

ANALISIS PREDIKSI KEPUTUSAN NASABAH DALAM BERLANGGANAN DEPOSITO BERJANGKA MENGGUNAKAN METODE MACHINE LEARNING PADA DATASET BANK MARKETING

*ANALYSIS OF CUSTOMER DECISION PREDICTIONS IN SUBSCRIBING TERM
DEPOSITS USING THE MACHINE LEARNING METHOD ON THE BANK MARKETING
DATASET*

Salom Indriyani Putri Saragih¹, Yelsi Herma Nova Ayu Safitri², Wahyu Dwi Lesmono³

¹⁻³Institut Teknologi Kalimantan, Fakultas Sains dan Teknologi Informasi, Jurusan Sains dan Analitika Data,
Ilmu Aktuaria

***Email Correspondence: 17231041@student.itk.ac.id.**

Abstract

Advances in digital technology have increased the availability of data in the banking sector, which can be leveraged to support data-driven decision-making. This study aims to compare the performance of machine learning algorithms in predicting customers' decisions to open time deposits using the Bank Marketing Dataset, which consists of 41,188 observations and 21 variables. The research stages included data cleaning, categorical data transformation, data standardization, splitting the data into training and testing sets, and addressing class imbalance using the Synthetic Minority Oversampling Technique. The algorithms used were Decision Tree, K-Nearest Neighbor, Logistic Regression, and Random Forest. The results show that Random Forest produced the best classification performance compared to the other models. Additionally, communication duration was the most influential factor affecting customers' decisions to subscribe to time deposits. Thus, Random Forest can be used as an effective approach to support data-driven marketing strategies in the banking sector.

Keywords: Bank Marketing Dataset, Classification, Machine Learning, Random Forest, Synthetic Minority Oversampling Technique.

Abstrak

Perkembangan teknologi digital meningkatkan ketersediaan data pada sektor perbankan yang dapat dimanfaatkan untuk mendukung pengambilan keputusan berbasis data. Penelitian ini bertujuan membandingkan performa algoritma *machine learning* dalam memprediksi keputusan nasabah untuk berlangganan deposito berjangka menggunakan Bank Marketing Dataset yang terdiri atas 41.188 observasi dan 21 variabel. Tahapan penelitian meliputi pembersihan data, transformasi data kategorik, standarisasi data, pembagian data pelatihan dan pengujian, serta penanganan ketidakseimbangan kelas menggunakan *Synthetic Minority Oversampling Technique*. Algoritma yang digunakan yaitu *Decision Tree*, *K-Nearest Neighbor*, *Logistic Regression*, dan *Random Forest*. Hasil penelitian menunjukkan bahwa *Random Forest* menghasilkan performa klasifikasi terbaik dibandingkan model lainnya. Selain itu, durasi komunikasi menjadi faktor yang paling berpengaruh terhadap keputusan nasabah dalam berlangganan deposito. Dengan demikian, *Random Forest* dapat digunakan sebagai pendekatan yang efektif dalam mendukung strategi pemasaran berbasis data pada sektor perbankan.

Kata kunci: Dataset Pemasaran Bank, Klasifikasi, Pembelajaran Mesin, Random Forest, Teknik Oversampling Minoritas Sintetis.

PENDAHULUAN

Perkembangan teknologi digital telah menyebabkan peningkatan jumlah data yang tersedia pada berbagai sektor, termasuk sektor perbankan. Data yang dihasilkan dari aktivitas transaksi, karakteristik nasabah, serta interaksi pemasaran dapat dimanfaatkan untuk memperoleh informasi yang mendukung pengambilan keputusan. Pengolahan data dalam jumlah besar membutuhkan pendekatan analisis yang mampu menemukan pola dan hubungan antar variabel secara lebih efektif. Oleh karena itu, penerapan teknologi berbasis data menjadi salah satu strategi penting dalam meningkatkan kualitas keputusan bisnis pada industri perbankan (Suryantari dkk, 2026).

Salah satu permasalahan yang dapat diselesaikan menggunakan machine learning dalam sektor perbankan adalah prediksi keputusan nasabah dalam menggunakan produk deposito berjangka. Deposito merupakan salah satu produk perbankan yang berperan dalam penghimpunan dana masyarakat sehingga bank perlu mengetahui karakteristik calon nasabah yang berpotensi menggunakan produk tersebut. Dataset Bank Marketing sering digunakan dalam penelitian klasifikasi karena menyediakan informasi mengenai karakteristik sosial, ekonomi, serta riwayat komunikasi pemasaran nasabah. Variabel-variabel tersebut dapat digunakan untuk membangun model prediksi terhadap keputusan nasabah dalam menerima atau menolak penawaran deposito berjangka.

Dalam proses klasifikasi menggunakan dataset Bank Marketing, terdapat permasalahan berupa ketidakseimbangan distribusi kelas (class imbalance) antara nasabah yang berlangganan deposito dan nasabah yang tidak berlangganan. Kondisi tersebut dapat menyebabkan model lebih banyak mempelajari pola dari kelas mayoritas sehingga kemampuan dalam mendeteksi kelas minoritas menjadi kurang optimal. Apabila masalah ini tidak ditangani, hasil prediksi model dapat menjadi bias dan kurang menggambarkan kondisi sebenarnya. Oleh sebab itu, diperlukan metode penanganan ketidakseimbangan data agar model klasifikasi dapat menghasilkan performa yang lebih baik (Erlangga dkk, 2024).

Salah satu teknik yang dapat digunakan untuk mengatasi permasalahan class imbalance adalah Synthetic Minority Oversampling Technique (SMOTE). SMOTE bekerja dengan menghasilkan data sintetis baru pada kelas minoritas berdasarkan karakteristik data yang sudah tersedia. Penerapan metode ini dapat meningkatkan jumlah observasi pada kelas minoritas sehingga distribusi data menjadi lebih seimbang sebelum dilakukan proses pemodelan. Dengan distribusi data yang lebih baik, algoritma machine learning diharapkan mampu melakukan proses klasifikasi secara lebih akurat dan stabil (Ardhana dkk, 2023).

Berbagai algoritma machine learning dapat diterapkan untuk melakukan prediksi keputusan nasabah, di antaranya Decision Tree, Random Forest, Logistic Regression, dan K-Nearest Neighbor (KNN). Decision Tree memiliki kemampuan dalam menghasilkan aturan keputusan yang mudah dipahami, sedangkan Random Forest mampu meningkatkan kemampuan prediksi melalui kombinasi beberapa pohon keputusan. Logistic Regression dapat digunakan untuk permasalahan klasifikasi biner karena mampu menggambarkan hubungan antara variabel independen dan variabel target. Sementara itu, KNN melakukan proses klasifikasi berdasarkan tingkat kedekatan karakteristik antar data sehingga dapat digunakan untuk mengenali pola berdasarkan kemiripan data (Ardhana dkk, 2023).

Evaluasi performa model merupakan tahap penting dalam menentukan kualitas algoritma klasifikasi yang digunakan. Pengukuran performa tidak hanya menggunakan nilai akurasi, tetapi juga mempertimbangkan metrik lain seperti precision, recall, F1-score, dan ROC-AUC. Penggunaan beberapa metrik evaluasi diperlukan karena dataset dengan distribusi kelas yang tidak seimbang dapat menghasilkan nilai akurasi yang tinggi

meskipun kemampuan model dalam mengenali kelas minoritas masih rendah. Oleh karena itu, evaluasi yang menyeluruh diperlukan untuk mengetahui algoritma yang memiliki kemampuan prediksi terbaik (Fauziah & Hernadi, 2024).

Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk menerapkan dan membandingkan beberapa algoritma machine learning dalam memprediksi keputusan nasabah untuk berlangganan deposito berjangka menggunakan dataset Bank Marketing. Tahapan penelitian meliputi proses eksplorasi data, preprocessing, penanganan ketidakseimbangan kelas menggunakan SMOTE, pembangunan model klasifikasi, serta evaluasi performa model. Perbandingan beberapa algoritma dilakukan untuk mengetahui metode yang mampu memberikan hasil prediksi terbaik berdasarkan metrik evaluasi yang digunakan. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pemanfaatan machine learning sebagai pendukung strategi pemasaran berbasis data pada sektor perbankan (Erin dkk, 2022).

TINJAUAN PUSTAKA

Machine Learning dalam Prediksi Keputusan Nasabah

Machine learning merupakan cabang dari artificial intelligence yang memungkinkan sistem komputer mempelajari pola dari data historis untuk melakukan prediksi atau pengambilan keputusan tanpa diprogram secara eksplisit pada setiap kondisi tertentu (Sarker, 2021). Perkembangan teknologi artificial intelligence dan deep learning memberikan kontribusi besar dalam kemampuan sistem komputer untuk memproses data berukuran besar dan mengenali pola yang kompleks pada berbagai bidang (Alzubaidi dkk, 2021). Berbagai metode supervised learning seperti Decision Tree, Random Forest, Logistic Regression, dan K-Nearest Neighbor (KNN) sering digunakan dalam permasalahan klasifikasi karena memiliki pendekatan dan kemampuan prediksi yang berbeda-beda sesuai dengan karakteristik data yang dianalisis (Chauhan dkk, 2023).

Dataset Bank Marketing dan Penanganan Ketidakseimbangan Data

Prediksi keputusan nasabah dalam menerima atau menolak penawaran deposito berjangka merupakan salah satu bentuk permasalahan klasifikasi biner. Dataset Bank Marketing menyediakan informasi karakteristik nasabah, kondisi ekonomi, serta riwayat kampanye pemasaran yang dapat digunakan untuk membangun model prediksi keputusan nasabah terhadap produk deposito. Penerapan algoritma machine learning pada dataset tersebut dapat membantu mengidentifikasi nasabah yang berpotensi menerima penawaran deposito sehingga mendukung efektivitas strategi pemasaran bank (Du dkk, 2025).

Pada proses klasifikasi menggunakan dataset Bank Marketing sering ditemukan permasalahan ketidakseimbangan kelas (class imbalance), yaitu jumlah nasabah yang tidak berlangganan deposito lebih banyak dibandingkan dengan nasabah yang berlangganan. Kondisi tersebut dapat menyebabkan model lebih dominan mempelajari pola dari kelas mayoritas sehingga kemampuan dalam mendeteksi kelas minoritas menjadi menurun. Oleh karena itu, teknik Synthetic Minority Oversampling Technique (SMOTE) digunakan untuk menghasilkan data sintesis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang dan performa model klasifikasi dapat meningkat (Abidin dkk, 2025).

Algoritma Klasifikasi

Decision Tree merupakan algoritma klasifikasi yang membentuk struktur menyerupai pohon melalui proses pembagian data berdasarkan fitur yang menghasilkan kelompok dengan tingkat homogenitas paling tinggi (Chauhan dkk, 2023). Kelebihan utama

Decision Tree adalah mampu menghasilkan aturan keputusan yang mudah dipahami dan diinterpretasikan sehingga sesuai digunakan pada bidang seperti perbankan yang membutuhkan penjelasan mengenai faktor-faktor yang mempengaruhi keputusan nasabah. Namun, Decision Tree memiliki kelemahan berupa kecenderungan mengalami overfitting apabila struktur pohon yang terbentuk terlalu kompleks sehingga diperlukan pengendalian kompleksitas model untuk memperoleh hasil klasifikasi yang lebih stabil (Chauhan dkk, 2023).

Logistic Regression merupakan metode klasifikasi yang digunakan untuk memprediksi probabilitas suatu observasi termasuk ke dalam kelas tertentu berdasarkan hubungan antara variabel prediktor dengan variabel target (Khanam & Foo, 2021). Logistic Regression sering digunakan sebagai model pembanding karena memiliki struktur yang sederhana, mudah diimplementasikan, serta mampu memberikan performa yang baik pada permasalahan klasifikasi biner. Perbandingan beberapa algoritma klasifikasi diperlukan karena setiap metode memiliki karakteristik serta performa yang berbeda bergantung pada jenis dan karakteristik data yang digunakan (Kurniawan & Megawaty, 2025).

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi yang menentukan kelas suatu data berdasarkan mayoritas kelas dari sejumlah data terdekat menggunakan ukuran jarak tertentu. Algoritma ini bekerja berdasarkan prinsip kemiripan antar data, yaitu data dengan karakteristik yang serupa cenderung berada pada kelas yang sama. KNN digunakan sebagai model pembanding karena memiliki pendekatan yang berbeda dengan Decision Tree sehingga hasil perbandingan dapat menunjukkan metode yang paling sesuai dengan karakteristik dataset Bank Marketing (Khanam & Foo, 2021).

Random Forest merupakan metode ensemble learning yang dibangun dari kumpulan beberapa Decision Tree yang dilatih menggunakan sampel data dan pemilihan fitur secara acak. Metode ini dikembangkan untuk mengatasi kelemahan Decision Tree tunggal, terutama masalah overfitting dan ketidakstabilan model. Dengan menggabungkan banyak pohon keputusan, Random Forest mampu meningkatkan kemampuan generalisasi dan menghasilkan performa prediksi yang lebih stabil dibandingkan Decision Tree tunggal. Oleh karena itu, perbandingan antara Decision Tree dan Random Forest dilakukan untuk mengetahui apakah penggunaan banyak pohon keputusan dapat memberikan hasil klasifikasi yang lebih optimal (Chauhan dkk, 2023).

Evaluasi Performa Model Klasifikasi

Evaluasi model klasifikasi dilakukan untuk mengetahui kemampuan model dalam memberikan prediksi yang tepat terhadap data baru. Pengukuran performa model tidak hanya menggunakan nilai accuracy, tetapi juga mempertimbangkan precision, recall, F1-score, dan ROC-AUC agar kemampuan model dalam mengenali setiap kelas dapat dievaluasi secara lebih menyeluruh, terutama pada data dengan kondisi ketidakseimbangan kelas (Chauhan dkk, 2023; Kurniawan & Megawaty, 2025).

Penggunaan beberapa metrik evaluasi diperlukan karena nilai accuracy yang tinggi belum tentu menunjukkan kemampuan model yang baik dalam mendeteksi kelas minoritas. Oleh sebab itu, precision, recall, dan F1-score menjadi metrik tambahan yang penting untuk memberikan penilaian performa model yang lebih komprehensif (Abidin dkk, 2025).

