

Perbandingan Naive Bayes dan Support Vector Machine dalam Klasifikasi Tingkat Kemiskinan di Indonesia

Zulfa Alviandri Mukharyahya^{1,*}, Yani Parti Astuti¹, Okta Nur Cahyani¹

¹ Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Indonesia

* Correspondence: 111202113735@mhs.dinus.ac.id

Copyright: © 2025 by the authors

Received: 4 Februari 2025 | Revised: 12 Februari 2025 | Accepted: 7 Maret 2025 | Published: 13 April 2025

Abstrak

Kemiskinan di Indonesia adalah sebuah masalah yang kompleks, dipengaruhi oleh berbagai faktor ekonomi dan sosio-budaya. Penelitian ini bertujuan untuk membandingkan kinerja *naive bayes* dan *Support Vector Machine* (SVM) dalam mengklasifikasikan tingkat kemiskinan di Indonesia, sekaligus mengevaluasi efektivitas random *oversampling* dalam mengatasi masalah ketidakseimbangan data. Dataset yang digunakan terdiri dari 514 sampel yang berasal dari berbagai kabupaten dan kota di Indonesia, dengan rincian kelas "tidak miskin" sebanyak 452 sampel dan "miskin" sebanyak 62 sampel. Setelah penerapan *oversampling*, jumlah sampel meningkat menjadi 730 dengan distribusi yang seimbang (365 untuk masing-masing kelas). Data yang diamati mencakup indikator sosial-ekonomi, seperti persentase penduduk miskin, pengeluaran per kapita, indeks pembangunan manusia, dan tingkat pengangguran terbuka. Penelitian ini membagi data dengan rasio 80:20 untuk keperluan pelatihan dan pengujian model. Hasil eksperimen menunjukkan bahwa SVM memiliki tingkat akurasi yang lebih tinggi, yaitu 81%, dibandingkan dengan *naive bayes* yang mencapai 76%. Selain itu, SVM juga menunjukkan keseimbangan *precision* dan *recall* yang lebih stabil. Di sisi lain, penerapan teknik *oversampling* terbukti meningkatkan kemampuan model dalam mengidentifikasi kelas minoritas, terutama pada *naive bayes* yang lebih responsif terhadap adanya duplikasi data. Temuan ini menegaskan pembelajaran mesin dalam merancang kebijakan sosial yang lebih efektif untuk pengelolaan data kemiskinan.

Kata kunci: kemiskinan; *naive bayes*; *svm*; *oversampling*

Abstract

Poverty in Indonesia is a complex issue influenced by various economic and socio-cultural factors. This study aims to compare the performance of Naïve Bayes and Support Vector Machine (SVM) in classifying poverty levels in Indonesia while also evaluating the effectiveness of random oversampling in addressing data imbalance. The dataset consists of 514 samples from various districts and cities in Indonesia, with 452 samples classified as "not poor" and 62 as "poor." After applying oversampling, the total number of samples increased to 730, with a balanced distribution (365 samples per class). The observed data include socio-economic indicators such as the percentage of the poor population, per capita expenditure, the Human Development Index, and the open unemployment rate. The study splits the data using an 80:20 ratio for training and testing. Experimental results show that SVM achieved a higher accuracy of 81% compared to naïve bayes, which reached 76%. Additionally, SVM demonstrated a more stable balance between precision and recall. On the other hand, the oversampling technique effectively improved the model's ability to identify the minority class, particularly for Naïve Bayes, which was more responsive to data duplication. These findings highlight the role of machine learning in designing more effective social policies for poverty data management.

Keywords: poverty; *naive bayes*; *svm*; *oversampling*



PENDAHULUAN

Kemiskinan merupakan masalah sosial yang kompleks yang masih menjadi tantangan utama di banyak negara, termasuk Indonesia. Berdasarkan Badan Pusat Statistik (BPS), angka kemiskinan Indonesia akan mencapai 9,36% pada tahun 2023, dengan distribusi yang tidak merata antara daerah perkotaan dan pedesaan. Faktor-faktor seperti akses terhadap pendidikan dan pekerjaan serta kualitas infrastruktur berkontribusi terhadap ketidaksetaraan ini. Meskipun berbagai langkah telah dilaksanakan untuk mengurangi angka kemiskinan, tantangan terbesar adalah mengidentifikasi kelompok populasi yang benar-benar membutuhkan dukungan sehingga bantuan sosial dapat didistribusikan dengan lebih baik. Oleh karena itu, pendekatan yang lebih sistematis dan berbasis data diperlukan untuk mengklasifikasikan tingkat kemiskinan secara lebih akurat (Anna, 2023; Putri et al., 2021; Triono & Sangaji, 2023).

