

**ANALISIS KOMPARASI ALGORITMA MACHINE
LEARNING PADA KLASIFIKASI KELAYAKAN PINJAMAN
NASABAH**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

MUHAMMAD WILDANU STAIFEN

22.11.4783

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**ANALISIS KOMPARASI ALGORITMA MACHINE
LEARNING PADA KLASIFIKASI KELAYAKAN PINJAMAN
NASABAH**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh
MUHAMMAD WILDANU STAIFEN

22.11.4783

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**ANALISIS KOMPARASI ALGORITMA MACHINE
LEARNING PADA KLASIFIKASI KELAYAKAN PINJAMAN
NASABAH**

yang disusun dan diajukan oleh

Muhammad Wildanu Staifen

22.11.4783

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 18 Mei 2026

Dosen Pembimbing,


Majid Rahardi, S.Kom., M.Eng.

NIK. 190302393

HALAMAN PENGESAHAN

SKRIPSI

**ANALISIS KOMPARASI ALGORITMA MACHINE
LEARNING PADA KLASIFIKASI KELAYAKAN PINJAMAN
NASABAH**

yang disusun dan diajukan oleh

Muhammad Wildanu Staifen

22.11.4783

Telah dipertahankan di depan Dewan Penguji
pada tanggal 18 Mei 2026

Susunan Dewan Penguji

Nama Penguji

Rumini, S.Kom., M.Kom.
NIK. 190302246

Yudha Riwanto, M.Kom.
NIK. 190302702

Majid Rahardi, S.Kom., M.Eng.
NIK. 190302393

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 18 Mei 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.

NIK. 190302106

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Wildanu Staifen

NIM : 22.11.4783

Menyatakan bahwa Skripsi dengan judul berikut:

Analisis Komparasi Algoritma Machine Learning Pada Klasifikasi Kelayakan Pinjaman Nasabah

Dosen Pembimbing : Majid Rahardi, S.Kom, M.Eng

1. Karya tulis ini adalah benar-benar **ASLI** dan **BELUM PERNAH** diajukan untuk mendapatkan gelar akademik, baik di Universitas **AMIKOM** Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian **SAYA** sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab **SAYA**, bukan tanggung jawab Universitas **AMIKOM** Yogyakarta.
5. Pernyataan ini **SAYA** buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka **SAYA** bersedia menerima **SANKSI AKADEMIK** dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

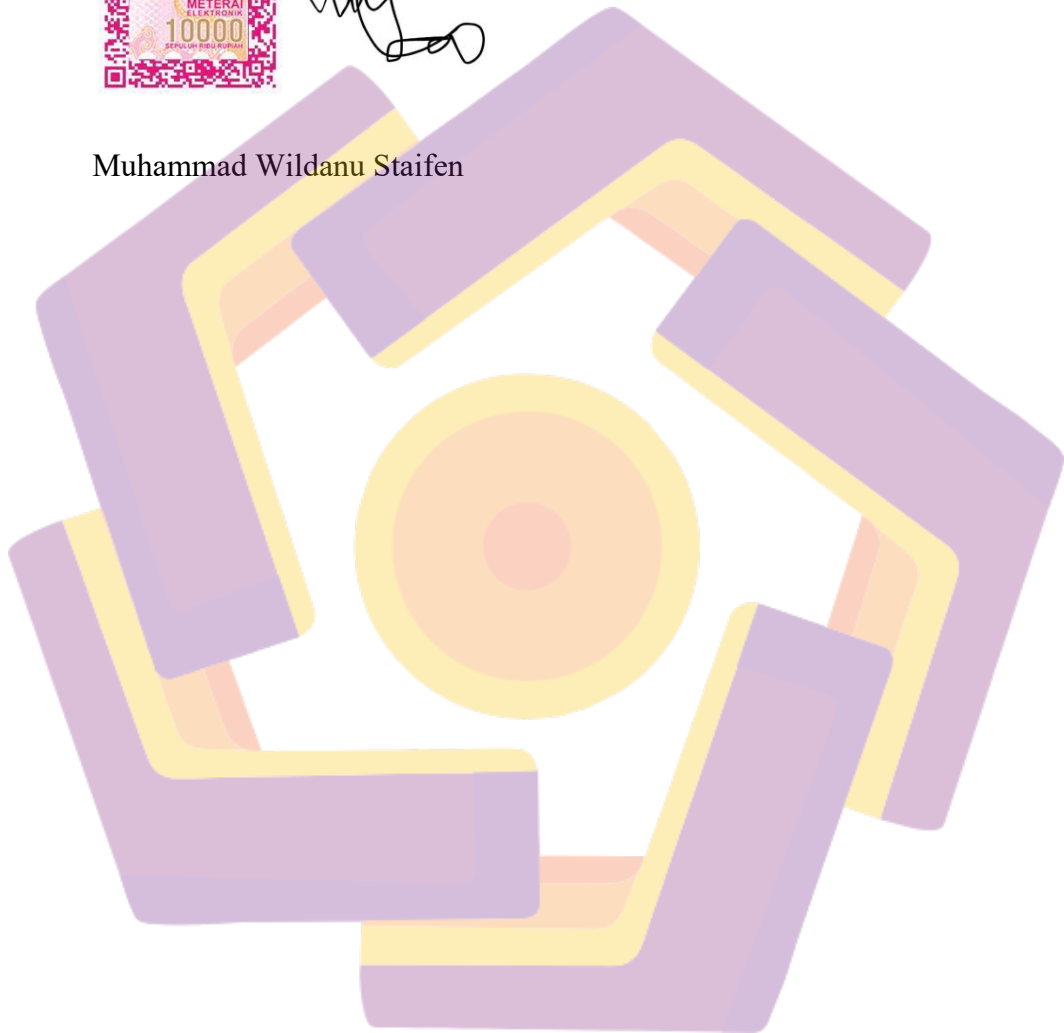
Yogyakarta, 18 Mei 2026

Yang Menyatakan,



A handwritten signature in black ink, appearing to read "Muhammad Wildanu Staifen".

