ANALISIS OVERFITTING PADA PREDIKSI RISIKO GAGAL BAYAR PINJAMAN MENGGUNAKAN ALGORITMA DECISION TREE

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1-Informatika

disusun oleh

OXA DEFRIZAL KHASAY

21.11.4388

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

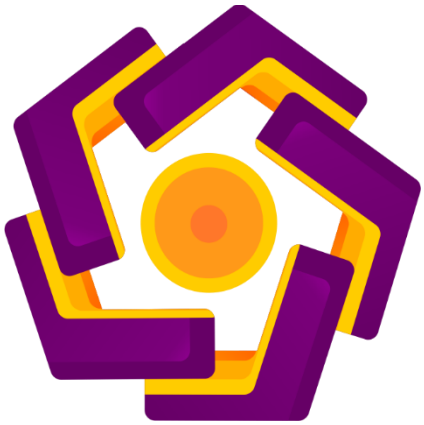
2025

ANALISIS OVERFITTING PADA PREDIKSI RISIKO GAGAL BAYAR PINJAMAN MENGGUNAKAN ALGORITMA DECISION TREE

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1-Informatika



disusun oleh

OXA DEFRIZAL KHASAY

21.11.4388

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

HALAMAN PERSETUJUAN

SKRIPSI

ANALISIS OVERFITTING PADA PREDIKSI RISIKO GAGAL BAYAR PINJAMAN MENGGUNAKAN ALGORITMA DECISION TREE

yang disusun dan diajukan oleh

Oxa Defrizal Khasay

21.11.4388

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 Januari 2025

Dosen Pembimbing,

Yuli Astuti, S.Kom., M.Kom.
NIK. 190302146

2

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

SKRIPSI

ANALISIS OVERFITTING PADA PREDIKSI RISIKO GAGAL BAYAR PINJAMAN MENGGUNAKAN ALGORITMA DECISION TREE

yang disusun dan diajukan oleh

Oxa Defrizal Khasay

21.11.4388

Telah dipertahankan di depan Dewan Penguji
pada tanggal 24 Januari 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Dr. Ferry Wahyu Wibowo, S.Si., M.Cs. **NIK. 190302235**	
Hastari Utama, S.Kom., M.Cs. **NIK. 190302230**	
Yuli Astuti, S.Kom., M.Cs. **NIK. 190302146**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 Januari 2025

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta,S.Kom., M.Kom., Ph.D.
NIK. 190302096

3

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Oxa Defrizal Khasay
NIM : 21.11.4388

Menyatakan bahwa Skripsi dengan judul berikut:

Analisis Overfitting pada Prediksi Risiko Gagal Bayar Pinjaman Menggunakan Algoritma Decision Tree

Dosen Pembimbing : Yuli Astuti, S.Kom., M.Kom.

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, 24 Januari 2025

Yang Menyatakan,

Oxa Defrizal Khasay

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat, nikmat, dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul *"Analisis Overfitting pada Prediksi Risiko Gagal Bayar Pinjaman Menggunakan Algoritma Decision Tree"* ini dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Informatika, Universitas Amikom Yogyakarta.

1. Allah SWT, atas segala rahmat, kemudahan, dan kekuatan yang diberikan selama proses pengerjaan skripsi ini.
2. Ibu Dosen Pembimbing, Yuli Astuti M.Kom, yang telah memberikan bimbingan, arahan, dan masukan yang sangat berarti selama proses penyusunan skripsi ini.
3. Kedua orang tua, yang selalu mendoakan, memberikan dukungan moral maupun material, serta menjadi sumber motivasi utama dalam setiap langkah perjalanan penulis.
4. Rekan-rekan mahasiswa angkatan 2021, yang senantiasa memberikan semangat, inspirasi, dan kebersamaan selama masa perkuliahan maupun proses penyelesaian skripsi ini.
5. Komunitas Kaggle dan Nikhil, yang menyediakan data serta sumber daya yang sangat membantu dalam proses penelitian ini.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan dan jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan saran dan kritik yang membangun untuk perbaikan karya ini di masa mendatang.

Akhir kata, penulis berharap semoga skripsi ini dapat memberikan manfaat, baik bagi perkembangan ilmu pengetahuan maupun bagi pembaca yang tertarik dengan bidang analisis data dan pembelajaran mesin.