Tingkat kemiskinan di Indonesia merupakan masalah yang sangat kompleks. Data seringkali tidak terstruktur seperti hasil survei atau laporan berbasis teks, dan dapat sulit dianalisis secara manual (Pahlawan & Ratna, 2020). Selain itu, pengolahan data tradisional memerlukan banyak waktu dan upaya yang mengurangi efisiensi pengolahan data yang terus meningkat. lain itu, pengolahan data secara tradisional memerlukan waktu dan tenaga yang sangat besar, sehingga menjadi kurang efisien untuk mengatasi volume data yang terus berkembang (Ikhsani et al., 2024). Hal ini menambah kompleksitas dalam membangun model klasifikasi yang akurat dan efektif. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengotomatisasi proses pengklasifikasian tingkat kemiskinan secara efisien dan sekaligus mengatasi masalah ketimpangan kelas guna meningkatkan akurasi hasil klasifikasi (Widodo et al., 2021).

Saat mengklasifikasikan kemiskinan, salah satu tantangan utama yang dihadapi adalah ketidakseimbangan data. Dalam banyak kasus, jumlah sampel kategori "tidak miskin" jauh lebih dominan dibandingkan dengan kategori "miskin". Ketidakseimbangan ini menyebabkan model pembelajaran mesin cenderung lebih akurat dalam mengenali kelompok mayoritas, namun kurang efektif dalam mengidentifikasi kelompok minoritas yang merupakan fokus utama kebijakan sosial (Al-Aluf & Fatah, 2025).. Selain itu, metode konvensional berbasis statistik sering kali memiliki keterbatasan dalam menangani data yang besar dan kompleks, sehingga diperlukan pendekatan yang lebih adaptif dan cerdas (Indey & Supangat, 2024; Mulyanan et al., 2024; Vebrian & Kustiyono, 2024). Pembelajaran mesin (*machine learning*) telah muncul sebagai solusi inovatif di berbagai bidang, termasuk analisis sosial-ekonomi. Algoritma klasifikasi seperti *naive bayes* dan *Support Vector Machine* (SVM) dapat membantu mengidentifikasi pola dalam data serta melakukan klasifikasi tingkat kemiskinan berdasarkan berbagai indikator sosial-ekonomi.

Naive bayes adalah metode probabilistik yang menggunakan Teorema Bayes, dengan asumsi bahwa setiap fitur dalam data bersifat independen (Armansyah & Ramli, 2022). Meskipun metode ini sederhana dan cepat, asumsi independensinya terkadang tidak sesuai dalam konteks data sosial-ekonomi. Di sisi lain, SVM berfungsi dengan mencari hyperplane optimal untuk memisahkan kelas dalam ruang fitur (Frenica et al., 2023; Pratama & Prasetyaningrum, 2025; Setyawan et al., 2024). SVM terbukti efektif dalam menangani data nonlinier, tetapi memerlukan optimasi parameter agar dapat berfungsi secara optimal pada dataset tertentu. Namun, karena adanya ketidakseimbangan data, model seringkali cenderung mengklasifikasikan data ke dalam kelas mayoritas ("tidak miskin"). Oleh karena itu, diperlukan metode penanganan khusus seperti *random oversampling*, yang dapat membantu menyeimbangkan distribusi kelas dan meningkatkan performa model dalam mengidentifikasi kelompok miskin (Khasanah, 2021; Prasetyono & Anggraini, 2021; Restiatun et al., 2023; Saragih et al., 2022).

Penelitian sebelumnya yang dilakukan oleh Chazar et al. (2022) menerapkan metode SVM untuk mengklasifikasikan keadaan penduduk miskin di Desa Taraju, Kabupaten

Tasikmalaya. Hasil penelitian mereka menunjukkan bahwa model yang dibuat memiliki tingkat akurasi yang rendah. Keterbatasan lain yang dihadapi adalah belum diterapkannya teknik oversampling karena adanya ketidakseimbangan yang tidak diseimbangkan sehingga kurang efektif. Di sisi lain, meskipun SVM telah digunakan dalam berbagai studi klasifikasi data sosial-ekonomi, masih sedikit penelitian yang secara langsung membandingkan performa SVM dengan *naive bayes* dalam konteks klasifikasi kemiskinan di Indonesia. Sebagian besar studi sebelumnya lebih memfokuskan pada penggunaan salah satu algoritma tanpa mengeksplorasi perbedaan performa keduanya. Selain itu, relatif sedikit penelitian yang menggunakan teknik oversampling untuk menangani ketidakseimbangan data dalam klasifikasi tingkat kemiskinan. Oleh karena itu, terdapat kesenjangan dalam penelitian yang perlu diisi dengan pendekatan yang lebih komprehensif, guna meningkatkan *Akurasi* dan efektivitas dalam klasifikasi kemiskinan.

Penelitian lain yang dilakukan oleh Nurmawati et al. (2021) menggunakan metode *naive bayes* untuk mengklasifikasikan kemiskinan di Desa Lepak, dan berhasil mencapai akurasi sebesar 96,63%, sensitivitas 94,87%, serta spesifisitas 99,1%. Meskipun hasil ini terlihat menjanjikan, penelitian mereka tidak membandingkan keakuratan model yang digunakan dengan studi sebelumnya atau dengan metode lain seperti *decision tree* atau SVM. Oleh karena itu, sulit untuk menilai apakah *naive bayes* adalah pilihan terbaik dalam klasifikasi kemiskinan. Selain itu, penelitian ini juga tidak menyertakan analisis ROC curve dan AUC yang sangat penting untuk mengevaluasi keseimbangan antara *True Positive Rate* dan *False Positive Rate*. Tanpa evaluasi tersebut, risiko bias klasifikasi tidak dapat diukur dengan optimal, yang berpotensi memengaruhi kebijakan, seperti dalam pemberian bantuan sosial. Dengan demikian, penelitian ini masih memerlukan pendekatan yang lebih komprehensif, termasuk perbandingan metode serta evaluasi model menggunakan ROC curve, guna meningkatkan keakuratan dan efektivitas dalam klasifikasi kemiskinan.

