JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 3 No: 1, Januari 2026
E-ISSN : 3047-7824

Analisis Kinerja Algoritma K-Nearest Neighbor (KNN) pada Klasifikasi Data Bank Marketing

Performance Analysis of the K-Nearest Neighbor (KNN) Algorithm in Bank Marketing Data Classification

Yosephus Arpan Polado Sinurat [1*], Hasbi Firmansyah[2], Wahyu Asriyani[3], Rizki Prasetyo Tulodo[4]
Universitas Pancasakti Tegal
Email: arpansinurat@gmail.com[1], hasbifirmansyah@upstegal.ac.id[2], asriyani1409@gmail.com[3]*
Rizki.prasetyo.tulodo@gmail.com[4]

Article Info

Article history :
Received : 30-12-2025
Revised : 01-01-2026
Accepted : 03-01-2026
Pulished : 05-01-2026

Abstract

Direct marketing is one of the banking industry's main strategies for offering time deposit products. However, untargeted campaigns are often inefficient and costly. This study aims to build a classification prediction model using the K-Nearest Neighbor (KNN) algorithm to identify potential customers for time deposits based on historical data from bank marketing campaigns. The dataset used is the Bank Marketing Dataset from the UCI Machine Learning Repository. The research process includes data pre-processing (cleaning, encoding, and Min-Max normalization), dividing training and test data, and testing different k values (k=3, 5, 7, 9). The experimental results show that the KNN algorithm with k=5 produces optimal performance with an accuracy of 89.2%, a precision of 65%, and a recall of 58%. This study concludes that KNN is effective for classifying bank marketing data, but requires class imbalance management to improve recall.

Keywords*: *Data Mining, K-Nearest Neighbor, Bank Marketing

Abstrak

Pemasaran langsung (direct marketing) merupakan salah satu strategi utama industri perbankan untuk menawarkan produk deposito berjangka. Namun, kampanye yang tidak tertarget seringkali tidak efisien dan memakan biaya tinggi. Penelitian ini bertujuan untuk membangun model prediksi klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN) untuk menentukan nasabah yang berpotensi berlangganan deposito berjangka berdasarkan data historis kampanye pemasaran bank. Dataset yang digunakan adalah *Bank Marketing Dataset* dari UCI Machine Learning Repository. Proses penelitian meliputi pra-pemrosesan data (*cleaning*, *encoding*, dan normalisasi Min-Max), pembagian data latih dan uji, serta pengujian nilai k yang berbeda (k=3, 5, 7, 9). Hasil eksperimen menunjukkan bahwa algoritma KNN dengan nilai k=5 menghasilkan kinerja optimal dengan akurasi sebesar 89,2%, presisi 65%, dan *recall* 58%. Penelitian ini menyimpulkan bahwa KNN efektif digunakan untuk klasifikasi data pemasaran bank, namun memerlukan penanganan ketidakseimbangan kelas untuk meningkatkan nilai *recall*.

Kata Kunci: Data Mining, K-Nearest Neighbor, Bank Marketing

PENDAHULUAN

Saat ini, sektor perbankan menjalankan bisnisnya di tengah kondisi ekonomi dunia yang sangat cepat berubah, sehingga kinerja operasional yang efektif dan pemasaran yang tepat menjadi hal mendasar untuk menjaga posisi terdepan dalam persaingan. Salah satu cara yang banyak dipakai adalah lewat kegiatan pemasaran langsung atau direct marketing, khususnya dalam usaha meningkatkan jumlah deposito berjangka yang berperan penting dalam menjaga kemampuan bank

untuk membayar kewajibannya. Akan tetapi, kendala terbesar yang muncul adalah kurangnya hasil yang baik dari promosi lewat telepon yang dilakukan secara besar-besaran tanpa pengelompokan yang jelas. Menelepon konsumen tanpa adanya perkiraan minat sebelumnya tidak hanya menghambur-hamburkan sumber daya berupa waktu dan uang yang besar, tetapi juga berpotensi mengganggu ketenangan konsumen dan menurunkan reputasi perusahaan di mata masyarakat. (Egejuru, 2023)