Yogyakarta, 24 Januari 2025

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMBANG DAN SINGKATAN

ALLKNN	All K-Nearest Neighbors
Bagging	Bootstrap Aggregating
CART	Classification and Regression Trees
CHAID	Chi-Square Automatic Interaction Detection
EDA	Exploratory Data Analysis
ENN	Edited Nearest Neighbors
FN	False Negative
FP	False Positive
IBM	International Business Machines Corporation
IMDb	Internet Movie Database
IPA	Ilmu Pengetahuan Alam
LSTM	Long Short-Term Memory
ML	Machine Learning
MSI	Minimum Surfeit and Inaccuracy
PUBLIC	Projection and Unification-Based Learning for Interpretable Classification
RENN	Repeated Edited Nearest Neighbors
RMSProp	Root Mean Square Propagation
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
SMA	Sekolah Menengah Atas
SMOTE	Synthetic Minority Over-sampling Technique
TP	True Positive
TN	True Negative
SNMPTN	Seleksi Nasional Masuk Perguruan Tinggi Negeri
WEB	World Wide Web
K	Jumlah lipatan (Folds) dalam cross-validation
$Akurasi_i$	Akurasi model pada lipatan ke-i
S	Himpunan kasus
A	Atribut

N	Jumlah Partisi atribut A
|Si|	Jumlah kasus pada partisi ke-i
|S|	Jumlah kasus dalam S
N	Jumlah partisi S
Pi	Proporsi dari Si terhadap S

DAFTAR ISTILAH

ALLKNN	Teknik *undersampling* untuk mengurangi *noise* dalam data.
Average Error	Kesalahan rata-rata yang dihitung selama proses evaluasi model.
Bagging	Teknik *ensemble* untuk meningkatkan stabilitas dan akurasi model.
Bayesian Optimization	Metode untuk mengoptimalkan *hyperparameter* model secara efisien.
Bayes Risk Post-Pruning	Teknik *post-pruning* untuk mengurangi *overfitting* dalam pohon keputusan.
Biner	Sistem numerik yang hanya menggunakan dua nilai, 0 dan 1.
Box Plot	Grafik yang digunakan untuk menggambarkan distribusi data dan mendeteksi *outlier*.
C4.5	Algoritma *Decision Tree* yang merupakan pengembangan dari algoritma ID3.
CART	Algoritma *Decision Tree* yang digunakan untuk klasifikasi dan regresi.
CHAID	Algoritma *Decision Tree* berbasis statistik Chi-squared.
Churn	Istilah yang mengacu pada pelanggan yang berhenti menggunakan layanan.
Confusion Matrix	Matriks untuk mengevaluasi performa model klasifikasi dengan membandingkan prediksi dan data aktual.
Criterion	Parameter dalam *Decision Tree* yang menentukan cara memisahkan data, seperti *Gini* atau *Entropy*.

Cross Validation	Teknik evaluasi model dengan membagi dataset menjadi beberapa bagian untuk melatih dan menguji model.
Data Mining	Proses menemukan pola atau informasi penting dari data besar.
Data Preprocessing	Proses membersihkan dan mempersiapkan data sebelum digunakan dalam model.
Data Testing	Bagian dataset yang digunakan untuk menguji performa model.
Data Training	Bagian dataset yang digunakan untuk melatih model.
Dataset	Kumpulan data yang digunakan untuk analisis atau pelatihan model.
Debt-to-Income Ratio	Rasio utang terhadap pendapatan yang menunjukkan kemampuan seseorang untuk membayar utang.
Decision Tree	Algoritma pembelajaran mesin berbasis pohon keputusan untuk klasifikasi dan regresi.
Drop Column	Proses menghapus kolom tertentu dari dataset.
Dropout	Teknik regulasi untuk mencegah *overfitting* dengan mengabaikan beberapa neuron selama pelatihan.
EDA	Analisis awal untuk memahami karakteristik dan distribusi data.
Ensemble Bagging	Teknik *ensemble* untuk meningkatkan performa model melalui *bagging*.
Entropy	Ukuran ketidakpastian dalam data yang digunakan untuk membangun *Decision Tree*.
ENN	Teknik *undersampling* untuk menghilangkan *noise* dari data.