METODE

Penelitian ini menggunakan data sekunder berupa Bank Marketing Dataset yang diperoleh dari UCI Machine Learning Repository. Dataset tersebut berisi informasi

karakteristik nasabah, aktivitas kampanye pemasaran, serta indikator ekonomi yang digunakan untuk memprediksi keputusan nasabah dalam berlangganan deposito berjangka. Dataset awal terdiri atas 41.188 observasi dengan 21 variabel, yang meliputi 20 variabel prediktor dan 1 variabel target (y). Variabel target memiliki dua kategori, yaitu yes yang menunjukkan nasabah berlangganan deposito dan no yang menunjukkan nasabah tidak berlangganan deposito.

Tahap awal penelitian dilakukan melalui proses eksplorasi dan pembersihan data (data cleaning). Hasil pemeriksaan menunjukkan bahwa tidak terdapat missing value pada seluruh variabel. Namun, ditemukan sebanyak 12 data duplikat sehingga data tersebut dihapus untuk menghindari bias dalam proses pemodelan. Setelah proses pembersihan data, jumlah observasi yang digunakan menjadi 41.176 data.

Selanjutnya dilakukan tahap prapemrosesan data (preprocessing). Variabel kategorik diubah ke dalam bentuk numerik menggunakan metode Label Encoding agar dapat diproses oleh algoritma machine learning. Setelah itu, dilakukan standardisasi pada variabel numerik menggunakan metode Standard Scaler. Proses standardisasi bertujuan untuk menyamakan skala antarvariabel sehingga tidak terjadi dominasi variabel tertentu dalam proses pembelajaran model. Standardisasi dilakukan menggunakan persamaan:

$$Z = \frac{X - \mu}{\sigma}$$

dengan X merupakan nilai observasi, μ adalah nilai rata-rata, dan σ adalah standar deviasi.

Data kemudian dibagi menjadi data pelatihan (training set) dan data pengujian (testing set) dengan perbandingan 80:20 menggunakan metode stratified sampling. Teknik ini digunakan untuk menjaga proporsi kelas pada data pelatihan dan data pengujian tetap sama dengan distribusi data asli. Berdasarkan proses tersebut diperoleh sebanyak 32.940 data pelatihan dan 8.236 data pengujian.

Distribusi kelas pada variabel target menunjukkan kondisi tidak seimbang (imbalanced data), dimana jumlah kelas no jauh lebih besar dibandingkan kelas yes. Oleh karena itu, dilakukan penanganan ketidakseimbangan data menggunakan metode Synthetic Minority Oversampling Technique (SMOTE). Metode SMOTE bekerja dengan membangkitkan data sintetis pada kelas minoritas berdasarkan kedekatan antara observasi sehingga distribusi data menjadi lebih seimbang. Sebelum penerapan SMOTE, jumlah kelas mayoritas dan minoritas pada data pelatihan masing-masing sebesar 29.229 dan 3.711 observasi. Setelah penerapan SMOTE, jumlah kedua kelas menjadi seimbang, yaitu masing-masing sebanyak 29.229 observasi.

Penelitian ini menggunakan empat algoritma klasifikasi, yaitu Decision Tree, K-Nearest Neighbor (KNN), Logistic Regression, dan Random Forest. Decision Tree digunakan untuk membentuk aturan keputusan berbentuk struktur pohon berdasarkan pemisahan data yang menghasilkan tingkat impuritas minimum. KNN melakukan klasifikasi berdasarkan kedekatan suatu observasi terhadap sejumlah tetangga terdekat menggunakan jarak Euclidean. Logistic Regression digunakan untuk memodelkan probabilitas suatu observasi masuk ke dalam kelas tertentu melalui fungsi logistik. Sementara itu, Random Forest merupakan metode ensemble learning yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih stabil dan akurat.

Untuk memperoleh parameter terbaik, dilakukan proses hyperparameter tuning menggunakan Grid Search Cross Validation. Pada model Decision Tree dilakukan pencarian parameter max_depth, sedangkan pada model KNN dilakukan pencarian nilai n_neighbors terbaik. Pada model Random Forest dilakukan optimasi parameter n_estimators, max_depth, min_samples_split, dan min_samples_leaf.

Kinerja model dievaluasi menggunakan confusion matrix dan beberapa metrik

evaluasi, yaitu accuracy, precision, recall, F1-score, serta Receiver Operating Characteristic Area Under Curve (ROC-AUC). Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data. Precision digunakan untuk mengukur ketepatan model dalam mengidentifikasi kelas positif. Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data positif. F1-score merupakan rata-rata harmonik antara precision dan recall, sedangkan ROC-AUC digunakan untuk mengukur kemampuan model dalam membedakan kelas positif dan negatif. Model terbaik ditentukan berdasarkan nilai accuracy, precision, recall, F1-score, dan ROC-AUC yang paling optimal.

HASIL DAN PEMBAHASAN

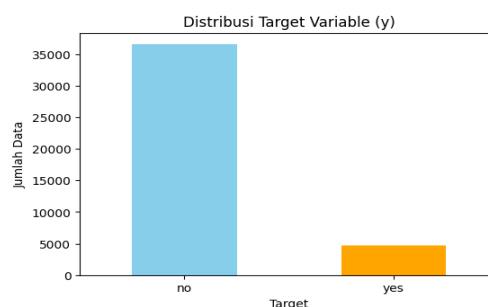
Karakteristik Dataset

Penelitian ini menggunakan Bank Marketing Dataset yang diperoleh dari UCI Machine Learning Repository. Dataset tersebut terdiri atas 41.188 observasi dengan 21 variabel yang mencakup karakteristik nasabah, informasi kampanye pemasaran, dan indikator ekonomi. Sebelum dilakukan pemodelan, data terlebih dahulu diperiksa untuk mengetahui kualitas data yang digunakan. Hasil pemeriksaan menunjukkan bahwa dataset memiliki kualitas yang cukup baik karena tidak ditemukan missing value pada seluruh variabel. Namun demikian, ditemukan sebanyak 12 data duplikat yang kemudian dihapus agar tidak memengaruhi hasil analisis.

Tabel 1. Karakteristik Dataset

Keterangan	Nilai
Jumlah variabel	21
<i>Missing value</i>	0
Data duplikat	12
Kelas <i>no</i>	36.548
Kelas <i>yes</i>	4.640

Berdasarkan Tabel 1, jumlah data yang digunakan setelah proses data cleaning adalah sebanyak 41.176 observasi. Penghapusan data duplikat dilakukan untuk menghindari pengaruh data yang berulang terhadap proses pembelajaran model. Selain itu, variabel target menunjukkan distribusi yang tidak seimbang antara kelas *yes* dan *no*. Ketidakseimbangan ini menjadi salah satu tantangan utama dalam penelitian karena dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas dibandingkan kelas minoritas.



Gambar 1. Distribusi Variabel Target

Berdasarkan Gambar 1, jumlah nasabah yang tidak berlangganan deposito (no) jauh lebih banyak dibandingkan nasabah yang berlangganan deposito (yes). Jumlah kelas no mencapai 36.548 observasi, sedangkan kelas yes hanya sebanyak 4.640 observasi. Perbedaan jumlah yang cukup besar ini menunjukkan adanya permasalahan class imbalance pada dataset. Kondisi tersebut berpotensi menyebabkan model klasifikasi menghasilkan bias terhadap kelas mayoritas sehingga diperlukan metode penyeimbangan data sebelum proses pemodelan dilakukan.