Penelitian kami ini bertujuan untuk membandingkan kinerja *naive bayes* dan SVM dalam mengklasifikasikan tingkat kemiskinan di Indonesia. Selain itu, penelitian ini juga menerapkan teknik *random oversampling* untuk mengatasi masalah ketidakseimbangan data, sehingga model yang dihasilkan dapat lebih akurat dalam mengidentifikasi kelompok miskin. Evaluasi dilakukan dengan menggunakan metrik performa seperti akurasi, presisi, *recall*, dan *f1-score*, untuk menentukan algoritma mana yang lebih unggul dalam klasifikasi kemiskinan. Melalui pendekatan ini, diharapkan penelitian ini dapat memberikan kontribusi signifikan dalam pemanfaatan pembelajaran mesin untuk mendukung pengambilan keputusan dalam kebijakan sosial, terutama dalam merumuskan strategi pengentasan kemiskinan yang lebih tepat sasaran.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasi, bertujuan untuk mengklasifikasikan tingkat kemiskinan di Indonesia dengan memanfaatkan algoritma SVM dan *naive bayes*. Dataset yang digunakan diambil dari Kaggle, mencakup 514 data dari kabupaten/kota yang dilengkapi dengan berbagai fitur sosial-ekonomi. Di antara fitur tersebut terdapat persentase penduduk miskin, pengeluaran per kapita, indeks pembangunan manusia, dan tingkat pengangguran terbuka. Data tersebut dibagi menjadi dua kategori, yaitu tidak miskin (0) dan miskin (1). Tujuan utama pengelompokan ini adalah untuk menganalisis faktor-faktor yang berkontribusi terhadap kemiskinan serta meningkatkan akurasi klasifikasi dengan menerapkan teknik *oversampling*. Proses dalam penelitian ini meliputi beberapa tahap, yaitu pengumpulan data, *preprocessing*, pembagian data, penanganan ketidakseimbangan data, pembangunan model, dan evaluasi model. Pada tahap pertama, dataset dikumpulkan dari sumber yang telah disebutkan, dengan mencakup informasi tentang ukuran dataset (514 sampel) dan distribusi kelas, agar memberikan gambaran mengenai ketidakseimbangan data. Selanjutnya, dilakukan *preprocessing* data, yang mencakup

penghapusan nilai-nilai kosong menggunakan metode yang sesuai, penyesuaian format angka agar bisa diproses oleh library statistik, serta konversi variabel target menjadi bentuk numerik (1 untuk miskin dan 0 untuk tidak miskin). Selain itu, diterapkan juga feature scaling menggunakan *StandardScaler* untuk memastikan bahwa SVM dapat beroperasi secara optimal. Setelah tahap preprocessing selesai, dataset dibagi menjadi fitur (X) dan target (y) dengan menggunakan rasio 80:20 untuk memisahkan data pelatihan dan pengujian. Pembagian ini dilakukan secara stratified agar proporsi masing-masing kelas tetap terjaga seimbang di kedua set data. Untuk mengatasi masalah ketidakseimbangan data, penelitian menerapkan teknik *Random Oversampling* (ROS) yang hanya diterapkan pada data pelatihan, untuk menghindari kebocoran informasi. Sampel pada kelas minoritas ditambahkan hingga tercapai keseimbangan dengan kelas mayoritas.

Naive bayes adalah algoritma klasifikasi yang mengandalkan Teorema Bayes dan mengasumsikan bahwa fitur-fitur dalam dataset bersifat independen satu sama lain. Dalam konteks klasifikasi kemiskinan, algoritma ini berperan dalam menentukan kategori "miskin" atau "tidak miskin" dengan mempertimbangkan probabilitas berdasarkan karakteristik sosial-ekonomi. Salah satu keunggulan *naive bayes* adalah kecepatan dalam pemrosesan data. Namun, terdapat kelemahan dalam asumsi independensi yang tidak selalu dapat dipertahankan. Di sisi lain, SVM digunakan karena kemampuannya untuk menangani data dengan dimensi tinggi dan pola nonlinier, sehingga dapat menghasilkan klasifikasi yang lebih akurat.

