Jurnal EurekaMatika
Journal homepage: https://ejournal.upi.edu/index.php/JEM

Metode *Hybrid Decision Tree – Adaptive Boosting* pada Klasifikasi *Credit Scoring*

Nurul Aini, Dewi Rachmatin, Entit Puspita*

Program Studi Matematika, Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam, Universitas Pendidikan Indonesia, Indonesia
*Correspondence: E-mail: nurulaini@upi.edu

A B S T R A K

Bank Indonesia mengungkapkan peningkatan penyaluran kredit baru oleh lembaga perbankan pada Agustus 2023. Semakin tinggi jumlah transaksi kredit, maka semakin tinggi risiko kredit bermasalah. Oleh karena itu, perusahaan pemberi kredit harus lebih cermat dalam memilih calon peminjam yang berkualitas agar dapat mengurangi risiko kredit. Salah satu cara dalam mengurangi risiko kredit adalah *credit scoring,* yakni suatu sistem penilaian risiko kredit yang banyak digunakan untuk membantu lembaga keuangan atau perusahaan pemberi kredit dalam mengevaluasi calon peminjam, baik individu ataupun perusahaan. Penelitian ini bertujuan untuk menentukan model terbaik dengan menghitung tingkat akurasi dari model *Decision Tree – AdaBoost* dan model *Logistic Regression – AdaBoost* dalam menentukan klasifikasi *credit scoring* pada perusahaan *Home Credit*. Hasil evaluasi model *Decision Tree – AdaBoost* menunjukkan performa terbaik dengan keseimbangan yang baik antara akurasi, *precision*, *recall*, *F1-Score*, dan *ROC-AUC*. Model ini berhasil mengungguli model *Logistic Regression – AdaBoost*. Tingkat akurasi model terbaik dari *Decision Tree – AdaBoost* dalam menentukan klasifikasi *credit scoring* pada perusahaan *Home Credit* sebesar 70% menunjukkan bahwa model *Decision Tree – AdaBoost* sudah cukup baik dalam menentukan klasifikasi *credit scoring*.

INFORMASI ARTIKEL

Sejarah Artikel:

Diterima 20 Oktober 2024
Direvisi 25 Oktober 2024
Disetujui 25 November 2024
Tersedia 1 Desember 2024
Dipublikasikan 1 Desember 2024

Kata Kunci:
AdaBoost,
Credit Scoring,
Klasifikasi,
Logistic Regression,
Pohon Keputusan.

A B S T R A C T

Bank Indonesia revealed indications of an increase in new credit disbursements by banking institutions in August 2023. As the number of credit transactions rises, the risk of problematic loans also increases. Therefore, credit providers must be more careful in selecting high-quality borrowers to reduce credit risk. One way to reduce credit risk is credit scoring, a widely used risk assessment system that helps financial institutions or credit providers evaluate potential borrowers, individuals, and companies. This study aims to identify the best model by calculating the accuracy levels of the Decision Tree C4.5 – AdaBoost model and the Logistic Regression – AdaBoost model in classifying credit scoring for Home Credit. The Decision Tree – AdaBoost model demonstrated the best performance, balancing accuracy, precision, recall, F1-Score, and ROC-AUC. This model outperformed the Logistic Regression - AdaBoost model. The accuracy level of the best Decision Tree – AdaBoost model in classifying credit scoring for Home Credit is 70%, indicating that the Decision Tree – AdaBoost model is quite effective in determining credit scoring classifications.

Keywords:
AdaBoost,
Classification,
Credit Scoring,
Decision Tree,
Logistic Regression.

1. PENDAHULUAN

Credit scoring merupakan suatu sistem penilaian risiko kredit yang banyak digunakan untuk membantu lembaga keuangan atau perusahaan pemberi kredit dalam mengevaluasi calon peminjam, baik individu ataupun perusahaan yang kemungkinan gagal melakukan pembayaran. *Credit scoring* dapat memutuskan apakah calon peminjam tersebut berhak mendapatkan kredit atau tidak. Tujuan dari *credit scoring* ini adalah untuk mengklasifikasikan calon peminjam berkualitas (*good credit*) dan calon peminjam tidak berkualitas (*bad credit*) yang kemungkinan gagal bayar (Bastos, 2022). *Credit scoring* telah menjadi faktor penting dalam industri keuangan, terutama dalam proses pengambilan keputusan pemberian kredit. *Credit scoring* membantu lembaga keuangan dalam mengevaluasi kemungkinan peminjam untuk membayar pinjaman secara tepat waktu yaitu dengan menganalisis data keuangan, perilaku pembayaran, dan informasi kredit lainnya. Oleh karena itu, sangat penting untuk mengembangkan model *credit scoring* yang dapat digunakan secara akurat dan efektif (Xiao, *et al*., 2021). Klasifikasi dalam *credit scoring* mencakup penentuan apakah seseorang layak atau tidak layak menerima pinjaman berdasarkan skor kredit mereka. Klasifikasi merupakan suatu bentuk analisis data yang mengekstrak model (atau fungsi) serta mendeskripsikan dan membedakan kelas atau konsep data. Klasifikasi juga dikenal sebagai *supervised learning* (Bastos, 2022).