F1-Score	Rata-rata harmonis antara presisi dan *recall* yang digunakan untuk mengevaluasi model pada data tidak seimbang.
Feature (X)	Variabel independen dalam dataset yang digunakan sebagai input untuk model.
Filter Noise	Proses menghapus data yang mengganggu atau tidak relevan dalam dataset.
Forecasting	Proses memperkirakan nilai atau tren masa depan berdasarkan data yang ada.
Gain	Ukuran yang digunakan untuk menentukan atribut terbaik dalam pembagian pada *Decision Tree*.
Gini	Ukuran ketidakmurnian dalam data yang digunakan dalam *Decision Tree*
Grid Search	Metode untuk menemukan kombinasi terbaik dari *hyperparameter* dengan mencoba setiap kemungkinan.
Hyperparameter	Parameter yang harus ditentukan sebelum proses pelatihan model.
Hyperparameter Tuning	Proses mengoptimalkan *hyperparameter* untuk meningkatkan performa model.
IBM	Perusahaan teknologi yang menyediakan berbagai dataset untuk analisis dan pengembangan model.
Imbalance Data	Ketidakseimbangan dalam distribusi data antar kelas.
Instance	Satu entri atau baris data dalam dataset.
Integer	Jenis data numerik yang tidak memiliki nilai desimal
K-Fold Cross-Validation	Teknik *cross-validation* yang membagi dataset menjadi K bagian untuk melatih dan menguji model secara bergantian.

Kaggle	Platform komunitas untuk berbagi dataset dan kompetisi *data science*
Label (y)	Variabel dependen dalam dataset yang menjadi target prediksi model.
Label Encoding	Teknik mengubah data kategorikal menjadi bentuk numerik
LSTM	Jenis jaringan saraf tiruan untuk data berurutan seperti teks dan waktu.
Maximum Error	Kesalahan terbesar yang ditemukan dalam prediksi model.
Min_Samples_Leaf	Parameter dalam *Decision Tree* yang menentukan jumlah minimum sampel di setiap daun.
Min_Samples_Split	Parameter dalam *Decision Tree* yang menentukan jumlah minimum sampel untuk memisahkan simpul.
Model	Representasi matematis yang digunakan untuk membuat prediksi berdasarkan data.
MSI	Teknik untuk menghindari *overfitting* tanpa menggunakan *hyperparameter*.
Naive Bayes	Algoritma klasifikasi berbasis teorema *Bayes* dengan asumsi independensi antar fitur.
Node	Titik dalam *Decision Tree* yang mewakili atribut atau kondisi
Node Daun	Titik akhir dalam *Decision Tree* yang menunjukkan kelas prediksi.
Noise	Data yang tidak relevan atau mengganggu dalam dataset.
Object	Jenis data yang berisi string atau nilai kategorikal.
Outlier	Data yang berbeda secara signifikan dari data lain dalam dataset.

Pie Chart	Grafik melingkar untuk menunjukkan proporsi data
Post-Pruning	Proses mengurangi ukuran *Decision Tree* setelah pohon selesai dibangun untuk mencegah overfitting.
Pre-Pruning	Proses menghentikan pembangunan *Decision Tree* lebih awal untuk mencegah overfitting.
Pruning	Teknik untuk mengurangi kompleksitas *Decision Tree* dengan menghapus cabang yang tidak relevan.
PUBLIC	Algoritma pohon keputusan yang sering digunakan untuk klasifikasi.
Random Forest	Algoritma ensemble yang terdiri dari beberapa *Decision Tree* untuk meningkatkan akurasi.
Random Search	Metode mencari *hyperparameter* terbaik dengan mencoba kombinasi secara acak.
Random_State	Parameter yang memastikan hasil pembagian data atau pengacakan konsisten.
Recall	Kemampuan model untuk mendeteksi semua data positif yang benar.
RENN	Teknik untuk menghapus *noise* dari data melalui beberapa iterasi.
RMSProp	Optimizer untuk jaringan saraf tiruan yang memperbaiki *learning rate* secara adaptif.
Scikit-Learn	Library Python untuk pembelajaran mesin dan analisis data.
SMOTE	Teknik untuk menyeimbangkan data dengan membuat sampel sintetis dari kelas minoritas.
Splitting	Proses membagi dataset menjadi bagian pelatihan dan pengujian.
Sqrt	Akar kuadrat, digunakan sebagai parameter dalam *Decision Tree*.