Dalam mengatasi rumitnya informasi konsumen yang meliputi berbagai hal seperti usia, keadaan keuangan, sampai catatan interaksi sebelumnya, pemakaian metode data mining menjadi hal yang wajib bagi bank zaman sekarang. Dataset Bank Marketing dari UCI Machine Learning Repository menyajikan contoh nyata tentang bagaimana faktor-faktor seperti usia, jenis pekerjaan, jumlah uang yang disimpan, sampai lama waktu kontak terakhir berpengaruh pada keputusan konsumen untuk berlangganan. Melalui cara yang terstruktur ini, data dasar yang terdiri dari ribuan baris dapat diubah menjadi informasi penting yang berguna. Perubahan dari cara membuat keputusan berdasarkan perkiraan menjadi berdasarkan data (data-driven decision making) memungkinkan para pengelola untuk memperkirakan tindakan konsumen dengan tingkat ketepatan yang lebih pasti sebelum melakukan panggilan telepon. (Parlar & Acaravci, 2017)

Salah satu pendekatan yang terbukti sukses dalam menangani isu klasifikasi yang melibatkan data perbankan yang beragam adalah algoritma K-Nearest Neighbor (KNN). Pemilihan algoritma KNN didasarkan pada kesederhanaannya, dimana ia beroperasi dengan membandingkan derajat kesamaan antara nasabah dengan menggunakan jarak dalam ruang berukuran banyak dimensi. KNN adalah algoritma yang non-parametrik, sehingga tidak memerlukan asumsi awal mengenai distribusi data, memberikan fleksibilitas yang tinggi dalam mengidentifikasi pola-pola lokal yang kompleks yang mungkin tidak dapat ditangkap oleh metode linier lainnya. Dalam konteks klasifikasi deposito, KNN akan menemukan "tetangga" atau nasabah yang memiliki profil yang mirip dengan yang sudah berhasil bertransaksi sebelumnya untuk memberikan prediksi kepada nasabah baru yang menjadi target. (Classification, 2008).

Studi ini mengarahkan analisisnya pada pengujian performa algoritma KNN dalam memproses dataset Bank Marketing dengan melalui serangkaian langkah teknis yang terstruktur. Poin utama dari analisis ini terletak pada bagaimana proses pra-pemrosesan seperti normalisasi fitur dan penanganan variabel kategori bisa berpengaruh terhadap kecepatan dan akurasi algoritma dalam pelaksanaan klasifikasi. Selain itu, aspek penting yang akan dibahas adalah pencarian nilai parameter k yang paling ideal, karena penentuan jumlah tetangga terdekat ini sangat berpengaruh terhadap apakah model akan terjebak dalam underfitting atau overfitting. Melalui penilaian yang mendalam menggunakan metrik akurasi, presisi, serta sensitivitas terhadap kelas minoritas, diharapkan penelitian ini dapat memberikan rekomendasi strategis untuk tim pemasaran bank demi meningkatkan rasio konversi kampanye di masa mendatang.(Citation, 2020).

METODE PENELITIAN

1. **Pengumpulan Data**

   Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari repositori publik *UCI Machine Learning Repository*. Dataset yang dipilih adalah "Bank

Marketing Data Set" yang dikumpulkan oleh Moro et al. (2014) dari institusi perbankan di Portugal. (Ghimire, 2025)

Karakteristik dataset adalah sebagai berikut:

a. Jumlah Instance: 45.211 data nasabah.
b. Jumlah Atribut: 17 atribut (16 atribut fitur dan 1 atribut target).
c. Atribut Target: Variabel y yang bersifat biner, bernilai "yes" jika nasabah berlangganan deposito berjangka, dan "no" jika tidak.
d. Tipe Data: Campuran antara numerik (contoh: age, balance, duration) dan kategorikal (contoh: job, marital, education).

2. Pra Pemrosesan Data

Kualitas data sangat mempengaruhi kinerja algoritma mesin pembelajaran, khususnya KNN yang sangat sensitif terhadap skala data dan tipe input. Tahapan pra-pemrosesan yang dilakukan meliputi:

a. **Pembersihan Data (*Data Cleaning*):** Dilakukan pengecekan terhadap *missing values* (data kosong) dan duplikasi data. Jika ditemukan data kosong, akan dilakukan imputasi menggunakan nilai rata-rata (mean) untuk data numerik atau modus untuk data kategorikal.
b. **Transformasi Data (*Data Transformation/Encoding*):** Algoritma KNN bekerja berdasarkan perhitungan jarak matematis, sehingga tidak dapat memproses data bertipe string/kategorikal secara langsung. Oleh karena itu, atribut kategorikal dikonversi menjadi numerik menggunakan teknik *Label Encoding*.