Beberapa metode yang dapat digunakan untuk klasifikasi dalam *credit scoring* yaitu, *Decision Tree*, Naïve Bayes, *K-Nearest Neighbor* (K-NN), *Support Vector Machines* (SVM), *Neural Networks*, *Logistic Regression* dan *Ensemble Methods* (Bastos, 2022). Salah satu metode dalam klasifikasi yang sering digunakan adalah *Decision Tree*. *Decision Tree* merupakan salah satu metode klasifikasi yang populer karena algoritmanya yang sederhana sehingga mudah dipahami dan mampu menangani tipe data campuran. Selain *Decision Tree, Logistic Regression* juga merupakan salah satu metode klasifikasi yang memerankan peran penting dalam mengevaluasi strategi *machine learning* dengan algoritmanya yang mudah dimengerti (Shah, *et al*., 2020).

Beberapa penelitian yang melakukan *credit scoring* seperti yang dilakukan oleh (Naufal, *et al*., 2023) yang mengusulkan metode Naïve Bayes dan *Decision Tree* untuk memprediksi potensi hilangnya nasabah bank. Akurasi terbaik diperoleh dari metode *Decision Tree* dengan nilai akurasi mencapai 93%. Pada penelitian (Jadhav & Chane, 2016) *Decision Tree* dinilai lebih akurat, memiliki tingkat kesalahan yang lebih rendah, dan lebih mudah dipahami dibandingkan K-NN dan Naïve Bayes. Pada kasus mendetekti penipuan kartu kredit oleh (Alenzi & Aljehane, 2020) metode *Logistic Regression* memperoleh hasil akurasi tertinggi sebesar 97,2% dibandingkan dengan metode K-NN dan *Voting Classifier*. Selain itu, metode *Logistic Regression* pada penelitian (Silva, *et al*., 2020) tentang *credit scoring* lembaga keuangan Portugis memperoleh hasil akurasi sebesar 89,79%.

Akan tetapi, model *Decision Tree* mudah dipengaruhi oleh *data noise* dan atribut data yang berlebihan sehingga membuat model *Decision Tree* menjadi lemah. Sedangkan pada model *Logistic Regression* kesulitan untuk bekerja dengan baik pada kumpulan data yang tidak seimbang (Zhang & Chen, 2021). Untuk mengatasi masalah tersebut dapat menggunakan metode *ensemble* seperti *Random Forest*, *Bagging* (*Bootstrap Aggregating*), *AdaBoost* (*Adaptive Boosting*), dan *Gradien Boosting* (Wang, *et al*., 2012). Wang, *et al*., (2012) mengusulkan metode RS-Bagging DT dan Bagging-RS DT yang didasarkan pada dua strategi *ensemble* yaitu *bagging* dan *random subspace*. Hasil penelitian ini terbukti dapat meningkatkan akurasi *Decision Tree* dibandingkan dengan *Decision Tree* tunggal. Berdasarkan Chopra & Bhilare (2018) metode *ensemble* juga terbukti mampu meningkatkan akurasi pada

Decision Tree. Bastos (2022) membahas mengenai prediksi *credit scores* menggunakan *Boosted Decision Tree* dengan yang menjadi Boostednya itu adalah metode *AdaBoost*. Hasil dari penelitian tersebut membuktikan bahwa metode *AdaBoost* mampu meningkatkan akurasi *Decision Tree* dan tingkat keakuratan *Boosted Decision Tree* ini lebih tinggi daripada *Multilayer Perceptron* dan *Support Vector Machine* (SVM). Selain dapat meningkatkan nilai akurasi dari *Decision Tree, AdaBoost* juga dapat meningkatkan nilai akurasi dari *Logistic Regression* (Zhang & Chen, 2021). Hal ini menunjukkan bahwa metode *AdaBoost* merupakan sebuah teknik yang kompetitif untuk meningkatkan kinerja model *Decision Tree* dan *Logistic Regression*. Oleh karena itu, berdasarkan pemaparan tersebut, peneliti tertarik mengkaji lebih dalam mengenai *credit scoring* menggunakan *ensemble learning* dengan membandingkan metode *Decision Tree – AdaBoost* dan *Logistic Regression – AdaBoost* dengan menggunakan variabel yang lebih beragam dan kompleks.