Hasil Preprocessing Data

Tahap preprocessing dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses klasifikasi. Proses yang dilakukan meliputi label encoding, feature scaling, pembagian data pelatihan dan pengujian, serta penanganan ketidakseimbangan data menggunakan SMOTE. Label encoding digunakan untuk mengubah variabel kategorik menjadi bentuk numerik agar dapat diproses oleh algoritma machine learning. Selanjutnya dilakukan standarisasi data menggunakan StandardScaler untuk menyamakan skala antarvariabel numerik. Tahapan ini penting terutama pada algoritma KNN yang sangat sensitif terhadap perbedaan skala data.

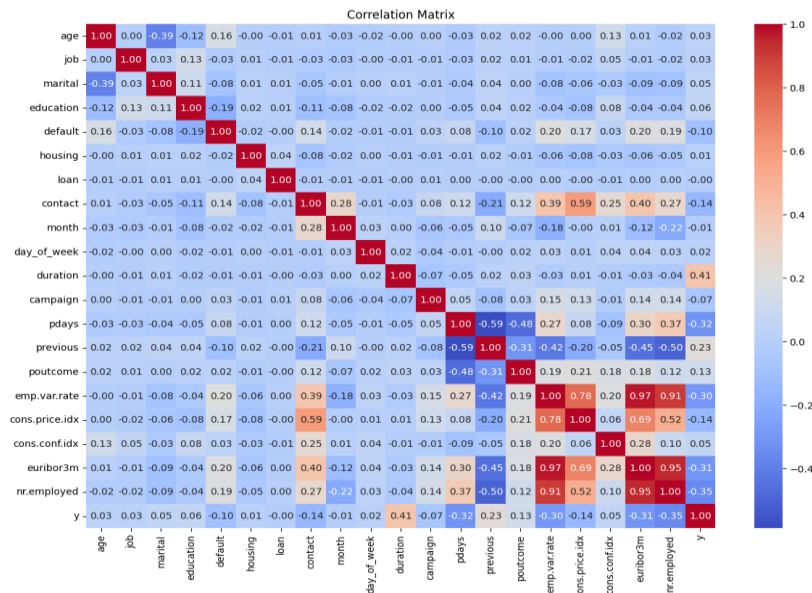
Tabel 2. Distribusi Kelas Sebelum dan Sesudah SMOTE

Kelas	Sebelum SMOTE	Sesudah SMOTE
No	29.229	29.229
Yes	3.711	29.229

Berdasarkan Tabel 2, terlihat bahwa jumlah kelas minoritas sebelum SMOTE hanya sebanyak 3.711 observasi, sedangkan kelas mayoritas berjumlah 29.229 observasi. Setelah penerapan SMOTE, jumlah kedua kelas menjadi sama yaitu masing-masing sebanyak 29.229 observasi. Hasil tersebut menunjukkan bahwa metode SMOTE berhasil mengatasi ketidakseimbangan data pada tahap pelatihan model. Dengan distribusi yang lebih seimbang, model diharapkan mampu mengenali pola pada kelas minoritas dengan lebih baik sehingga meningkatkan kualitas prediksi.

Analisis Korelasi Antar Variabel

Analisis korelasi dilakukan untuk mengetahui hubungan antarvariabel dalam dataset. Hubungan tersebut dapat memberikan gambaran awal mengenai variabel yang berpotensi memengaruhi keputusan nasabah dalam berlangganan deposito. Korelasi dihitung menggunakan matriks korelasi yang divisualisasikan dalam bentuk heatmap. Nilai korelasi berkisar antara -1 hingga 1, dimana nilai positif menunjukkan hubungan searah dan nilai negatif menunjukkan hubungan berlawanan arah.

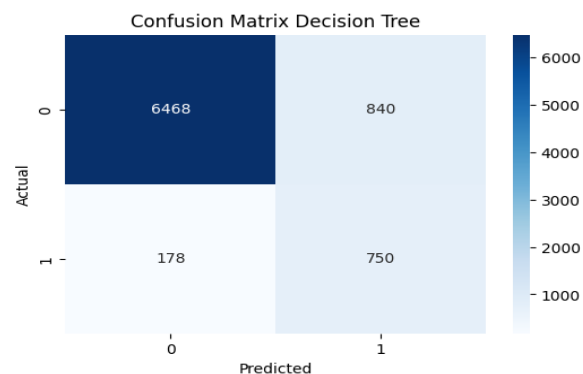


Gambar 2. Correlation Heatmap

Berdasarkan Gambar 2, variabel duration memiliki korelasi positif terbesar terhadap variabel target dengan nilai sekitar 0,41. Hasil tersebut menunjukkan bahwa semakin lama durasi komunikasi antara pihak bank dan nasabah, maka peluang nasabah untuk berlangganan deposito cenderung meningkat. Selain itu, variabel nr.employed memiliki korelasi negatif sebesar -0,35, sedangkan variabel pdays dan euribor3m memiliki korelasi negatif masing-masing sebesar -0,32 dan -0,31. Temuan ini mengindikasikan bahwa faktor ekonomi dan riwayat kontak sebelumnya juga berkontribusi terhadap keputusan nasabah.

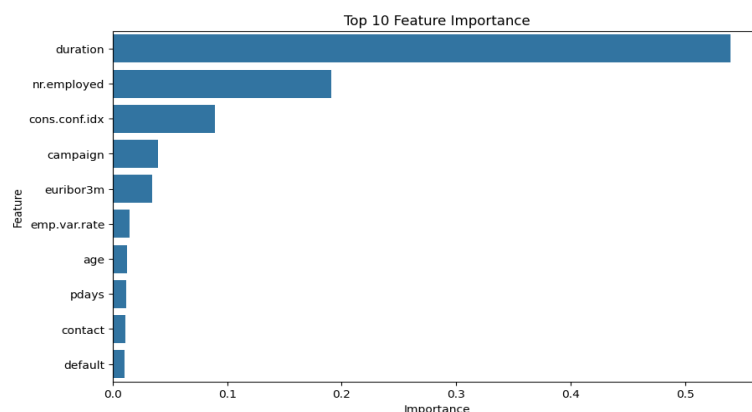
Hasil Pemodelan Decision Tree

Model Decision Tree dibangun menggunakan data hasil preprocessing dan penyeimbangan kelas dengan SMOTE. Algoritma ini bekerja dengan membentuk struktur pohon keputusan berdasarkan variabel yang memiliki kemampuan terbaik dalam memisahkan kelas target. Evaluasi model dilakukan menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Hasil evaluasi menunjukkan bahwa model mampu memberikan performa yang cukup baik dalam mengidentifikasi nasabah yang berpotensi berlangganan deposito.



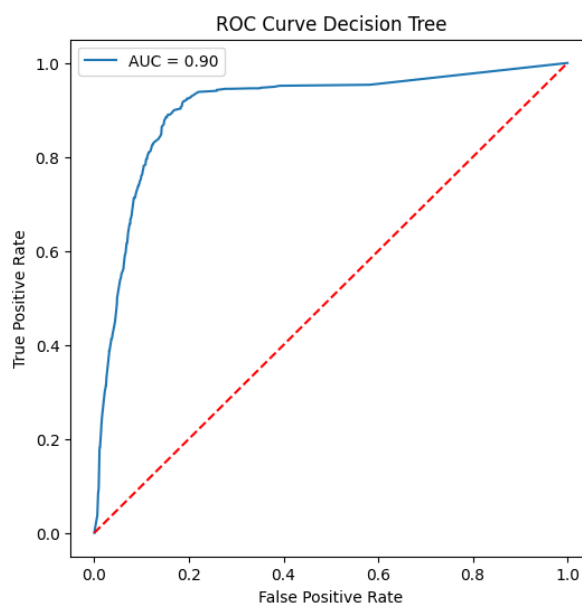
Gambar 3. Confusion Matrix Decision Tree

Berdasarkan Gambar 3, hasil evaluasi, model Decision Tree memperoleh accuracy sebesar 86,12%, precision sebesar 44,10%, recall sebesar 86,53%, dan F1-score sebesar 58,42%. Nilai recall yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar nasabah yang benar-benar berlangganan deposito. Namun demikian, nilai precision yang relatif rendah menunjukkan masih terdapat sejumlah prediksi positif yang tidak sesuai dengan kondisi sebenarnya. Kondisi tersebut menunjukkan bahwa model lebih fokus dalam menangkap kelas positif meskipun harus mengorbankan ketepatan prediksi.