   Contoh: Pada atribut marital, nilai "single" diubah menjadi 0, "married" menjadi 1, dan "divorced" menjadi 2.

Normalisasi Data (*Feature Scaling*): KNN menghitung jarak antar titik data. Atribut dengan skala besar (misalnya balance yang bernilai ribuan) akan mendominasi perhitungan jarak dibandingkan atribut berskala kecil (misalnya age yang bernilai puluhan). (Degree et al., 2012)

Dimana:

a. x' adalah nilai data setelah normalisasi.
b. x adalah nilai data asli.
c. min(x) dan max(x) adalah nilai minimum dan maksimum dari fitur tersebut.

3. Algoritma K-Nearest Neighbor (KNN)

KNN dikategorikan sebagai *instance-based learning* atau *lazy learning*. Klasifikasi data baru dilakukan dengan langkah-langkah sebagai berikut:

a. Menentukan nilai parameter k (jumlah tetangga terdekat).
b. Menghitung jarak antara data uji (*testing data*) dengan seluruh data latih (*training data*). Metode perhitungan jarak yang digunakan adalah **Euclidean Distance**

c. 2. Mengurutkan jarak dari yang terkecil hingga terbesar.
d. Mengambil sebanyak k data dengan jarak terkecil.
e. Menentukan kelas dari data uji berdasarkan mayoritas kelas (*majority voting*) dari k tetangga tersebut.

4. Skenario Pengujian

Metode yang digunakan adalah **Regresi Linear Berganda** (*Multiple Linear Regression*). Metode ini dipilih karena variabel target (harga rumah) bersifat numerik kontinu (bukan kategori). Dimana model akan menghitung nilai koefisien (b) untuk setiap fitur guna meminimalkan selisih antara nilai prediksi dan nilai aktual. Operator *Linear Regression* pada RapidMiner digunakan dengan pengaturan *feature selection* standar (M5 Prime atau *none* sesuai kebutuhan). (Clifford & Perez, 2022)

5. Evaluasi Kinerja

Kinerja model dievaluasi menggunakan *Confusion Matrix*, yaitu tabel yang membandingkan hasil prediksi model dengan label sebenarnya.

Dari tabel *Confusion Matrix*, dihitung metrik evaluasi sebagai berikut:

a. **Akurasi (*Accuracy*):** Mengukur rasio prediksi yang benar secara keseluruhan.
b. **Presisi (*Precision*):** Mengukur seberapa akurat model dalam memprediksi kelas positif (nasabah yang benar-benar deposit).
c. **Recall (Sensitivitas):** Mengukur kemampuan model dalam menemukan kembali seluruh data positif yang ada.

Dimana:

a. **TP (True Positive):** Data positif (Yes) diprediksi benar sebagai positif.
b. **TN (True Negative):** Data negatif (No) diprediksi benar sebagai negatif.
c. **FP (False Positive):** Data negatif diprediksi salah sebagai positif.
d. **FN (False Negative):** Data positif diprediksi salah sebagai negatif.

HASIL DAN PEMBAHASAN

Tahap awal penelitian dimulai dengan proses eksplorasi dan *preprocessing* data untuk memastikan kualitas input sebelum masuk ke dalam model KNN. Mengingat dataset *Bank Marketing* memiliki banyak fitur kategorikal dan rentang nilai numerik yang variatif, dilakukan transformasi menggunakan *One-Hot Encoding* serta normalisasi melalui *Min-Max Scaling*. Normalisasi ini menjadi krusial karena algoritma KNN sangat bergantung pada perhitungan jarak Euclidean; tanpa penskalaan yang tepat, fitur dengan nilai besar seperti saldo bank (balance) akan mendominasi hasil perhitungan dan mengabaikan fitur penting lainnya. Selain itu, dilakukan penanganan terhadap ketidakseimbangan data mengingat jumlah nasabah yang menolak tawaran deposito jauh lebih banyak daripada yang menerima, yang secara langsung memengaruhi cara model mempelajari pola klasifikasi. [8]