Standard Deviation	Ukuran seberapa jauh data tersebar dari nilai rata-rata.
Supervised Machine Learning	Jenis pembelajaran mesin di mana model dilatih menggunakan data berlabel.
Train-Test Split	Teknik membagi dataset menjadi data pelatihan dan pengujian dengan rasio tertentu.
Validation Loss	Kesalahan model pada data validasi selama pelatihan.
Variabel Dependen	Variabel target yang diprediksi oleh model.
Variabel Independen	Variabel yang digunakan sebagai input untuk model.
Web	Platform untuk mengakses informasi secara online.
Weight Decay	Teknik regulasi untuk mencegah overfitting dengan menambahkan penalti pada bobot besar dalam model.
X_test	Data fitur untuk pengujian model.
X_train	Data fitur untuk melatih model.
y_pred_test	Hasil prediksi model pada data pengujian.
y_pred_train	Hasil prediksi model pada data pelatihan.
y_test	Label sebenarnya pada data pengujian.
y_train	Label sebenarnya pada data pelatihan.

INTISARI

Pertumbuhan ekonomi suatu negara dapat terganggu oleh berbagai permasalahan, seperti pengangguran, inflasi, dan ketimpangan pendapatan. Dalam konteks perekonomian Indonesia, tingkat pengangguran yang tinggi sering kali mendorong terjadinya tindakan kriminal, termasuk penipuan dan penggelapan uang di sektor keuangan. Salah satu bentuk kejahatan yang sering terjadi adalah gagal bayar pinjaman, yang tidak hanya merugikan lembaga keuangan, tetapi juga memengaruhi nasabah baru yang beritikad baik. Teknologi *machine learning* telah diterapkan secara luas dalam mendukung pengambilan keputusan, khususnya dalam prediksi risiko gagal bayar pinjaman. Salah satu permasalahan utama dalam pengembangan model prediksi adalah *overfitting*, yaitu kondisi ketika model terlalu menyesuaikan data pelatihan sehingga kinerjanya menurun pada data baru. Penelitian ini bertujuan untuk menganalisis dan mengurangi *overfitting* pada model prediksi gagal bayar pinjaman menggunakan algoritma *Decision Tree*. Metode penelitian meliputi pengumpulan data dari *platform Kaggle*, eksplorasi data untuk memahami karakteristik dataset, *preprocessing* untuk membersihkan dan menyiapkan data, serta pembagian data menggunakan *train-test split* dan *k-fold cross-validation*. Selain itu, *tuning* parameter dilakukan melalui *Grid Search* untuk menemukan kombinasi parameter terbaik. Hasil penelitian menunjukkan bahwa kombinasi teknik SMOTE, dan *hyperparameter tuning* mampu menghasilkan model dengan akurasi 90,06%, presisi 90,52%, *recall* 90,06%, f*1-score* 90.03%, dan perbedaan kecil antara akurasi pelatihan dan uji. Penerapan teknik SMOTE terbukti efektif dalam menangani ketidakseimbangan data, yang berkontribusi pada peningkatan kinerja model. Penggunaan *k-fold cross-validation* juga berperan penting dalam memastikan stabilitas model saat memprediksi risiko gagal bayar pinjaman.

Kata kunci: *Overfitting*, *Decision Tree*, Prediksi, SMOTE, *Hyperparameter Tuning*

ABSTRACT

Economic growth in a country can be disrupted by various issues, such as unemployment, inflation, and income inequality. In the context of Indonesia's economy, the high unemployment rate often leads to criminal activities, including fraud and embezzlement in the financial sector. One of the most common crimes is loan default, which not only harms financial institutions but also affects new borrowers with good intentions. Machine learning technology has been widely applied to support decision-making, particularly in predicting loan default risk. One of the main challenges in developing predictive models is overfitting, a condition where the model overly fits the training data, resulting in poor performance on new data. This study aims to analyze and mitigate overfitting in predicting loan default risk using the Decision Tree algorithm. The research methodology includes data collection from the Kaggle platform, data exploration to understand dataset characteristics, preprocessing to clean and prepare the data, and data splitting using train-test split and k-fold cross-validation. Additionally, parameter tuning was conducted using Grid Search to find the best parameter combinations. The results show that the combination of SMOTE and hyperparameter tuning achieved a model with an accuracy of 90.06%, precision of 90.52%, recall of 90.06%, and an F1-score of 90.03%, with a minimal difference between training and testing accuracy. The application of SMOTE proved effective in handling imbalanced data, contributing to improved model. The use of k-fold cross-validation also played a crucial role in ensuring model stability when predicting loan default risk.

Keyword**:** *Overfitting, Decision Tree, Prediction,* SMOTE*, Hyperparameter Tuning*