Gambar 4. Feature Importance Decision Tree

Berdasarkan Gambar 4, variabel duration merupakan variabel yang paling berpengaruh dengan nilai importance sebesar 0,6117. Variabel berikutnya yang memiliki kontribusi besar adalah nr.employed sebesar 0,2250 dan cons.conf.idx sebesar 0,1037. Dominasi variabel duration menunjukkan bahwa lama komunikasi antara pihak bank dan nasabah menjadi faktor utama dalam menentukan keberhasilan penawaran deposito. Hasil ini juga sejalan dengan analisis korelasi yang menunjukkan bahwa duration memiliki hubungan positif terbesar terhadap variabel target.

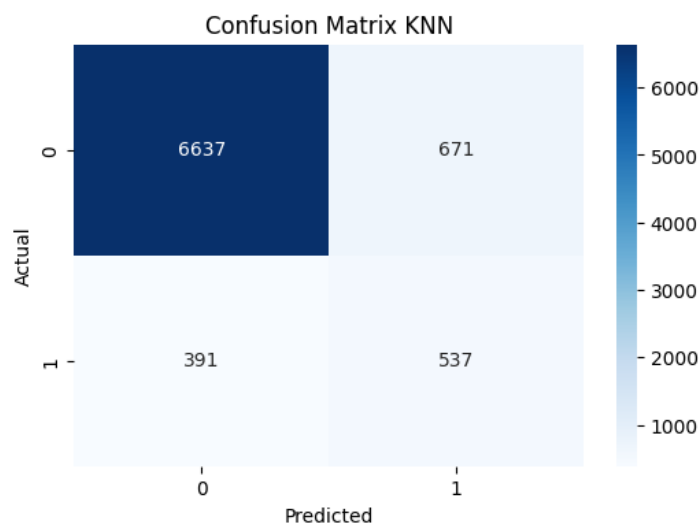


Gambar 5. ROC Curve Decision Tree

Berdasarkan Gambar 5, nilai ROC-AUC Decision Tree sebesar 0,9264 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan nasabah yang berlangganan dan tidak berlangganan deposito. Nilai tersebut menunjukkan bahwa Decision Tree mampu melakukan klasifikasi dengan tingkat diskriminasi yang tinggi. Selain itu, perbandingan nilai train accuracy sebesar 92,68% dan test accuracy sebesar 87,64% menunjukkan adanya selisih sekitar 5,04%, sehingga model memiliki sedikit indikasi overfitting, namun masih memiliki kemampuan generalisasi yang cukup baik terhadap data baru.

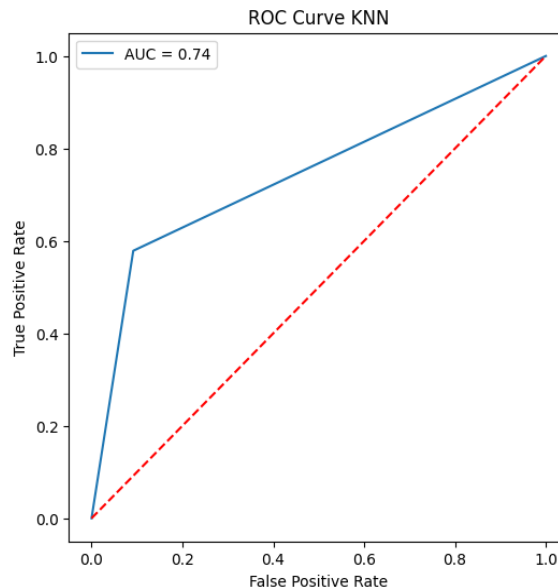
Hasil Pemodelan K-Nearest Neighbor (KNN)

Metode K-Nearest Neighbor (KNN) digunakan untuk mengklasifikasikan data berdasarkan kedekatan jarak antarobservasi. Pada penelitian ini, data terlebih dahulu distandarisasi menggunakan StandardScaler karena KNN sangat sensitif terhadap perbedaan skala data. Selain itu, dilakukan proses hyperparameter tuning untuk menentukan nilai k terbaik yang dapat menghasilkan performa optimal. Hasil tuning menunjukkan bahwa nilai k terbaik adalah 1 dengan nilai rata-rata akurasi validasi silang sebesar 94,05%.



Gambar 6. Confusion Matrix KNN

Berdasarkan Gambar 6, hasil evaluasi, model KNN memperoleh accuracy sebesar 84,80%, precision sebesar 41,45%, recall sebesar 84,59%, dan F1-score sebesar 55,63%. Nilai recall yang cukup tinggi menunjukkan bahwa model mampu mengenali sebagian besar nasabah yang benar-benar berlangganan deposito. Namun demikian, nilai precision yang relatif rendah mengindikasikan masih terdapat banyak kesalahan prediksi pada kelas positif. Selain itu, nilai F1-score yang lebih rendah dibandingkan model lainnya menunjukkan bahwa performa KNN kurang optimal dalam menyeimbangkan ketepatan dan kemampuan deteksi kelas positif.

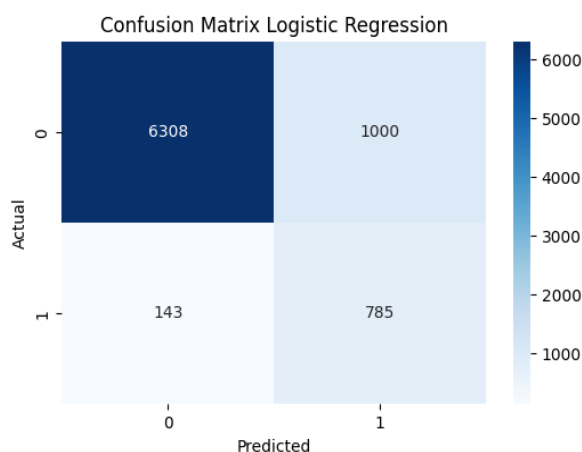


Gambar 7. ROC Curve KNN

Berdasarkan Gambar 7, hasil evaluasi ROC-AUC menunjukkan nilai sebesar 0,9060 yang termasuk dalam kategori sangat baik. Nilai tersebut menunjukkan bahwa model KNN memiliki kemampuan yang cukup baik dalam membedakan nasabah yang berlangganan dan tidak berlangganan deposito. Selain itu, perbandingan nilai train accuracy sebesar 100% dan test accuracy sebesar 87,11% menunjukkan selisih yang cukup besar, yaitu sekitar 12,89%, sehingga model mengalami indikasi overfitting dan memiliki kemampuan generalisasi yang kurang optimal terhadap data baru.