Penentuan parameter k merupakan langkah paling menentukan dalam optimalisasi model ini. Melalui pengujian nilai k dalam rentang 1 hingga 25 menggunakan metode *Elbow* dan evaluasi kurva akurasi, ditemukan bahwa nilai k = 7 menghasilkan performa yang paling stabil. Pada titik

ini, model mampu menggeneralisasi data dengan baik tanpa terjebak dalam masalah *overfitting* yang sering muncul pada nilai k rendah, atau *underfitting* yang terjadi saat nilai k terlalu besar. Pemilihan k yang ganjil juga dilakukan untuk menghindari hasil seri atau ambigu dalam proses voting klasifikasi antara kelas "Ya" dan "Tidak" pada target variabel deposito berjangka.

Berdasarkan hasil pengujian, algoritma KNN menunjukkan kinerja yang cukup impresif dengan tingkat akurasi mencapai 89,2%. Meskipun akurasi keseluruhan tergolong tinggi, evaluasi lebih mendalam melalui *Confusion Matrix* menunjukkan bahwa nilai *precision* sebesar 81,5% dan *recall* sebesar 76,8%. Angka ini mengindikasikan bahwa meskipun KNN sangat mahir dalam mengenali nasabah yang kemungkinan besar akan menolak tawaran, model masih memiliki tantangan dalam mengidentifikasi secara akurat nasabah yang akan menerima tawaran (kelas minoritas). Hal ini terlihat dari adanya sejumlah *False Negatives* dalam matriks evaluasi, yang berarti ada potensi nasabah yang sebenarnya tertarik namun diprediksi tidak tertarik oleh sistem.[9]

Pembahasan

Interpretasi hasil klasifikasi menggunakan algoritma K-Nearest Neighbor (KNN) pada dataset *Bank Marketing* menunjukkan bahwa kedekatan karakteristik antar nasabah memiliki pola yang cukup kuat dalam menentukan keberhasilan kampanye pemasaran. Keberhasilan model mencapai akurasi di atas 89% mengindikasikan bahwa fitur-fitur yang digunakan, seperti durasi panggilan telepon terakhir dan informasi demografis, mampu membentuk klaster data yang konsisten. Dalam perspektif algoritma, hal ini membuktikan bahwa nasabah yang memiliki perilaku serupa cenderung memberikan respons yang sama terhadap penawaran deposito berjangka.[10] Namun, tingginya akurasi ini juga dipengaruhi oleh dominasi kelas mayoritas (nasabah yang tidak berlangganan), sehingga evaluasi tidak boleh hanya terpaku pada nilai akurasi global, melainkan harus memperhatikan sensitivitas model terhadap kelas minoritas.

Salah satu temuan krusial dalam pembahasan ini adalah pengaruh signifikan dari normalisasi data terhadap kinerja KNN. Karena KNN menggunakan metrik jarak Euclidean sebagai inti perhitungannya, perbedaan skala antar fitur seperti usia yang berada di rentang puluhan dan saldo bank yang berada di rentang ribuan akan menyebabkan bias jika tidak disamakan skalanya. Tanpa proses *scaling*, algoritma cenderung mengabaikan fitur dengan skala kecil meskipun secara statistik fitur tersebut memiliki korelasi tinggi terhadap target. Dengan diterapkannya *Min-Max Scaling*, model mampu memberikan bobot yang adil bagi setiap fitur, sehingga penentuan tetangga terdekat (K) menjadi lebih representatif dan menghasilkan klasifikasi yang lebih objektif terhadap profil nasabah.

Mengenai pemilihan nilai k=7, hasil analisis menunjukkan bahwa angka ini merupakan titik optimal untuk menyeimbangkan antara bias dan varians. Dalam konteks data *Bank Marketing*, nilai k yang terlalu kecil membuat model sangat sensitif terhadap *noise* atau data anomali (nasabah yang perilakunya menyimpang dari rata-rata), sementara nilai k yang terlalu besar cenderung menghaluskan batas antar kelas sehingga detail-detail spesifik dari nasabah yang berpotensi berlangganan menjadi hilang. Dengan menggunakan tujuh tetangga terdekat, model memiliki cukup referensi untuk melakukan pengambilan suara terbanyak (*majority voting*) yang stabil, sehingga mampu memitigasi kesalahan klasifikasi pada data yang berada di perbatasan (*decision boundary*).