Muhammad Wildanu Staifen



HALAMAN PERSEMBAHAN

Skripsi ini penulis persembahkan dengan segenap rasa syukur dan kerendahan hati kepada Tuhan Yang Maha Esa, yang atas segala rahmat, kemudahan, dan kekuatan-Nya karya ini dapat diselesaikan.

Kepada kedua orang tua, terima kasih atas setiap doa yang tidak pernah berhenti dipanjatkan, atas pengorbanan yang tidak pernah dikeluhkan, dan atas kasih sayang yang menjadi kekuatan terbesar dalam setiap langkah perjalanan ini. Tidak ada kata yang cukup untuk membalas semua yang telah diberikan. Semoga karya ini menjadi salah satu bukti kecil dari bakti yang tulus.

Kepada seluruh keluarga yang selalu hadir dengan dukungan, perhatian, dan semangat yang tulus. Kehadiran kalian menjadi tempat pulang yang hangat di tengah perjalanan yang penuh tantangan.

Kepada dosen pembimbing dan seluruh dosen, terima kasih atas bimbingan, ilmu, dan kesabaran yang diberikan dengan tulus selama masa studi. Setiap arahan yang diberikan menjadi cahaya yang menuntun penulis untuk terus berkembang dan tidak mudah menyerah.

Kepada sahabat dan rekan-rekan seperjuangan, terima kasih atas kebersamaan, semangat yang saling dibagikan, dan tawa yang selalu menjadi pengingat bahwa perjalanan ini tidak harus ditempuh sendirian.

Serta kepada almamater tercinta, semoga karya ini dapat memberikan manfaat dan menjadi bagian kecil dari kontribusi bagi perkembangan ilmu pengetahuan.

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga skripsi yang berjudul *Analisis Komparasi Algoritma Machine Learning Pada Klasifikasi Kelayakan Pinjaman Nasabah* dapat diselesaikan dengan baik.

Penyusunan skripsi ini tidak terlepas dari bantuan dan dukungan berbagai pihak. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada Dosen Pembimbing atas bimbingan, arahan, dan masukan yang diberikan selama proses penyusunan skripsi. Ucapan terima kasih juga disampaikan kepada Tim Dosen Penguji atas saran dan kritik yang membangun, kepada seluruh dosen dan staf akademik yang telah memberikan ilmu dan pelayanan selama masa studi, serta kepada kedua orang tua dan keluarga atas doa dan dukungan yang senantiasa diberikan. Penulis juga mengucapkan terima kasih kepada semua pihak yang telah membantu secara langsung maupun tidak langsung dalam penyelesaian skripsi ini.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan. Oleh karena itu, saran dan kritik yang membangun sangat diharapkan demi penyempurnaan di masa yang akan datang. Semoga skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan dan pihak-pihak yang membutuhkan.