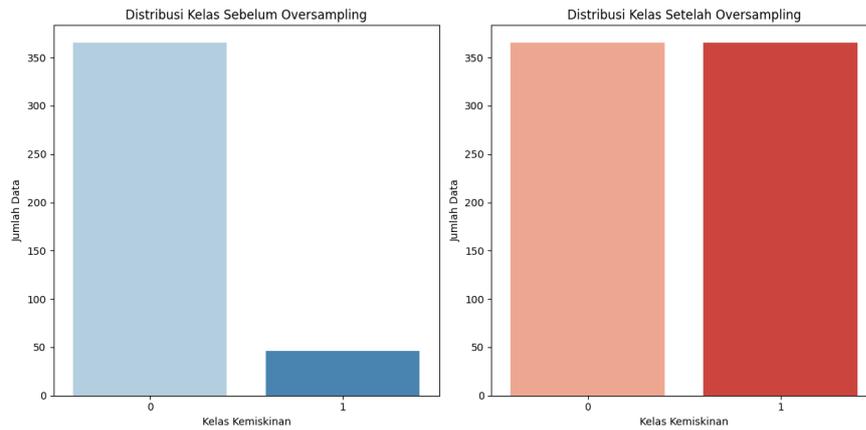
Evaluasi model pada penelitian ini menggunakan beberapa metrik utama, yaitu akurasi, presisi, *recall*, dan *f1-score*. Akurasi diukur dengan persentase prediksi yang benar di seluruh dataset uji, sedangkan presisi berfungsi untuk memastikan bahwa prediksi kelas "miskin" sesuai dengan kenyataan. *Recall* sangat krusial dalam penelitian ini, karena berfokus pada identifikasi semua individu yang benar-benar miskin. *F1-score* berguna untuk menyeimbangkan antara *presisi* dan *recall*, khususnya dalam situasi di mana terdapat ketidakseimbangan kelas. Dengan menerapkan teknik-teknik ini, penelitian ini bertujuan untuk meningkatkan kinerja model dalam mendeteksi kelas minoritas "miskin" serta memahami faktor-faktor sosial-ekonomi yang berkontribusi terhadap kemiskinan.

HASIL DAN PEMBAHASAN

Hasil

Dataset pada penelitian ini diambil dari data klasifikasi kemiskinan Indonesia dan mencakup karakteristik seperti persentase jumlah penduduk miskin, pengeluaran per kapita, indeks pembangunan manusia, dan tingkat pengangguran serta klasifikasi kemiskinan sebagai variabel target. Data kemudian diproses melalui tahap *preprocessing* ini menghapus nilai kosong, mengubah format numerik, dan mengonversi variabel target ke format numerik. Untuk mengatasi ketidakseimbangan data, teknik *random oversampling* diterapkan pada kelas-kelas yang "miskin" untuk memastikan model belajar tanpa bias terhadap kelas mayoritas. Data yang diproses dan diseimbangkan digunakan untuk melatih dan menguji model.

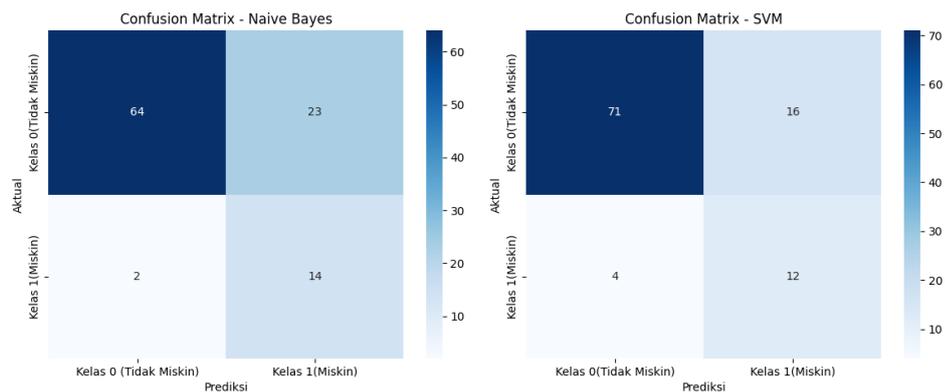
Pada gambar 1 menunjukkan distribusi data kelas sebelum dan sesudah proses *oversampling*. Pada grafik di sebelah kiri, kita dapat melihat bahwa jumlah data di kelas 0 jauh lebih tinggi daripada kelas 1, yang membuat data sangat tidak seimbang. Ketidakseimbangan ini dapat menyebabkan model pembelajaran mesin cenderung memihak pada kelas mayoritas. Hal ini membuat sulit untuk mengidentifikasi pola minoritas. Di sisi lain, gambar di sebelah kanan menunjukkan hasil setelah *oversampling*, yang membuat jumlah data untuk kelas 0 dan kelas 1 sama. Proses *oversampling* ini bertujuan untuk mengatasi masalah ketidakseimbangan data dan memungkinkan model mempelajari kedua kelas secara lebih adil dan mencapai kinerja klasifikasi yang lebih baik.



Gambar 1. Hasil *oversampling*

Penelitian ini membandingkan kinerja model *naive bayes* dan SVM dalam klasifikasi data. *Naive Bayes* menggunakan probabilitas kondisional antar fitur, sementara SVM memisahkan kelas dengan hyperplane berdasarkan margin maksimum. Evaluasi kedua model dilakukan menggunakan *confusion matrix* yang terdiri dari empat komponen utama: *True Positive* (TP), yaitu jumlah sampel yang benar-benar teridentifikasi sebagai "Miskin"; *True Negative* (TN), yakni jumlah sampel yang tepat diklasifikasikan sebagai "Tidak Miskin"; *False Positive* (FP), yaitu jumlah sampel "Tidak Miskin" yang keliru diklasifikasikan sebagai "Miskin"; dan *False negative* (FN), yaitu jumlah sampel "Miskin" yang salah terklasifikasi sebagai "Tidak Miskin".