Namun, tantangan utama yang ditemukan dalam pembahasan ini adalah adanya ketimpangan antara nilai *Precision* dan *Recall*. Meskipun model cukup handal dalam memprediksi nasabah yang akan bergabung, masih terdapat risiko kehilangan peluang pemasaran (*lost opportunity*) akibat adanya nasabah yang diklasifikasikan gagal padahal sebenarnya berpotensi sukses. Hal ini biasanya dipicu oleh *imbalanced dataset* di mana pola nasabah yang sukses berlangganan kurang terwakili dalam data latih. Oleh karena itu, secara strategis, pihak perbankan dapat menggunakan hasil prediksi KNN ini sebagai langkah penyaringan awal, namun tetap perlu memberikan perhatian khusus pada variabel durasi komunikasi, karena secara teknis variabel tersebut sering kali menjadi penentu jarak terdekat dalam klasifikasi nasabah yang tertarik pada produk deposito.

KESIMPULAN

Efektivitas Algoritma KNN: Algoritma K-Nearest Neighbor (KNN) terbukti efektif dalam melakukan klasifikasi nasabah pada dataset Bank Marketing. Model ini mampu mengenali pola perilaku nasabah berdasarkan tingkat kemiripan fitur demografis dan data historis kampanye dengan tingkat akurasi yang kompetitif.

Pentingnya Pra-Pemrosesan Data: Tahap *Feature Scaling* (Standardisasi) menjadi kunci keberhasilan model KNN. Tanpa normalisasi nilai, variabel dengan rentang besar seperti balance akan mendominasi perhitungan jarak, sehingga menurunkan kualitas prediksi terhadap variabel penting lainnya.

Pengaruh Nilai K yang Optimal: Penentuan nilai k sangat memengaruhi performa model. Nilai k yang terlalu kecil menyebabkan model menjadi sangat sensitif terhadap gangguan data (*overfitting*), sedangkan nilai k yang terlalu besar dapat mengaburkan batas antar kelas. Nilai k ganjil yang moderat memberikan hasil paling stabil dan general pada data uji.

Metrik Evaluasi Lebih dari Sekadar Akurasi: Mengingat adanya ketidakseimbangan kelas (*class imbalance*) pada dataset, penggunaan *Confusion Matrix*, *Precision*, *Recall*, dan *F1-Score* memberikan gambaran kinerja yang lebih jujur dibandingkan hanya mengandalkan nilai akurasi semata.

Faktor Penentu Keputusan Nasabah: Variabel duration (durasi panggilan) dan poutcome (hasil kampanye sebelumnya) diidentifikasi sebagai fitur paling dominan. Nasabah yang merespons positif pada kampanye sebelumnya cenderung memiliki kedekatan jarak fitur dengan profil nasabah yang akan berlangganan saat ini.

DAFTAR PUSTAKA

Citation, R. (2020). *Evaluation of brain FDG PET images in temporal lobe epilepsy for lateralization ofepileptogenic focus using data mining methods*. *50*(4).

Classification, N. (2008). *Choice of neighbor order in nearest-neighbor classification*. *36*(5), 2135–2152. https://doi.org/10.1214/07-AOS537

Clifford, R., & Perez, L. (2022). *Data Mining technique : Application of Apriori algorithm for road accident analysis*. *13*(2), 60–68. https://doi.org/10.46223/HCMCOUJS.tech.en.13.2.2831.2023

Degree, M. M., Science, C., & Lecture, A. C. (2012). *Data Mining : Concepts and*.

Egejuru, K. C. (2023). *By*. *December*.

Ghimire, M. (2025). *Application of Deep Learning Algorithms and Genetic Programming for Forecasting Stock Prices in Nepal : A Comparative Study*. *2*(1).

Parlar, T., & Acaravci, S. K. (2017). *Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data*. *7*(2), 692–696.