Yogyakarta, 18 Mei 2026

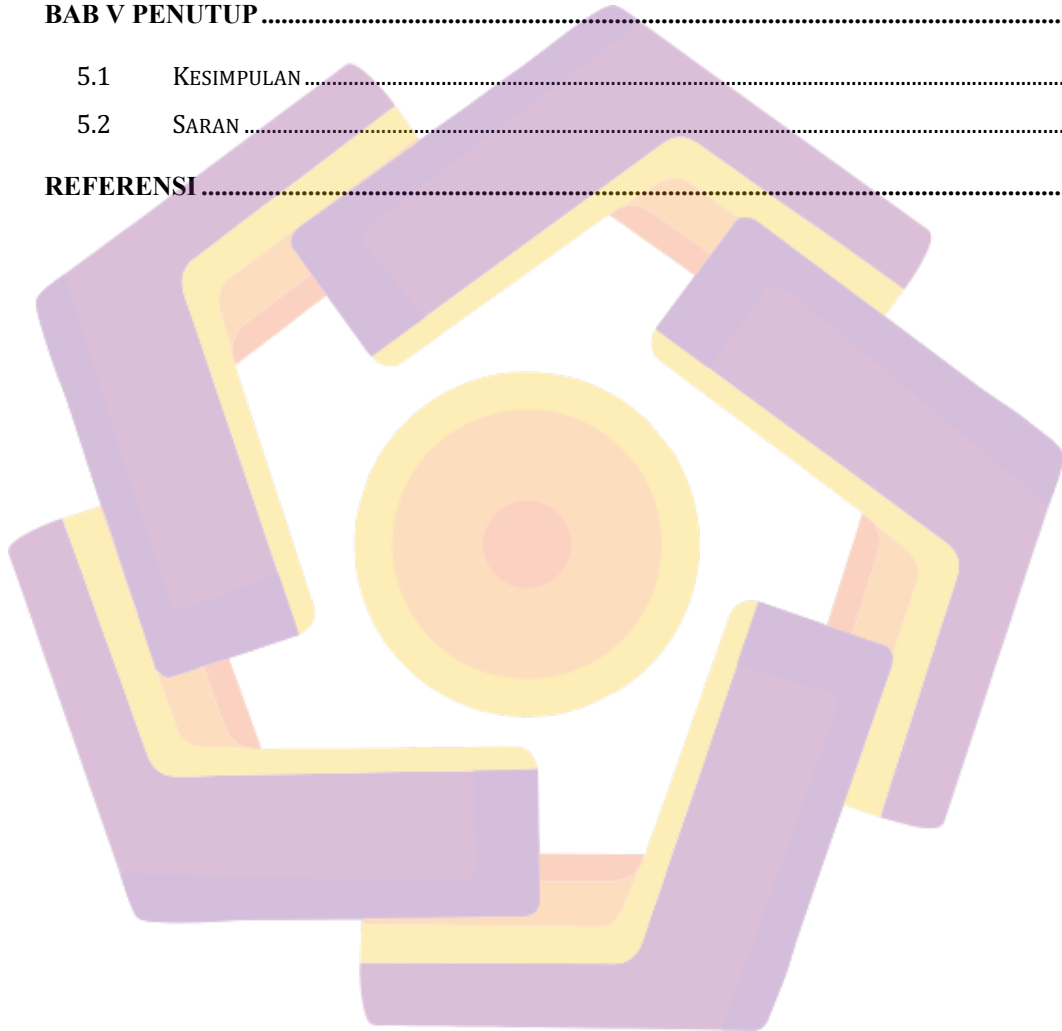
Muhammad Wildanu Staifen

DAFTAR ISI

HALAMAN JUDUL.....	I
HALAMAN PERSETUJUAN.....	II
HALAMAN PENGESAHAN	III
HALAMAN PERNYATAAN KEASLIAN SKRIPSI. ERROR! BOOKMARK NOT DEFINED.	
HALAMAN PERSEMBAHAN.....	VI
KATA PENGANTAR.....	VII
DAFTAR ISI.....	VIII
DAFTAR TABEL.....	XI
DAFTAR GAMBAR.....	XIII
DAFTAR LAMBANG DAN SINGKATAN.....	XIV
DAFTAR ISTILAH.....	XV
INTISARI.....	XXIII
ABSTRACT.....	XXIV
BAB I PENDAHULUAN	1
1.1 LATAR BELAKANG.....	1
1.2 RUMUSAN MASALAH.....	2
1.3 BATASAN MASALAH.....	2
1.4 TUJUAN PENELITIAN.....	3
1.5 MANFAAT PENELITIAN.....	4
1.6 SISTEMATIKA PENULISAN	4
BAB II TINJAUAN PUSTAKA.....	6
2.1 STUDI LITERATUR	6
2.2 DASAR TEORI.....	14
2.2.1 Machine Learning	14
2.2.2 Klasifikasi.....	15
2.2.3 Kelayakan Pinjaman Nasabah	15
2.2.5 Data Leakage	16
2.2.6 Overfitting.....	17

2.2.7	<i>Logistic Regression</i>	17
2.2.8	<i>Random Forest</i>	18
2.2.9	<i>LightGBM</i>	19
2.2.10	<i>CatBoost</i>	21
2.2.11	<i>Evaluasi Model Klasifikasi</i>	22
2.2.12	<i>Fine Tuning Hyperparameter</i>	24
BAB III METODE PENELITIAN		25
3.1	OBJEK PENELITIAN.....	25
3.2	ALUR PENELITIAN	27
3.2.1	<i>Identifikasi Masalah dan Studi Literatur</i>	29
3.2.2	<i>Pengumpulan Dataset</i>	29
3.2.3	<i>Profil Dataset dan Pemahaman Variabel</i>	30
3.2.4	<i>Exploratory Data Analysis dan Visualisasi</i>	40
3.2.5	<i>Preprocessing Data</i>	40
3.2.6	<i>Pemisahan Data</i>	41
3.2.7	<i>Pelatihan Model Dasar</i>	42
3.2.8	<i>Hyperparameter Tuning</i>	45
3.2.9	<i>Evaluasi Model</i>	45
3.2.10	<i>Pengendalian Data Leakage dan Overfitting</i>	46
3.3	ALAT DAN BAHAN	47
3.3.1	<i>Data Penelitian</i>	47
3.3.2	<i>Alat Penelitian</i>	47
3.3.3	<i>Bahan Penelitian</i>	49
BAB IV HASIL DAN PEMBAHASAN		50
4.1	HASIL PERSIAPAN DATA	50
4.2	HASIL EDA.....	52
4.3	HASIL PELATIHAN MODEL BASELINE	54
4.3	HASIL FINE TUNING HYPERPARAMETER	56
4.4	PERBANDINGAN BASELINE DAN FINE TUNING.....	60
4.5	ANALISIS OVERFITTING.....	62
4.6	ANALISIS PENCEGAHAN DATA LEAKAGE.....	64
4.7	ANALISIS CONFUSION MATRIX.....	65
4.8	ANALISIS MENGAPA AKURASI SANGAT TINGGI.....	68
4.9	ANALISIS FITUR PENTING.....	73
4.10	PEMBAHASAN KETERKAITAN DENGAN PENELITIAN TERDAHULU	76