Hasil analisis *confusion matrix*, model *naive bayes* mencatat 64 TN, 14 TP, dengan 23 FP dan 2 FN. Sedangkan model SVM mencatat 71 TN dan 12 TP, dengan 16 FP dan 4 FN. *Naive bayes* memiliki FN yang lebih rendah dibandingkan dengan SVM, yang berarti lebih sedikit individu "Miskin" yang salah diklasifikasikan sebagai "Tidak Miskin". Namun, model ini juga memiliki FP yang lebih tinggi, yang berarti lebih banyak individu "Tidak Miskin" yang diklasifikasikan secara keliru sebagai "Miskin". Di sisi lain, SVM menunjukkan jumlah TN yang lebih tinggi dibandingkan dengan *Naive Bayes*, menunjukkan bahwa model ini lebih baik dalam mengidentifikasi individu "Tidak Miskin" secara benar. Namun, FN yang lebih tinggi pada SVM menunjukkan bahwa model ini lebih sering gagal mendeteksi individu "Miskin", yang dapat berimplikasi pada ketidaktepatan distribusi bantuan sosial. Dalam konteks kebijakan sosial, meminimalkan FN menjadi prioritas karena kesalahan dalam mengklasifikasikan individu "Miskin" sebagai "Tidak Miskin" dapat menghambat penyaluran bantuan sosial yang tepat sasaran. Dengan FN yang lebih rendah, *Naive Bayes* lebih unggul dalam mendeteksi individu "Miskin", meskipun memiliki tingkat FP yang lebih tinggi dibandingkan dengan SVM. Hasil dari percobaan ini dapat dilihat dari gambar 2 dan tabel 1.

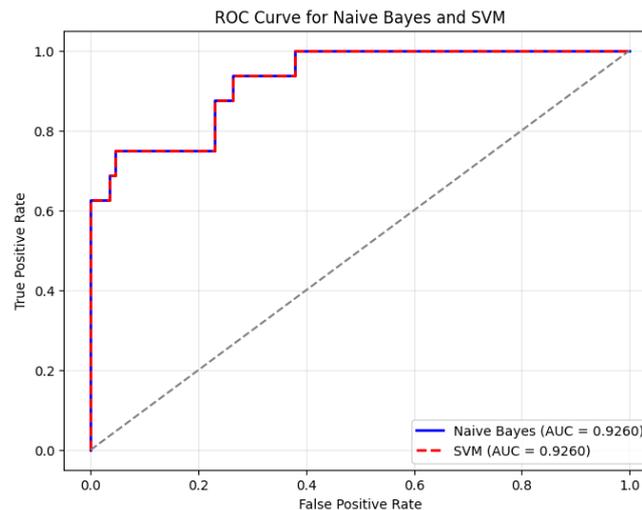


Gambar 2. *Confusion matrix*

Tabel 1. Hasil evaluasi klasifikasi

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	76%	67%	81%	68%
SVM	81%	69%	78%	71%

Hasil pada tabel 1 menunjukkan perbandingan performa antara *Naive Bayes* dan SVM berdasarkan beberapa indikator, yaitu akurasi *accuracy*, *precision*, *recall*, dan *f1-score*. Secara keseluruhan, SVM menunjukkan performa yang lebih baik dalam hal *accuracy* (81%) dan *f1-score* (71%), yang menggambarkan keseimbangan antara *precision* dan *recall*. Di sisi lain, *naive bayes* memiliki nilai *recall* yang lebih tinggi (81%) dibandingkan SVM (78%), menunjukkan kemampuannya untuk menangkap lebih banyak individu dari kelas "miskin". Dalam konteks klasifikasi kemiskinan, *recall* menjadi sangat penting karena penelitian ini berfokus pada deteksi kelas minoritas "miskin" demi mengurangi risiko tidak terdeteksinya individu yang memerlukan bantuan. *F1-score* juga menjadi pertimbangan yang signifikan, karena metrik ini menggabungkan aspek *recall* dan *precision* untuk menilai keseimbangan performa model. Walaupun *naive bayes* memiliki *precision* yang lebih tinggi, keunggulan SVM dalam *accuracy* dan *f1-score* menunjukkan tingkat stabilitasnya dalam mengklasifikasikan data secara menyeluruh. Performa masing-masing model juga sangat bergantung pada pemilihan parameter, seperti jenis kernel pada SVM dan asumsi independensi fitur pada *naive bayes*. Jika optimasi parameter belum dilakukan, masih ada peluang untuk meningkatkan hasil yang diperoleh. Dari sudut pandang implementasi, jika tujuan utama adalah mendeteksi sebanyak mungkin individu yang miskin, maka *naive bayes* lebih unggul. Namun, jika keseimbangan performa keseluruhan menjadi prioritas, maka SVM adalah pilihan yang lebih tepat untuk mendukung kebijakan sosial berbasis data.