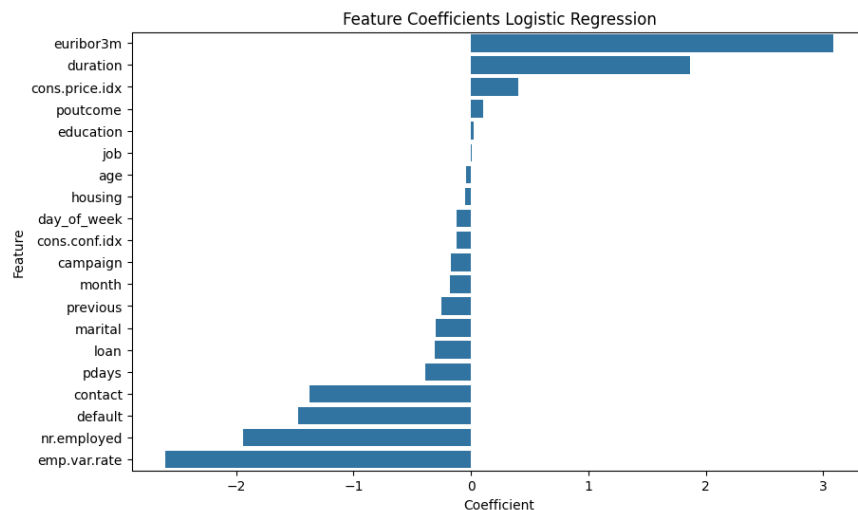
Hasil Pemodelan Logistic Regression

Model Logistic Regression dibangun menggunakan data hasil preprocessing dan penyeimbangan kelas menggunakan metode SMOTE. Evaluasi model dilakukan menggunakan confusion matrix, accuracy, precision, recall, F1-score, dan ROC-AUC untuk mengetahui kemampuan model dalam mengklasifikasikan keputusan nasabah dalam berlangganan deposito.



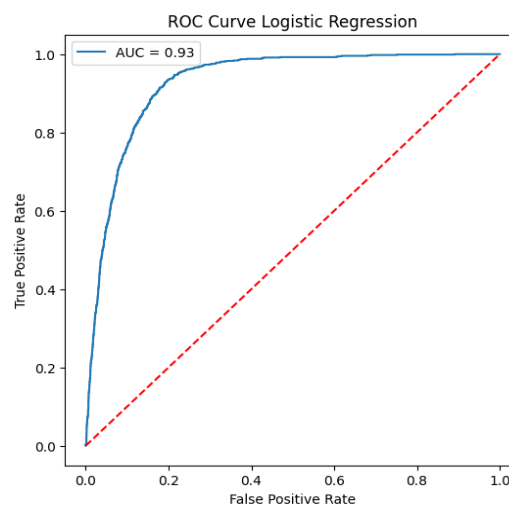
Gambar 8. Confusion Matrix Logistic Regression

Berdasarkan Gambar 8, hasil evaluasi, model Logistic Regression memperoleh accuracy sebesar 86,12%, precision sebesar 44,00%, recall sebesar 84,59%, dan F1-score sebesar 58,00%. Nilai recall yang tinggi menunjukkan bahwa model mampu mengenali sebagian besar nasabah yang benar-benar berlangganan deposito. Namun, nilai precision yang relatif rendah menunjukkan masih terdapat sejumlah prediksi positif yang tidak sesuai dengan kondisi sebenarnya. Hasil tersebut menunjukkan bahwa Logistic Regression memiliki kemampuan yang cukup baik dalam mendeteksi kelas positif.



Gambar 9. Feature Importance Logistic Regression

Berdasarkan Gambar 9, terdapat nilai koefisien model, variabel euribor3m memiliki pengaruh positif terbesar dengan nilai koefisien sebesar 3,0902, diikuti oleh variabel duration sebesar 1,8692 dan cons.price.idx sebesar 0,4026. Sebaliknya, variabel emp.var.rate memiliki koefisien negatif terbesar sebesar -2,6036, diikuti oleh nr.employed sebesar -1,9437, default sebesar -1,4675, dan contact sebesar -1,3766. Hasil tersebut menunjukkan bahwa variabel-variabel tersebut memiliki kontribusi terbesar dalam meningkatkan maupun menurunkan peluang nasabah untuk berlangganan deposito.

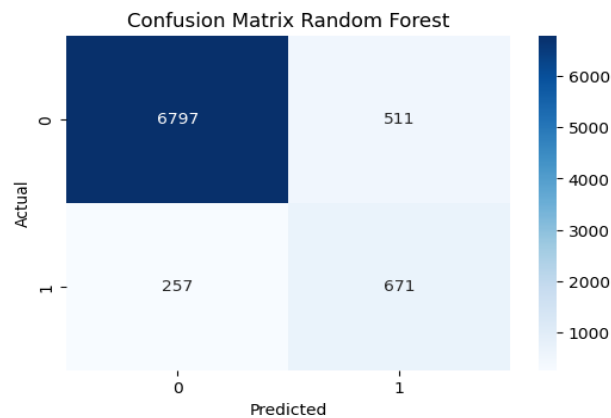


Gambar 10. ROC Curve Logistic Regression

Berdasarkan Gambar 10, terdapat hasil evaluasi menggunakan kurva ROC, model Logistic Regression memperoleh nilai ROC-AUC sebesar 0,9283 yang menunjukkan kemampuan sangat baik dalam membedakan nasabah yang berlangganan dan tidak berlangganan deposito. Selain itu, analisis overfitting menunjukkan bahwa nilai train accuracy sebesar 87,87% dan test accuracy sebesar 86,12% dengan selisih sekitar 1,75%, sehingga model tidak mengalami overfitting yang signifikan dan memiliki kemampuan generalisasi yang baik terhadap data baru.

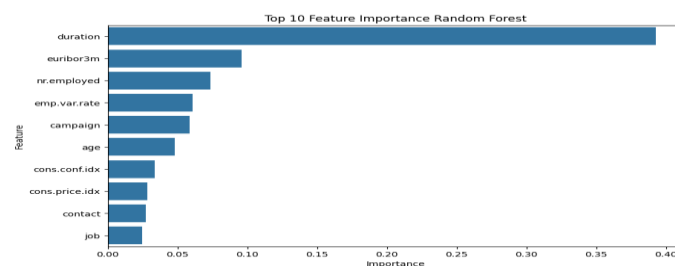
Hasil Pemodelan Random Forest

Random Forest merupakan metode ensemble learning yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih stabil dan akurat. Pada penelitian ini dilakukan proses hyperparameter tuning menggunakan GridSearchCV untuk memperoleh kombinasi parameter terbaik. Hasil tuning menunjukkan bahwa model terbaik menggunakan 200 pohon keputusan ($n_estimators = 200$) dengan $max_depth = None$, $min_samples_split = 2$, dan $min_samples_leaf = 1$. Parameter tersebut menghasilkan nilai rata-rata F1-score validasi silang sebesar 94,82%.