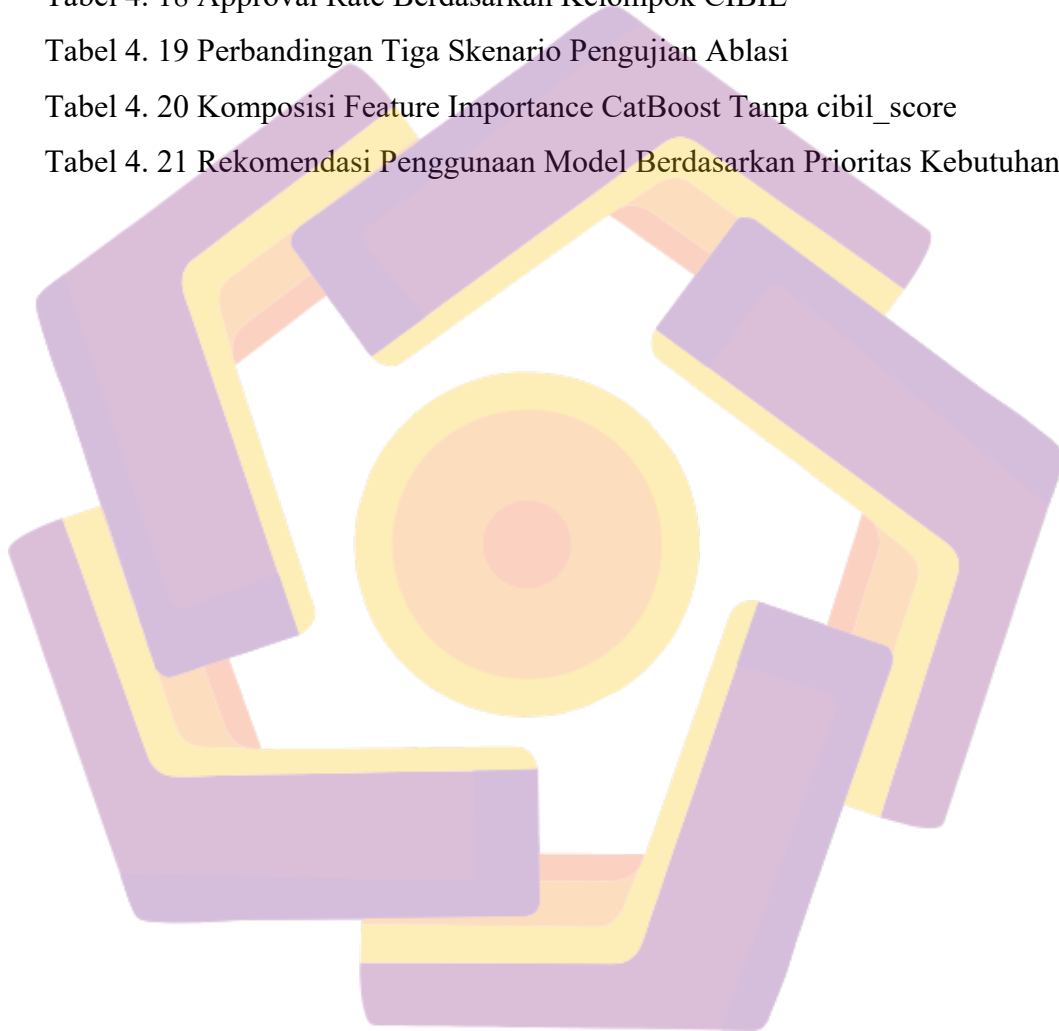
4.11	IMPLEMENTASI SISTEM.....	78
4.11.1	<i>Halaman Prediksi</i>	79
4.11.2	<i>Hasil Prediksi dan Perbandingan Model</i>	80
4.11.3	<i>Riwayat Prediksi</i>	80
4.11.4	<i>Dashboard Metrik Model</i>	82
4.11.5	<i>Halaman Tentang</i>	83
BAB V	PENUTUP.....	85
5.1	KESIMPULAN.....	85
5.2	SARAN.....	86
REFERENSI		88



DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian	9
Tabel 3. 1 Profil Objek Penelitian	26
Tabel 3. 2 Deskripsi Fitur Dataset Mentah	30
Tabel 3. 3 Contoh Lima Baris Dataset Mentah	32
Tabel 3. 4 Fitur Turunan Hasil Rekayasa Fitur	33
Tabel 3. 5 Struktur Dataset Setelah Preprocessing	35
Tabel 3. 6 Statistik Deskriptif Fitur Numerik Dataset Mentah	37
Tabel 3. 7 Distribusi Fitur Kategorikal dan Approval Rate	38
Tabel 3. 8 Komponen Pipeline Preprocessing per Algoritma	39
Tabel 3. 9 Distribusi Kelas pada Setiap Subset Data	42
Tabel 3. 10 Data Contoh Perhitungan Manual Logistic Regression	43
Tabel 3. 11 Hasil Prediksi Langkah 3	44
Tabel 3. 12 Perangkat Keras Penelitian	47
Tabel 3. 13 Perangkat Lunak Penelitian	48
Tabel 4. 1 Distribusi Kelas Setelah Split	50
Tabel 4. 2 Ringkasan Hasil Pemeriksaan dan Persiapan Data	51
Tabel 4. 3 Korelasi Fitur Numerik Terhadap Target Kelakyakan	52
Tabel 4. 4 Statistik Deskriptif Fitur Turunan	53
Tabel 4. 5 Hasil Baseline pada Data Uji	54
Tabel 4. 6 Stabilitas Model Baseline Berdasarkan F1-score	55
Tabel 4. 7 Hasil Cross-Validation 5-Fold Model Baseline	56
Tabel 4. 8 Parameter Terbaik Hasil Fine Tuning	57
Tabel 4. 9 Hasil Fine Tuning pada Data Uji	57
Tabel 4. 10 Perbandingan F1 Validasi Kandidat Hyperparameter	59
Tabel 4. 11 Perbandingan Baseline dan Tuning pada Data Uji	61
Tabel 4. 12 Perbandingan ROC-AUC Baseline vs Tuning pada Data Uji	62

Tabel 4. 13 Analisis Overfitting Model Hasil Tuning	63
Tabel 4. 14 Interpretasi Status Generalisasi Model Hasil Tuning	64
Tabel 4. 15 Confusion Matrix Model Hasil Tuning	65
Tabel 4. 16 Analisis FPR dan FNR Model Hasil Tuning	67
Tabel 4. 17 Top 8 Feature Importance (%) Model Berbasis Pohon	69
Tabel 4. 18 Approval Rate Berdasarkan Kelompok CIBIL	70
Tabel 4. 19 Perbandingan Tiga Skenario Pengujian Ablasi	72
Tabel 4. 20 Komposisi Feature Importance CatBoost Tanpa cibil_score	73
Tabel 4. 21 Rekomendasi Penggunaan Model Berdasarkan Prioritas Kebutuhan	75



DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian	28
Gambar 4. 1 Perbandingan Metrik Utama Model Hasil Tuning	60
Gambar 4. 2 Tabel Perbandingan Hasil Tuning	60
Gambar 4. 3 Confusion Matrix Hyperparameter Tuning	66
Gambar 4. 4 Fitur terpenting Dari Setiap Model	74
Gambar 4. 5 Tampilan Halaman Prediksi	79
Gambar 4. 6 Tampilan Hasil Prediksi dan Perbandingan Keempat Model	80
Gambar 4. 7 Tampilan Halaman Riwayat Prediksi	81
Gambar 4. 8 Tampilan Modal Detail Prediksi	82
Gambar 4. 9 Tampilan Dashboard Metrik Evaluasi Keempat Model	83
Gambar 4. 10 Tampilan Halaman Tentang Aplikasi	84

DAFTAR LAMBANG DAN SINGKATAN

ML : Machine Learning
LR : Logistic Regression
RF : Random Forest
LGBM : LightGBM — Light Gradient Boosting Machine
CB : CatBoost — Categorical Boosting
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative
FPR : False Positive Rate
FNR : False Negative Rate
AUC : Area Under the Curve
ROC : Receiver Operating Characteristic
CV : Cross-Validation
NPL : Non-Performing Loan — kredit bermasalah
CIBIL : Credit Information Bureau (India) Limited
EDA : Exploratory Data Analysis
EYD : Ejaan yang Disempurnakan

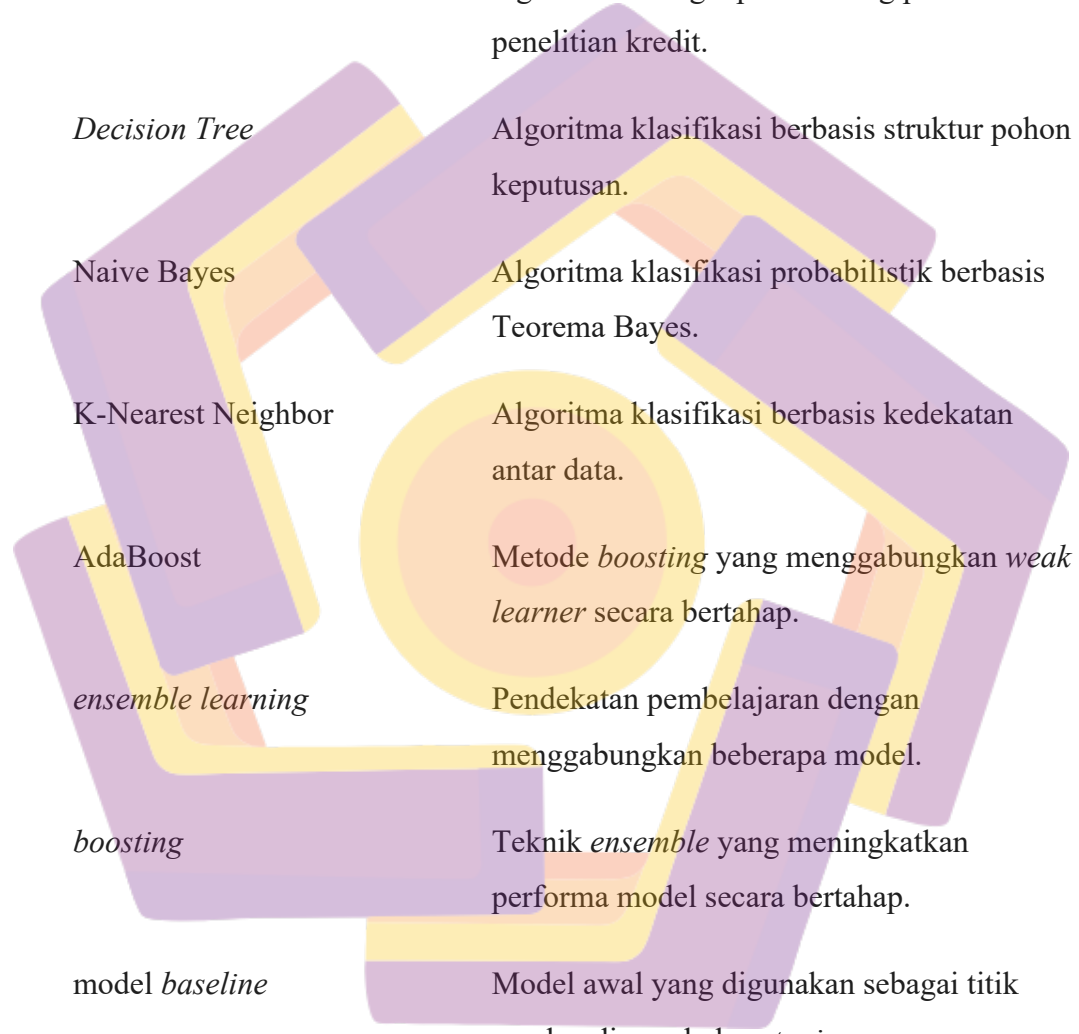
DAFTAR ISTILAH



<i>loan_id</i>	ID unik untuk setiap pengajuan pinjaman.
<i>no_of_dependents</i>	Jumlah tanggungan nasabah.
<i>education</i>	Tingkat pendidikan nasabah.
<i>self_employed</i>	Status pekerjaan mandiri nasabah.
<i>income_annum</i>	Pendapatan tahunan nasabah.
<i>loan_amount</i>	Jumlah pinjaman yang diajukan.
<i>loan_term</i>	Jangka waktu pinjaman.
<i>cibil_score</i>	Skor CIBIL atau skor kelayakan kredit nasabah.
<i>residential_assets_value</i>	Nilai aset residensial yang dimiliki nasabah.
<i>commercial_assets_value</i>	Nilai aset komersial yang dimiliki nasabah.
<i>luxury_assets_value</i>	Nilai aset mewah yang dimiliki nasabah.
<i>bank_asset_value</i>	Nilai aset nasabah yang tersimpan di bank.
<i>loan_status</i>	Status persetujuan pinjaman.
<i>Approved</i>	Status pinjaman disetujui.
<i>Rejected</i>	Status pinjaman ditolak.