Gambar 3. Kurva roc

Pada gambar 3 menampilkan kurva *Receiver Operating Characteristic* (ROC) untuk model *naive bayes* dan SVM, di mana keduanya memiliki nilai *Area Under the Curve* (AUC) sebesar 0,926. Nilai ini menunjukkan bahwa kedua model memiliki kemampuan yang hampir setara dalam membedakan antara kategori "miskin" dan "tidak miskin". Kurva yang mendekati sudut kiri atas adalah indikasi dari performa klasifikasi yang baik. Meskipun SVM menunjukkan *accuracy* dan *f1-score* yang lebih tinggi, *naive bayes* justru unggul dalam *recall*, yang sangat penting untuk mendeteksi individu miskin. Oleh karena untuk meminimalkan kesalahan dalam identifikasi kelas miskin, maka model *naive bayes* lebih direkomendasikan untuk digunakan.

Pembahasan

Penelitian ini menggunakan dataset yang berkaitan dengan klasifikasi tingkat kemiskinan di Indonesia. Dataset ini diperoleh dari Kaggle dan terdiri dari 514 data yang diambil dari berbagai kabupaten dan kota. Data tersebut dilengkapi dengan fitur sosial-ekonomi, seperti persentase penduduk miskin, pengeluaran per kapita, indeks pembangunan manusia, serta tingkat pengangguran terbuka. Data ini dikelompokkan ke dalam dua kategori, yaitu "tidak miskin" (0) dan "miskin" (1). Dataset awal terdapat ketidakseimbangan dalam distribusi kelas, dengan jumlah sampel untuk kategori "tidak miskin" mencapai 452, sementara kategori "miskin" hanya memiliki 62 sampel. Ketidakseimbangan ini berpotensi membuat model lebih condong dalam mengklasifikasikan data sebagai "tidak miskin," sehingga mengakibatkan penurunan akurasi dalam mendeteksi individu yang sebenarnya miskin. Untuk mengatasi permasalahan ini, diterapkan teknik *random oversampling* yang meningkatkan jumlah sampel dalam kelas "miskin" sampai menjadi 365. Dengan demikian, total dataset pasca-*oversampling* mencapai 730 sampel dengan distribusi yang lebih seimbang (365:365). Teknik ini memberikan kesempatan bagi model untuk lebih memahami pola yang berkaitan dengan kemiskinan.

Hasil klasifikasi menunjukkan bahwa model *naïve bayes* memiliki keunggulan dalam mendeteksi individu miskin jika dibandingkan dengan SVM. Meskipun SVM unggul dalam hal *accuracy* (81% berbanding 76%) dan *f1-score* yang lebih baik (71% dibandingkan 68%), *Naïve Bayes* menunjukkan keunggulan dalam *recall*, yaitu 81% berbanding 78%. *Recall* yang lebih tinggi menunjukkan bahwa *naïve bayes* lebih efektif dalam mengenali individu miskin dengan tingkat kesalahan klasifikasi yang lebih rendah (*false negative*). Variabel yang paling berpengaruh dalam klasifikasi ini adalah Pengeluaran per Kapita Disesuaikan, Indeks Pembangunan Manusia (IPM), Tingkat Pengangguran Terbuka, dan Persentase Penduduk Miskin. Pengeluaran per Kapita menjadi faktor utama, di mana semakin rendah pengeluaran seseorang, semakin besar kemungkinan untuk diklasifikasikan sebagai miskin. IPM juga berperan penting karena daerah dengan nilai IPM rendah cenderung memiliki tingkat kemiskinan yang lebih tinggi. Selain itu, Tingkat Pengangguran Terbuka dan Persentase Penduduk Miskin juga turut memperkuat prediksi model.

Berdasarkan hasil analisis kami, Pengeluaran per Kapita dan IPM menjadi faktor dominan dalam membedakan individu miskin dan tidak miskin. *Naïve bayes* lebih efektif dalam mengenali pola keterkaitan antarvariabel ini, sehingga mampu mencapai *recall* yang lebih tinggi jika dibandingkan dengan SVM. Keunggulan *naïve bayes* dalam *recall* dapat dijelaskan oleh pendekatan probabilistik yang digunakannya, yang lebih responsif terhadap data yang tidak seimbang dibandingkan dengan SVM. Model ini mampu menangkap pola distribusi fitur dengan lebih baik, sehingga lebih akurat dalam mengidentifikasi individu miskin. Di sisi lain, SVM cenderung lebih sensitif terhadap kelas mayoritas, sehingga berisiko lebih tinggi untuk salah mengklasifikasikan individu miskin sebagai bukan miskin. Dalam konteks kebijakan sosial, *recall* yang tinggi sangatlah penting, karena kesalahan dalam mengidentifikasi individu miskin dapat mengakibatkan mereka tidak menerima bantuan sosial yang mereka perlukan. Oleh karena itu, penelitian ini lebih mengedepankan penggunaan *naïve bayes*, karena model ini lebih dapat diandalkan dalam mendeteksi individu miskin dengan lebih akurat, serta mengurangi risiko kesalahan dalam pengklasifikasian kelas minoritas.