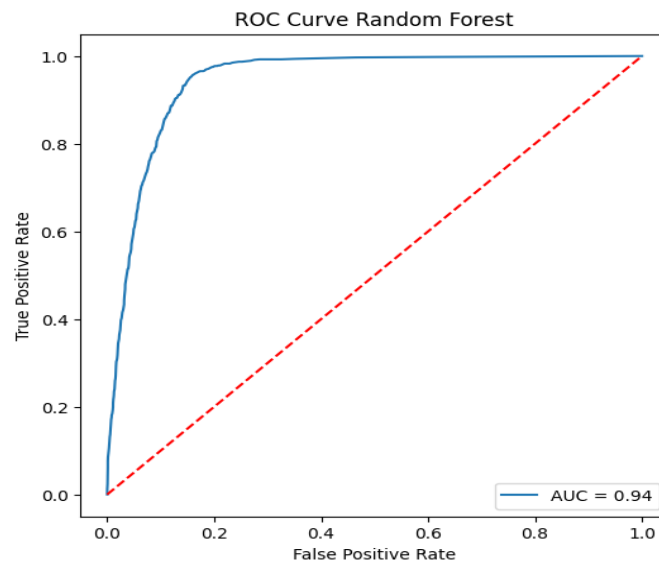
Gambar 11. Confusion Matrix Random Forest

Berdasarkan Gambar 11, hasil evaluasi, Random Forest memperoleh accuracy sebesar 90,68%, precision sebesar 56,77%, recall sebesar 72,31%, dan F1-score sebesar 63,60%. Hasil confusion matrix menunjukkan bahwa model berhasil mengidentifikasi dengan benar 6.797 nasabah yang tidak berlangganan deposito dan 671 nasabah yang berlangganan deposito. Dibandingkan model lainnya, Random Forest menghasilkan kombinasi precision dan F1-score yang paling tinggi. Hal ini menunjukkan bahwa model memiliki kemampuan yang lebih baik dalam menghasilkan prediksi yang akurat pada kedua kelas.



Gambar 12. Feature Importance Random Forest

Berdasarkan Gambar 12, variabel *duration* kembali menjadi variabel paling penting dengan nilai *importance* sebesar 0,3925. Variabel lain yang juga memberikan kontribusi besar adalah *euribor3m* sebesar 0,0957, *nr.employed* sebesar 0,0734, dan *emp.var.rate* sebesar 0,0606. Hasil ini menunjukkan bahwa keputusan nasabah dalam menerima penawaran deposito tidak hanya dipengaruhi oleh durasi komunikasi, tetapi juga oleh kondisi ekonomi pada saat kampanye berlangsung. Dengan demikian, faktor komunikasi dan indikator ekonomi menjadi komponen utama yang mempengaruhi keberhasilan prediksi model Random Forest.



Gambar 13. ROC Curve Random Forest

Berdasarkan Gambar 13, Random Forest menghasilkan nilai ROC-AUC sebesar 0,9447, yang merupakan nilai tertinggi dibandingkan model lainnya. Nilai tersebut menunjukkan bahwa Random Forest memiliki kemampuan yang sangat baik dalam membedakan nasabah yang berpotensi berlangganan deposito dan yang tidak berlangganan deposito. Semakin tinggi nilai ROC-AUC, semakin baik kemampuan model dalam melakukan klasifikasi pada kedua kelas. Dengan demikian, Random Forest dapat dianggap sebagai model yang paling optimal dalam penelitian ini.

Perbandingan Performa Model

Untuk menentukan model terbaik, dilakukan perbandingan performa berdasarkan accuracy, precision, recall, F1-score, dan ROC-AUC. Perbandingan ini penting karena setiap metrik memiliki fungsi yang berbeda dalam mengevaluasi model klasifikasi. Accuracy digunakan untuk melihat tingkat ketepatan prediksi secara keseluruhan, sedangkan precision dan recall digunakan untuk mengevaluasi kemampuan model dalam mengidentifikasi kelas positif. Selain itu, F1-score digunakan untuk menyeimbangkan precision dan recall, sedangkan ROC-AUC digunakan untuk mengukur kemampuan model dalam membedakan kedua kelas.

Tabel 3. Perbandingan Performa Model

Model	Accuracy	Precision	Recall	F1-Score
<i>Decision Tree</i>	92.68%	47,17%	80.82%	59.57%
KNN	100%	44,45%	57.87%	50.28%
<i>Logistic Regression</i>	87.87%	43,98%	84.59%	57.87%
<i>Random Forest</i>	100%	56,77%	72.31%	63.60%

Berdasarkan Tabel 3, Decision Tree memperoleh nilai recall sebesar 80,82% dan F1-score sebesar 59,57%, yang menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mendeteksi nasabah yang berlangganan deposito. Namun, nilai precision sebesar 47,17% menunjukkan bahwa masih terdapat beberapa prediksi positif yang kurang sesuai dengan kondisi sebenarnya.

Logistic Regression memperoleh nilai recall tertinggi sebesar 84,59%, yang menunjukkan kemampuan yang sangat baik dalam mengenali nasabah yang benar-benar berlangganan deposito. Akan tetapi, nilai precision yang diperoleh masih relatif rendah yaitu sebesar 43,98%, sehingga masih terdapat cukup banyak kesalahan dalam memprediksi nasabah yang akan berlangganan deposito.

Meskipun KNN memperoleh accuracy sebesar 100%, model ini memiliki nilai precision sebesar 44,45%, recall sebesar 57,87%, dan F1-score sebesar 50,28% yang lebih rendah dibandingkan model lainnya. Selain itu, hasil analisis overfitting sebelumnya menunjukkan adanya perbedaan yang cukup besar antara accuracy training dan accuracy testing, sehingga kemampuan generalisasi KNN terhadap data baru masih kurang optimal.

Sementara itu, Random Forest juga memperoleh accuracy training sebesar 100% dengan nilai precision sebesar 56,77%, recall sebesar 72,31%, dan F1-score sebesar 63,60%. Nilai tersebut menunjukkan bahwa Random Forest memiliki kemampuan yang baik dalam menghasilkan prediksi positif yang tepat serta menjaga keseimbangan antara precision dan recall dibandingkan model lainnya pada tahap evaluasi ini.

Tabel 4. Perbandingan Nilai ROC-AUC

Model	ROC-AUC
<i>Decision Tree</i>	0,9264
KNN	0,9060
<i>Logistic Regression</i>	0,9283
<i>Random Forest</i>	0,9447

Berdasarkan Tabel 4, Random Forest memperoleh nilai ROC-AUC tertinggi sebesar 0,9447, yang menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan nasabah yang berlangganan dan tidak berlangganan deposito.

Logistic Regression memperoleh nilai ROC-AUC sebesar 0,9283, diikuti oleh Decision Tree sebesar 0,9264, sedangkan KNN memiliki nilai ROC-AUC terendah sebesar 0,9060. Meskipun seluruh model menunjukkan kemampuan klasifikasi yang sangat baik karena memiliki nilai ROC-AUC di atas 0,90, Random Forest memberikan kemampuan diskriminasi yang paling tinggi dibandingkan model lainnya.

Berdasarkan keseluruhan hasil evaluasi yang meliputi accuracy, precision, recall, F1-score, dan ROC-AUC, Random Forest dipilih sebagai model terbaik pada penelitian ini. Selain memiliki nilai precision dan F1-score yang paling tinggi, Random Forest juga memperoleh nilai ROC-AUC tertinggi serta menunjukkan kemampuan generalisasi yang baik dalam melakukan prediksi terhadap data baru. Hal tersebut menunjukkan bahwa pendekatan ensemble learning yang menggabungkan beberapa pohon keputusan mampu menghasilkan model yang lebih stabil dan efektif dalam memprediksi keputusan nasabah untuk berlangganan deposito berjangka.