<i>Approved = 1</i>	Representasi numerik untuk kelas pinjaman disetujui.
<i>Rejected = 0</i>	Representasi numerik untuk kelas pinjaman ditolak.
<i>Graduate</i>	Kategori nasabah lulusan pendidikan formal.
<i>Not Graduate</i>	Kategori nasabah belum lulus atau tidak termasuk lulusan formal.
<i>self_employed = Yes</i>	Nasabah bekerja mandiri.
<i>self_employed = No</i>	Nasabah tidak bekerja mandiri.
<i>total_assets_value</i>	Total seluruh aset nasabah dari komponen aset yang tersedia pada <i>dataset</i> .
<i>rasio_pinjaman_pendapatan</i>	Rasio antara jumlah pinjaman dan pendapatan tahunan nasabah.
<i>rasio_agunan_pinjaman</i>	Rasio antara total aset atau agunan dan jumlah pinjaman.
<i>target_kelayakan</i>	Variabel target biner hasil transformasi dari <i>loan_status</i> .
<i>Logistic Regression</i>	Algoritma klasifikasi linear yang digunakan untuk memprediksi probabilitas suatu kelas.
<i>Random Forest</i>	Algoritma <i>ensemble</i> berbasis banyak <i>decision tree</i> .
LightGBM	Algoritma <i>gradient boosting</i> yang dirancang efisien untuk data tabular.

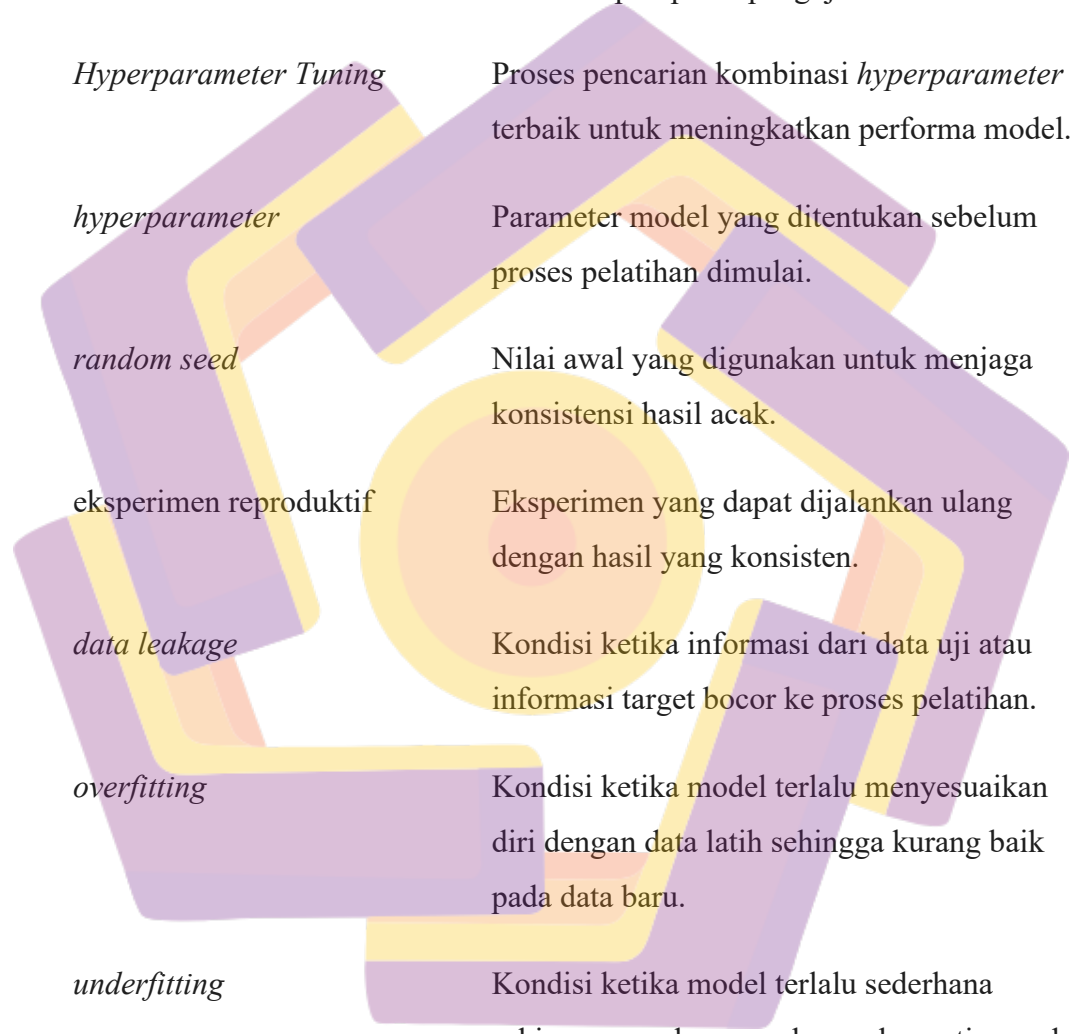


CatBoost	Algoritma <i>boosting</i> yang kuat untuk data tabular dan efektif menangani fitur kategorikal.
XGBoost	Algoritma <i>gradient boosting</i> yang sering digunakan sebagai pembanding pada penelitian kredit.
<i>Decision Tree</i>	Algoritma klasifikasi berbasis struktur pohon keputusan.
Naive Bayes	Algoritma klasifikasi probabilistik berbasis Teorema Bayes.
K-Nearest Neighbor	Algoritma klasifikasi berbasis kedekatan antar data.
AdaBoost	Metode <i>boosting</i> yang menggabungkan <i>weak learner</i> secara bertahap.
<i>ensemble learning</i>	Pendekatan pembelajaran dengan menggabungkan beberapa model.
<i>boosting</i>	Teknik <i>ensemble</i> yang meningkatkan performa model secara bertahap.
model <i>baseline</i>	Model awal yang digunakan sebagai titik pembanding sebelum <i>tuning</i> .
model terbaik	Model dengan performa evaluasi paling unggul berdasarkan metrik yang ditetapkan.
<i>Accuracy</i>	Persentase prediksi yang benar dari seluruh data.