Penelitian yang dilakukan oleh Chazar et al. (2022) menerapkan metode SVM untuk mengklasifikasikan keadaan penduduk miskin di Desa Taraju, Kabupaten Tasikmalaya. Hasil penelitian mereka menunjukkan bahwa model yang dibuat memiliki tingkat akurasi yang rendah. serta tidak diterapkannya teknik *oversampling* untuk mengatasi masalah ketidakseimbangan data. Selain itu, penelitian mereka tidak melakukan perbandingan performa *naïve bayes* dengan metode lain seperti SVM, yang mengakibatkan kesulitan dalam menilai keunggulan model yang digunakan. Di sisi lain, penelitian yang dilakukan oleh Nurmayanti et

al. (2021) juga menerapkan metode *naive bayes* untuk mengklasifikasikan kemiskinan di Desa Lepak, dan mencatat *accuracy* yang lebih tinggi, yaitu 96,63%. Namun, penelitian ini pun tidak membandingkan tingkat keakuratan *naive bayes* dengan metode lain seperti *decision tree* atau SVM.

Selain itu, ketiadaan analisis *ROC curve* dan *AUC* membuat evaluasi terkait keseimbangan *True Positive Rate* dan *False Positive Rate* menjadi kurang mendalam. Dengan demikian, baik penelitian Nurmayanti et al. (2021) dan Chazar et al. (2022) masih memiliki keterbatasan dalam evaluasi model dan perbandingan metode. Oleh karena itu, diperlukan pendekatan yang lebih komprehensif untuk meningkatkan akurasi dan efektivitas klasifikasi kemiskinan. Untuk pengembangan lebih lanjut, penelitian ini bisa ditingkatkan dengan menerapkan teknik SMOTE atau ADASYN guna mengatasi masalah ketidakseimbangan data secara lebih efektif, serta melakukan validasi silang (*k-fold cross-validation*) untuk meningkatkan stabilitas model. Eksplorasi terhadap model yang lebih kompleks seperti *random forest*, *XGBoost*, atau pendekatan berbasis *deep learning* (CNN, LSTM) juga perlu dipertimbangkan. Analisis pengaruh fitur tambahan yang lebih spesifik terhadap klasifikasi kemiskinan juga dapat memberikan wawasan yang berharga. Dengan model yang lebih akurat dan andal, sistem ini berpotensi menjadi dasar bagi pengembangan sistem informasi kemiskinan yang dapat diintegrasikan dalam kebijakan pemerintah, guna membantu masyarakat miskin dengan lebih efektif.

SIMPULAN

Penelitian ini menunjukkan bahwa metode *random oversampling* efektif dalam menyeimbangkan data, sehingga meningkatkan akurasi model dalam mengidentifikasi kelompok masyarakat yang berada dalam kondisi miskin. Dalam hal ini, algoritma *naive bayes* menunjukkan performa yang lebih baik dibandingkan SVM, dengan nilai *recall* yang sangat tinggi mencapai 98%. Hal ini memastikan bahwa kesalahan klasifikasi terhadap individu yang miskin dapat diminimalkan. Di sisi lain, SVM mengalami keterbatasan, dengan nilai *recall* yang lebih rendah yaitu 78% serta tingkat FN yang lebih tinggi. Metode *naive bayes* terbukti lebih optimal untuk analisis data sosial ekonomi karena kemampuannya dalam menangani pola distribusi data yang kompleks secara probabilistik. Temuan dari penelitian ini memiliki implikasi penting bagi kebijakan sosial, di mana model dengan tingkat *recall* yang tinggi dapat membantu pemerintah dalam menyalurkan bantuan dengan lebih tepat sasaran. Penelitian selanjutnya diharapkan dapat mengeksplorasi penggunaan dataset yang lebih luas dan memanfaatkan algoritma yang lebih canggih, sehingga model yang dikembangkan dapat diterapkan dalam skala yang lebih besar.

REFERENSI

- Al-Aluf, M. W., & Fatah, Z. (2025). Klasifikasi Algoritma Decision Tree Untuk Tingkat Kemiskinan Di Indonesia. *Journal of Computer Science and Technology*, 3, 55–62. <https://doi.org/10.59435/jocstec.v3i1.404>
- 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>
- Armansyah, A., & Ramli, R. K. (2022). Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes. *Edumatic: Jurnal Pendidikan Informatika*, 6(1), 1–10. <https://doi.org/10.29408/edumatic.v6i1.4789>
- Chazar, C., Hendra Gunawan, & Sumpena. (2022). Implementasi Support Vektor Machine Pada Klasifikasi Penduduk Miskin Wilayah Desa Taraju Kabupaten Tasikmalaya. *INFORMASI: Jurnal Informatika Dan Sistem Informasi*, 14(1), 80–93. <https://doi.org/10.37424/informasi.v14i1.150>