KESIMPULAN

Penelitian ini berhasil menerapkan empat algoritma machine learning, yaitu Decision Tree, K-Nearest Neighbor (KNN), Logistic Regression, dan Random Forest dalam memprediksi keputusan nasabah untuk berlangganan deposito berjangka menggunakan Bank Marketing Dataset. Sebelum proses pemodelan dilakukan, data melalui tahap preprocessing yang meliputi data cleaning, label encoding, standardisasi data, pembagian data pelatihan dan pengujian, serta penanganan ketidakseimbangan kelas menggunakan metode SMOTE. Hasil preprocessing menunjukkan bahwa jumlah data setelah proses pembersihan menjadi sebanyak 41.176 observasi dan distribusi kelas yang sebelumnya tidak seimbang berhasil diseimbangkan menggunakan metode SMOTE.

Hasil evaluasi menunjukkan bahwa setiap algoritma memiliki karakteristik performa yang berbeda. Decision Tree menghasilkan nilai F1-score sebesar 59,57% dengan kemampuan yang baik dalam mendeteksi nasabah yang berpotensi berlangganan deposito. Logistic Regression memiliki kemampuan deteksi kelas positif tertinggi dengan nilai recall sebesar 84,59%, namun masih memiliki nilai precision yang relatif rendah. Sementara itu, KNN menunjukkan indikasi overfitting karena memperoleh accuracy sebesar 100%, tetapi memiliki penurunan performa pada data pengujian dengan nilai F1-score sebesar 50,28%, sehingga kemampuan generalisasinya terhadap data baru masih kurang optimal.

Berdasarkan keseluruhan metrik evaluasi yang meliputi accuracy, precision, recall, F1-score, dan ROC-AUC, Random Forest memberikan performa terbaik dibandingkan model lainnya dengan nilai precision sebesar 56,77%, recall sebesar 72,31%, F1-score sebesar 63,60%, dan ROC-AUC sebesar 94,47%. Hasil tersebut menunjukkan bahwa Random Forest memiliki kemampuan klasifikasi yang lebih seimbang dan kemampuan diskriminasi yang sangat baik dalam membedakan nasabah yang berlangganan dan tidak berlangganan deposito.

Selain itu, hasil analisis korelasi dan feature importance menunjukkan bahwa variabel duration menjadi salah satu faktor yang paling berpengaruh terhadap keputusan nasabah dalam berlangganan deposito. Temuan ini mengindikasikan bahwa lamanya komunikasi antara pihak bank dan nasabah memiliki peranan penting terhadap keberhasilan pemasaran produk deposito berjangka. Dengan demikian, penerapan machine learning, khususnya metode Random Forest, dapat dimanfaatkan sebagai alat bantu dalam mendukung strategi pemasaran berbasis data pada sektor perbankan.

SARAN

Berdasarkan hasil penelitian yang telah dilakukan, penelitian selanjutnya disarankan untuk mengeksplorasi algoritma machine learning lain yang lebih kompleks, seperti XGBoost, LightGBM, atau Artificial Neural Network (ANN), untuk memperoleh performa prediksi yang lebih baik. Selain itu, dapat dilakukan perbandingan berbagai teknik penanganan data tidak seimbang, seperti ADASYN atau Borderline-SMOTE, guna meningkatkan kemampuan model dalam mengklasifikasikan kelas minoritas. Penelitian berikutnya juga disarankan untuk melakukan seleksi fitur dan tuning hyperparameter yang lebih mendalam agar diperoleh model yang lebih optimal dan efisien. Di samping itu, penggunaan dataset yang lebih terbaru atau data dari institusi perbankan yang berbeda dapat dilakukan untuk menguji konsistensi dan kemampuan generalisasi model pada kondisi yang lebih beragam.

DAFTAR PUSTAKA

- Abidin, Dodo Zaenal, Maria Rosario, Ali Sadikin, Nurhadi, dan Jasmir. 2025. "Improving Term Deposit Customer Prediction Using Support Vector Machine with SMOTE and Hyperparameter Tuning in Bank Marketing Campaigns." *Jurnal Teknik Informatika (JUTIF)*, 6(3): 1267–1278. DOI: <https://doi.org/10.52436/1.jutif.2025.6.3.4585>
- Alzubaidi, Laith, et al. 2021. "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions." *Journal of Big Data*, 8: 53. DOI: <https://doi.org/10.1186/s40537-021-00444-8>
- Ardhana, S. R., Widihari, T., dan Saputra, B. A. (2023). "Klasifikasi Menggunakan Algoritma K-Nearest Neighbor pada Imbalance Class Data dengan SMOTE." *Indonesian Journal of Applied Statistics*, 6(2), 152–161. DOI: <https://doi.org/10.13057/ijas.v6i2.79389>
- Chauhan, A. S., et al. 2023. "Comparative Analysis of Supervised Machine and Deep Learning Algorithms for Classification." *Applied Sciences*, 13(8): 5012. Diakses dari: <https://www.mdpi.com/2076-3417/13/8/5012>
- Du, S. 2025. "Findings from the Bank Marketing Dataset: Using Machine Learning to Forecast Term Deposit Subscriptions." *Advances in Economics, Management and Political Sciences*, 240: 62–69. DOI: <https://doi.org/10.54254/2754-1169/2025.BL29270>
- Erlangga, A. A. G. W. S., Gunadi, I. G. A., dan Sunarya, I. M. G. (2024). "Kombinasi Oversampling dan Undersampling dalam Menangani Class Imbalanced dan Overlapping pada Klasifikasi Data Bank Marketing." *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, 7(1). DOI: <https://doi.org/10.31598/jurnalresistor.v7i1.1515>
- Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., dan Zoromi, F. (2022). "Dampak SMOTE terhadap Kinerja Random Forest Classifier Berdasarkan Data Tidak Seimbang." *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 21(3), 667–678. DOI: <https://doi.org/10.30812/matrik.v21i3.1726>
- Fauziah, A., dan Hernadi, J. (2024). "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Random Forest dengan SMOTE dan SMOTE-ENN." *Jurnal Teknomatika*, 17(2). DOI: <https://doi.org/10.30989/teknomatika.v17i2.1530>
- Khanam, J., dan S. Y. Foo. 2021. "A Comparison of Machine Learning Algorithms for Diabetes Prediction." *ICT Express*, 7(4): 432–439. DOI: <https://doi.org/10.1016/j.icte.2021.02.004>
- Kurniawan, M. F., dan D. A. Megawaty. 2025. "Comparison of Logistic Regression, Random Forest, Support Vector Machine and K-Nearest Neighbor Algorithms in

-
- Diabetes Prediction.” *Journal of Applied Informatics and Computing*. DOI: <https://doi.org/10.30871/jaic.v9i5.9815>
- Sarker, Iqbal H. 2021. “Machine Learning: Algorithms, Real-World Applications and Research Directions.” *SN Computer Science*, 2(3): 160. DOI: <https://doi.org/10.1007/s42979-021-00592-x>
- Suryantari, P. A., Muttaqin, F., dan Rahajoe, A. D. (2026). “Peran Big Data dalam Manajemen Data dan Informasi sebagai Sistem Pendukung Keputusan.” *Jurnal Informatika dan Teknik Elektro Terapan*, 14(1). DOI: <https://doi.org/10.23960/jitet.v14i1.8899>