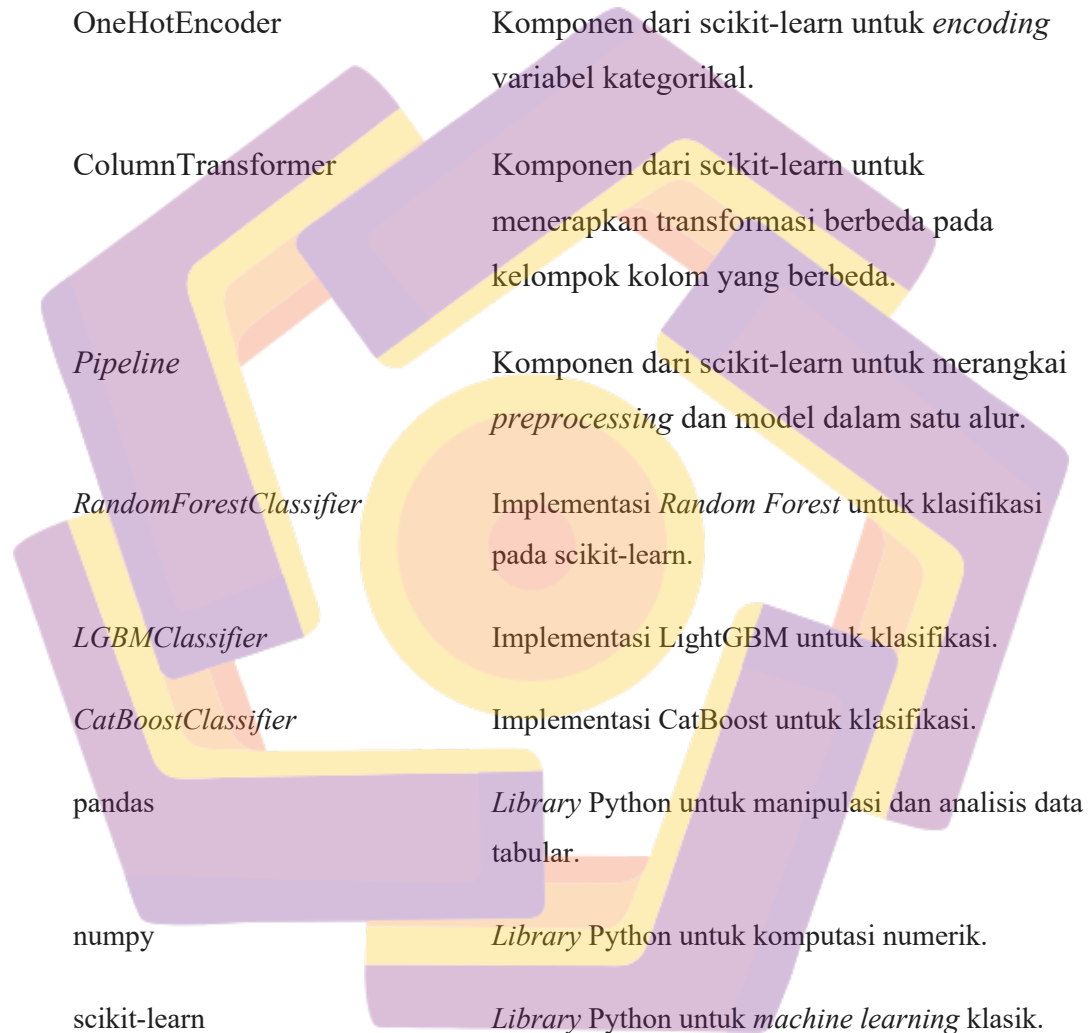


<i>Precision</i>	Rasio prediksi positif yang benar terhadap seluruh prediksi positif.
<i>Recall</i>	Rasio prediksi positif yang benar terhadap seluruh data positif aktual.
<i>F1-score</i>	Rata-rata harmonik antara <i>Precision</i> dan <i>Recall</i> .
<i>ROC-AUC</i>	Luas area di bawah kurva ROC yang menunjukkan kemampuan model membedakan dua kelas.
<i>Confusion Matrix</i>	Tabel evaluasi yang menampilkan jumlah prediksi benar dan salah pada tiap kelas.
<i>true positive (TP)</i>	Data positif yang diprediksi positif secara benar.
<i>true negative (TN)</i>	Data negatif yang diprediksi negatif secara benar.
<i>false positive (FP)</i>	Data negatif yang keliru diprediksi positif.
<i>false negative (FN)</i>	Data positif yang keliru diprediksi negatif.
Kurva ROC	Kurva yang menunjukkan hubungan antara <i>true positive rate</i> dan <i>false positive rate</i> .
<i>feature importance</i>	Ukuran kontribusi relatif fitur terhadap keputusan model.
generalisasi model	Kemampuan model mempertahankan performa pada data baru yang tidak digunakan saat pelatihan.

<i>gap latih-validasi</i>	Selisih performa antara data latih dan data validasi.
<i>gap latih-uji</i>	Selisih performa antara data latih dan data uji.
<i>preprocessing</i>	Tahap awal pengolahan data sebelum model dilatih.
<i>cleaning</i>	Proses pembersihan data dari nilai tidak logis, duplikasi, atau inkonsistensi.
<i>encoding</i>	Proses mengubah data kategorikal menjadi representasi numerik.
<i>scaling</i>	Proses penyesuaian skala nilai fitur numerik.
<i>imputation</i>	Proses pengisian nilai hilang menggunakan aturan tertentu.
<i>pipeline</i>	Rangkaian proses transformasi data dan pelatihan model yang dijalankan secara terstruktur.
<i>train set</i>	Data latih yang digunakan untuk membangun model.
<i>validation set</i>	Data validasi yang digunakan untuk menilai model selama proses <i>tuning</i> .
<i>test set</i>	Data uji yang digunakan untuk evaluasi akhir model.
<i>hold-out data</i>	Data yang disisihkan dan tidak digunakan selama pelatihan maupun <i>tuning</i> .



