 \times \frac{(100 \times 75)}{(100+75)} \times 100\%$$

$$F1\text{-}Score = 85{,}71\%$$

Sehingga di dapatkan hasil evaluasi algoritma K-NN dengan *accuracy* 97,09%, *precision* 100%, *recall* 75%, dan *F1-score* 85,71%.

Selanjutnya, berdasarkan hasil *confusion matrix* pada Gambar 13, dilakukan perhitungan untuk algoritma *Naive Bayes* guna memperoleh matriks evaluasi model. Nilai-nilai seperti *accuracy, precision, recall,* dan *F1-score* dihitung berdasarkan rumus evaluasi klasifikasi dengan dasar jumlah TP, TN, FP, dan FN. Berikut penghitungan masing-masing metrik tersebut:

1. *Accuracy*

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\%$$

$$Accuracy = \frac{(11+87)}{(11+87+4+1)} \times 100\%$$

$$Accuracy = \frac{98}{103} \times 100\%$$

$$Accuracy = 95{,}15\%$$

2. *Precission*

$$Precission = \frac{TP}{TP+FP} \times 100\%$$

$$Precission = \frac{11}{11+4} \times 100\%$$

$$Precission = \frac{11}{15} \times 100\%$$

$$Precission = 73{,}33\%$$

3. *Recall*

$$Recall = \frac{TP}{TP+FN} \times 100\%$$

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

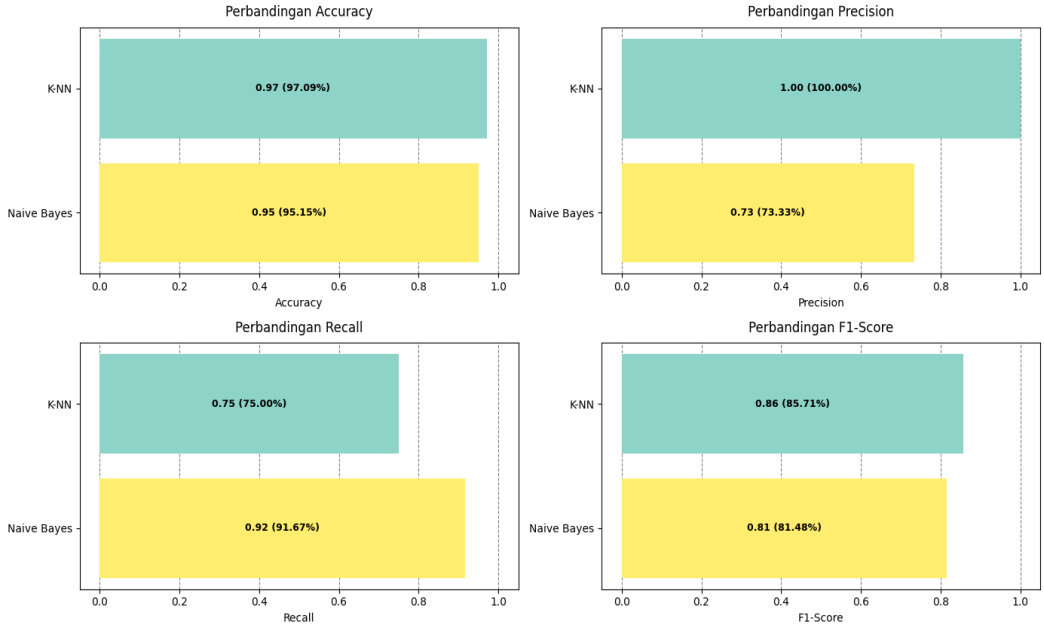
$$Recall = \frac{11}{11+1} \times 100\%$$
$$Recall = \frac{11}{12} \times 100\%$$
$$Recall = 91{,}67\%$$

4. *F1-Score*

$$F1\text{-}Score = 2 \times \frac{(persisi \times recall)}{(persisi + recall)} \times 100\%$$
$$F1\text{-}Score = 2 \times \frac{(73{,}33 \times 91{,}67)}{(73{,}33 + 91{,}67)} \times 100\%$$
$$F1\text{-}Score = 81{,}48\%$$

Sehingga di dapatkan hasil evaluasi algoritma *Naive Bayes* dengan *accuracy* 95,15%, *precision* 73,33%, *recall* 91,67%, dan *F1-score* 81,48%.

Selanjutnya, pada Gambar 14 berikut ini merupakan diagram batang perbandingan hasil evaluasi dari algoritma *K-Nearest Neighbor* (K-NN) dan *Naive Bayes* yang mencakup *accuracy*, *precision*, *recall*, dan *F1-score*.



Gambar 14 Perbandingan *Accuracy, Precision, Recall,* dan *F1-Score*
Sumber: Penulis, 20 Oktober 2025

Berdasarkan Gambar 14, hasil perbandingan menunjukkan bahwa *K-Nearest Neighbor* memberikan kinerja lebih baik daripada *Naive Bayes*. Pada penelitian ini, K-NN mencapai *accuracy* 97,09%, *precision* 100%, *recall* 75,00%, dan *F1-score* 85,71%, lebih baik dibandingkan *Naive Bayes* yang memperoleh *accuracy* 95,15%, *precision* 73,33%, *recall* 91,67%, dan *F1-score* 81,48%. Hasil ini menunjukkan bahwa algoritma K-NN memiliki kinerja yang sangat baik dalam mengklasifikasikan data tingkat kemiskinan di wilayah Indonesia secara akurat.

Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat diketahui bahwa algoritma *K-Nearest Neighbor* menunjukkan kinerja yang lebih baik dibandingkan dengan *Naive Bayes* dalam mengklasifikasikan tingkat kemiskinan di Indonesia. K-NN mencapai akurasi sebesar 97,09%, sedangkan *Naive Bayes* hanya mencapai akurasi 95,15%. Hal ini menunjukkan bahwa K-NN lebih efektif dalam mengenali pola data sosial-ekonomi dan memberikan hasil klasifikasi yang lebih akurat, terutama karena kemampuannya mempertimbangkan kedekatan antar data dalam ruang fitur. Hasil penelitian ini mengonfirmasi bahwa pemilihan parameter K yang optimal serta penerapan tahap pra-pemrosesan yang tepat berperan penting dalam meningkatkan performa model klasifikasi kemiskinan berbasis data. Selain itu, performa model ini lebih baik dibandingkan hasil dari penelitian sebelumnya, sehingga K-NN dapat dijadikan acuan dalam pengembangan model prediktif untuk mendukung kebijakan sosial berbasis data.

Penelitian selanjutnya disarankan untuk memperluas jumlah data serta menambahkan variabel sosial dan ekonomi lain seperti tingkat inflasi, tingkat pendidikan, dan akses terhadap lapangan kerja guna meningkatkan ketepatan model. Selain itu, perlu dilakukan pengujian terhadap algoritma lain seperti *Random Forest* atau *Support Vector Machine* (SVM) untuk membandingkan kinerja dan mendapatkan model yang lebih robust. Implementasi sistem berbasis web atau dashboard interaktif juga direkomendasikan agar hasil klasifikasi dapat dimanfaatkan oleh pemerintah daerah dalam perumusan kebijakan penanggulangan kemiskinan yang lebih tepat sasaran.

DAFTAR PUSTAKA

Adane, M. D., Deku, J. K., & Asare, E. K. (2023). Performance Analysis of Machine Learning Algorithms in Prediction of Student Academic Performance. *Journal of Advances in Mathematics and Computer Science*, *38*(5), 74–86. https://doi.org/10.9734/jamcs/2023/v38i51762

Agus Triono, T., & Sangaji, R. C. (2023). Faktor Mempengaruhi Tingkat Kemiskinan di Indonesia: Studi Literatur Laporan Data Kemiskinan BPS Tahun 2022. *Journal of Society Bridge*, *1*(1), 59–67. https://doi.org/10.59012/jsb.v1i1.5

Alfitriana Riska, Purnawansyah, Darwis, H., & Astuti, W. (2023). Studi Perbandingan Kombinasi GMI, HSV, KNN, dan CNN pada Klasifikasi Daun Herbal. *The Indonesian Journal of Computer Science*, *12*(3), 1201–1215. https://doi.org/10.33022/ijcs.v12i3.3210

Amalia, W., Nur, C., Abdullah, D., & Meiyanti, R. (2025). Performance of K-Nearest Neighbor Algorithm and C4 . 5 Algorithm in Classifying Citizens Eligible to Receive Direct Cash Assistance in Bandar Mahligai Village. *International Journal of Engineering, Science and Information Technology*, *5*(1), 368–372.

Anna, A. (2023). Pengujian Teknik Algoritma Klasifikasi Terhadap Tingkat Kemiskinan Penduduk. *JTIK (Jurnal Teknik Informatika Kaputama)*, *7*(1), 61–66. https://doi.org/10.59697/jtik.v7i1.35

Djafar, N. M., & Fauzan, A. (2024). Implementation of K-Nearest Neighbor using the

Vol 6 No 1 (2026) 156–172 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10368

oversampling technique on mixed data for the classification of household welfare status. *Statistics in Transition New Series*, *25*(1), 109–124. https://doi.org/10.59170/stattrans-2024-007

Duwo Jiwo Saputro, A., Darmawan, A., & Nurina Sari, B. (2024). Klasifikasi Persentase Kemiskinan Di Jawa Barat Menggunakan Data Mining Algoritma K-Nearest Neighbor (Knn). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(4), 2718–2723. https://doi.org/10.36040/jati.v7i4.7178

Fauziah, Tiro, M. A., & Ruliana. (2022). Comparison of k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) Methods for Classification of Poverty Data in Papua. *ARRUS Journal of Mathematics and Applied Science*, *2*(2), 83–91. https://doi.org/10.35877/mathscience741

Fitra, R., & Rusdi, I. (2022). Penerapan Metode Algoritma K-Nearest Neighbor Menggunakan Rapidminer Studio Pada Klasifikasi Status Sosial Ekonomi Studi Kasus : Kelurahan Kapuk Muara Rt 010 Rw 04. *Smart Comp: Jurnalnya Orang Pintar Komputer*, *11*(4), 653–660. https://doi.org/10.30591/smartcomp.v11i4.4250

Mardiah, A., Defni, D., Lestari, A. H., Junaldi, J., & Ritmi, T. (2024). Classification of Population Data of Nagari Based on Economic Level Using The K-Nearest Neighbor Method. *International Journal of Advanced Science Computing and Engineering*, *6*(1), 32–35. https://doi.org/10.62527/ijasce.6.1.191

Rivaldo, Vito Junivan, T. A. Y., & Pranoto, W. J. (2024). Perbaikan Akurasi Naïve Bayes dengan Chi-Square dan SMOTE Dalam Mengatasi High Dimensional dan Imbalanced Data Banjir. *Jurnal Media Informatika Budidarma*, *8*(3), 1656. https://doi.org/10.30865/mib.v8i3.7886

Suci Mulyani, Pajri, A. E., & Fikram, M. (2024). Klasisifikasi Tingkat Kemiskinan Di Indonesia Menggunakan Algoritma Naives Bayes. *Scientific : Journal of Computer Science and Informatics*, *1*(2), 53–57. https://doi.org/10.34304/scientific.v1i2.333