- Frenica, A., Lindawati, L., & Soim, S. (2023). Implementasi Algoritma Support Vector Machine (SVM) untuk Deteksi Banjir. *INOVTEK Polbeng - Seri Informatika*, 8(2), 291. <https://doi.org/10.35314/isi.v8i2.3443>
- Ikhsani, Y., Permana, I., Nur Salisah, F., & Evrilyan Rozanda, N. (2024). Perbandingan Algoritma Support Vector Machine dan Naïve Bayes dalam Menganalisis Sentimen Pinjaman Online di Twitter. *Technology and Science (BITS)*, 6(3), 1413–1426.
- Indey, J. F., & Supangat, S. (2024). Implementasi Algoritma Naïve Bayes dalam Sistem Pengarsipan Surat Berbasis AI di GPI Papua Klasik Mimika Papua Tengah. *JURNAL ILMIAH INFORMATIKA*, 12(02), 102–113. <https://doi.org/10.33884/jif.v12i02.9087>
- Khasanah, L. (2021). Dampak Ketimpangan Pendapatan, Tata Kelola Pemerintahan dan Korupsi terhadap Tingkat Kemiskinan di Indonesia. *Bharanomics*, 1(2). <https://doi.org/10.46821/bharanomics.v1i2.156>
- Mulyanan, I. N., Setyawan, M. Y. H., & Rahayu, W. I. (2024). Penerapan Metode Naive Bayes untuk Merekomendasikan Pekerjaan yang Sesuai Terhadap Fresh Graduate. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(5), 3453–3460. <https://doi.org/10.36040/jati.v7i5.7155>
- Nurmayanti, W. P., Saky, D. A. L., Malthuf, M., Gazali, M., & Hirzi, R. H. (2021). Penerapan Naive Bayes dalam Mengklasifikasikan Masyarakat Miskin di Desa Lepak. *Geodika: Jurnal Kajian Ilmu Dan Pendidikan Geografi*, 5(1), 123–132. <https://doi.org/10.29408/geodika.v5i1.3430>
- Pahlawan, P. Y., & Ratna, R. (2020). Pengaruh Tingkat Pendidikan, Tingkat Pengangguran dan Pertumbuhan Ekonomi Terhadap Tingkat Kemiskinan di Indonesia Periode 2012-2017. *Journal of Malikussaleh Public Economics*, 1(2), 44–49. <https://doi.org/10.29103/jmpe.v1i2.882>
- Prasetyono, R. I., & Anggraini, D. (2021). Analisis Peramalan Tingkat Kemiskinan Di Indonesia Dengan Model Arima. *Jurnal Ilmiah Informatika Komputer*, 26(2), 95-110. <https://doi.org/10.35760/ik.2021.v26i2.3699>
- Pratama, H. I., & Prasetyaningrum, P. T. (2025). Penerapan Support Vector Machine untuk Analisis Sentimen pada Google Review Hotel. *Journal of Information System Research*, 6(2), 1246–1254.
- Putri, A. C., Hariyanto, F. E., Andini, N. L. E., & Zulkarnaen, Z. C. S. (2021). Klasifikasi Rumah Tangga Miskin di Provinsi Papua Tahun 2017 Menggunakan Metode Naive Bayes. *Jurnal Sains Matematika Dan Statistika*, 7(1), 89-95. <https://doi.org/10.24014/jsms.v7i1.11924>
- Restiatun, R., Udi, K., & Rosyadi, R. (2023). Pengaruh Pertumbuhan Sektor Pertanian, Jumlah Pekerja Sektor Pertanian dan Nilai Tukar Petani Terhadap Tingkat Kemiskinan Perdesaan di Indonesia. *Jurnal Ekonomi Pembangunan*, 12(1), 42-53. <https://doi.org/10.23960/jep.v12i1.977>
- Saragih, R. F., Silalahi, P. R., & Tambunan, K. (2022). Pengaruh Indeks Pembangunan Manusia, Tingkat Pengangguran Terbuka Terhadap Tingkat Kemiskinan di Indonesia Tahun 2007 – 2021. *PESHUM : Jurnal Pendidikan, Sosial Dan Humaniora*, 1(2), 71-79. <https://doi.org/10.56799/peshum.v1i2.36>
- Setyawan, M. F., Oktawijaya, J. D., & Agustin, S. (2024). Implementasi Support Vector Machine (SVM) pada Klasifikasi Jenis Tanah Memanfaatkan Fitur RGB. *JURNAL SISFOTENIKA*, 14(2), 175–184.
- Triono, T. A., & 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>
- Vebrian, Y. Z., & Kustiyono, K. (2024). Analisis Sentimen Masyarakat Terhadap Program Makan Siang Gratis Pada Media Sosial X Menggunakan Algoritma Naive Bayes. *Jurnal*

Informatika Dan Teknik Elektro Terapan, 12(3), 355–366.
<https://doi.org/10.23960/jitet.v12i3.4902>

Widodo, E., Ermayani, P., Laila, L. N., & Madani, A. T. (2021). Pengelompokan Provinsi di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Analisis Hierarchical Agglomerative Clustering. *Seminar Nasional Official Statistics, 2021*(1), 557–566.
<https://doi.org/10.34123/semnasoffstat.v2021i1.971>