<i>stratified split</i>	Teknik pembagian data dengan menjaga proporsi kelas tetap konsisten pada tiap <i>subset</i> .
<i>cross-validation</i>	Teknik evaluasi model dengan membagi data ke beberapa lipatan pengujian.
<i>Hyperparameter Tuning</i>	Proses pencarian kombinasi <i>hyperparameter</i> terbaik untuk meningkatkan performa model.
<i>hyperparameter</i>	Parameter model yang ditentukan sebelum proses pelatihan dimulai.
<i>random seed</i>	Nilai awal yang digunakan untuk menjaga konsistensi hasil acak.
eksperimen reproduktif	Eksperimen yang dapat dijalankan ulang dengan hasil yang konsisten.
<i>data leakage</i>	Kondisi ketika informasi dari data uji atau informasi target bocor ke proses pelatihan.
<i>overfitting</i>	Kondisi ketika model terlalu menyesuaikan diri dengan data latih sehingga kurang baik pada data baru.
<i>underfitting</i>	Kondisi ketika model terlalu sederhana sehingga gagal menangkap pola penting pada data.
interpretabilitas	Tingkat kemudahan model atau fitur untuk dijelaskan secara logis.



SimpleImputer	Komponen dari scikit-learn untuk mengisi nilai hilang.
StandardScaler	Komponen dari scikit-learn untuk standardisasi fitur numerik.
OneHotEncoder	Komponen dari scikit-learn untuk <i>encoding</i> variabel kategorikal.
ColumnTransformer	Komponen dari scikit-learn untuk menerapkan transformasi berbeda pada kelompok kolom yang berbeda.
<i>Pipeline</i>	Komponen dari scikit-learn untuk merangkai <i>preprocessing</i> dan model dalam satu alur.
<i>RandomForestClassifier</i>	Implementasi <i>Random Forest</i> untuk klasifikasi pada scikit-learn.
<i>LGBMClassifier</i>	Implementasi LightGBM untuk klasifikasi.
<i>CatBoostClassifier</i>	Implementasi CatBoost untuk klasifikasi.
pandas	<i>Library</i> Python untuk manipulasi dan analisis data tabular.
numpy	<i>Library</i> Python untuk komputasi numerik.
scikit-learn	<i>Library</i> Python untuk <i>machine learning</i> klasik.
matplotlib	<i>Library</i> Python untuk visualisasi data.
seaborn	<i>Library</i> Python untuk visualisasi statistik berbasis matplotlib.
cibil_score	Skor CIBIL



INTISARI

Penelitian ini membahas perbandingan empat algoritma *machine learning*, yaitu *Logistic Regression*, *Random Forest*, *LightGBM*, dan *CatBoost* pada klasifikasi kelayakan pinjaman nasabah menggunakan *Loan Approval Prediction Dataset* dari Kaggle. Permasalahan utama pada penelitian ini adalah menentukan algoritma yang paling efektif untuk memprediksi status persetujuan pinjaman, sekaligus memastikan bahwa rancangan eksperimen terbebas dari *data leakage* dan tidak menunjukkan *overfitting* yang berlebihan. Tahapan penelitian meliputi pembersihan data, rekayasa fitur, *exploratory data analysis*, pembagian data secara stratified ke dalam subset train, validasi, dan uji, pelatihan model *baseline*, serta *fine tuning hyperparameter*. Evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*, serta analisis gap performa antara data latih, validasi, dan data uji. Hasil penelitian menunjukkan bahwa *CatBoost* hasil *fine tuning* memberikan performa terbaik dengan *accuracy* 0,9977, *F1-score* 0,9981, dan *ROC-AUC* 1,0000 pada data uji. *Random Forest* menempati posisi berikutnya, diikuti oleh *LightGBM* dan *Logistic Regression*. Tingginya akurasi model terutama dipengaruhi oleh karakter *dataset* yang sangat mudah dipisahkan, terutama karena dominasi variabel *cibil_score*. Penelitian ini memberikan kontribusi berupa komparasi model yang disusun secara metodologis lebih ketat dan dapat dimanfaatkan sebagai referensi akademik maupun praktis pada pengembangan sistem prediksi kelayakan pinjaman berbasis data.

Kata kunci: Risiko Kredit, *Machine Learning*, Klasifikasi Pinjaman, Analisis Komparasi, Institusi Keuangan.

ABSTRACT

This study compares four machine learning algorithms, namely Logistic Regression, Random Forest, LightGBM, and CatBoost, for loan eligibility classification using the Loan Approval Prediction Dataset from Kaggle. The main objective is to identify the most effective algorithm for predicting loan approval status while ensuring that the experimental design is protected from data leakage and excessive overfitting. The research stages include data cleaning, feature engineering, exploratory data analysis, stratified train-validation-test splitting, baseline training, and hyperparameter fine tuning. Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, along with train-validation-test gap analysis. The results show that the tuned CatBoost model achieved the best performance with 0.9977 accuracy, 0.9981 F1-score, and 1.0000 ROC-AUC on the test set. Random Forest ranked second, followed by LightGBM and Logistic Regression. The very high accuracy was mainly driven by the highly separable structure of the dataset, especially the dominant role of `cibil_score`. This study contributes a tighter experimental comparison framework and can serve as an academic and practical reference for data-driven loan eligibility prediction systems.

Keyword: *Credit Risk, Machine Learning, Loan Classification, Comparative Analysis, Financial Institutions.*