2. METODE

Metode yang digunakan dan dibahas pada penelitian ini adalah *Decision Tree*, *Logistic Regression*, dan *Adaptive Boosting*. Data yang digunakan dalam penelitian ini diperoleh dari situs web *Kaggle* (https://www.kaggle.com/). Penelitian ini menggunakan 356.255 data peminjam dan 122 variabel atau atribut riwayat kredit peminjam.

2.1 *Decision Tree*

Salah satu metode klasifikasi yang populer adalah *Decision Tree* karena algoritmanya yang sederhana sehingga mudah dipahami dan mudah diinterpretasikan (Lee & Cheang, 2021). *Decision Tree* adalah struktur pohon yang berakar dan terarah, mirip dengan diagram alur, di mana setiap simpul internal (*nonleaf node*) menunjukkan atribut, setiap cabang mewakili hasil pengujian, dan setiap simpul daun (*leaf node*) menunjukkan label kelas. Simpul paling atas dalam sebuah pohon adalah simpul akar (*root node*) (Bansal *et al*., 2022).

Decision Tree dapat menangani berbagai bentuk data baik itu data kategori maupun numerik, serta tidak memerlukan asumsi tentang distribusi data atau hubungan linier, sehingga dapat digunakan dalam berbagai aplikasi, termasuk yang melibatkan data non-linier. Tidak ada batasan khusus pada jumlah variabel, tetapi jika jumlah atribut terlalu banyak, *Decision Tree* dapat menjadi terlalu kompleks (*overfitting*). Oleh karena itu, diperlukan pemilihan fitur atau teknik *pruning* untuk mengurangi kompleksitas dan meningkatkan kinerja model (Bansal *et al*., 2022). Menurut Hssina *et al.,* (2014), Algoritma *Decision Tree* secara umum dinyatakan pada Persamaan (1), (2) dan (3) sebagai berikut:

$$Entropy(X) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

$$Gain(A) = Entropy(X) - \sum_{i=1}^{k} \frac{|X_i|}{|X|} . Entropy\ (X_i) \quad (2)$$

di mana:

$$Entropy(X_i) = -\sum_{i,j=1}^{m} p_{ij} \log_2(p_{ij}) \quad (3)$$

Keterangan:

X : Himpunan data pelatihan
X_i : Subset dari X yang berisi sampel pada kategori ke-i dari atribut A
A : Himpunan atribut
m : Jumlah kelas dalam himpunan data
k : Jumlah kategori pada atribut A
p_i : Proporsi sampel dalam kelas ke-i dari himpunan data X
p_{ij} : Proporsi sampel dalam kelas ke-j dari subset X_i
$|X|$: Jumlah total sampel dalam himpunan data X pada atribut A
$|X_i|$: Jumlah sampel pada kategori ke-i dari atribut A
$Entropy\ (X_i)$: *Entropy* dari subset X_i, yang berisi sampel dari kategori ke-i atribut A

2.2 *Logisric Regression* (Regresi Logistik)

Regresi logistik adalah salah satu metode analisis statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dengan variabel dependen yang bersifat kategorikal (Muflihah, 2017). Regresi logistik dibagi menjadi tiga jenis yaitu regresi logistik biner, regregi logistik multinomial, dan regresi logistik ordinal. Regresi logistik biner digunakan ketika variabel dependen memiliki dua kategori, regresi logistik multinomial digunakan ketika variabel dependen memiliki lebih dari dua kategori, dan regresi logistik ordinal digunakan ketika variabel dependen berskala ordinal. Dataset pada Regresi Logistik harus seimbang dalam hal jumlah observasi untuk setiap kategori dari variabel dependen. Jika satu kategori jauh lebih banyak daripada yang lain (misalnya, banyak "tidak" dan sedikit "ya"), ini dapat menyebabkan model bias dan mengurangi akurasi prediksi. Regresi logistik menggunakan fungsi logit untuk menghubungkan variabel dependen dengan satu atau lebih variabel independen. Bentuk umum persamaan regresi logistik biner menurut (Silva, *et al.*, 2020) pada Persamaan (4) sebagai berikut.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_r x_r}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_r x_r}} \tag{4}$$

Di mana:
$\pi(x)$: probabilitas variabel independent
β_i : parameter-parameter regresi logistik dan perlu diestimasi, untuk $i = 1, 2, 3, \dots, r$
x_i : pengamatan variabel independen, untuk $i = 1, 2, 3, \dots, r$

Ukuran yang digunakan untuk menginterpretasikan hubungan antara variabel independen dan variabel dependen biner dalam regresi logistik disebut dengan *Odds Ratio*. *Odds Ratio* mengukur seberapa besar perubahan rasio antara probabilitas terjadinya suatu peristiwa dengan probabilitas tidak terjadinya peristiwa tersebut dari kejadian terjadi untuk setiap unit perubahan dalam variabel independen. *Odds Ratio* dinotasikan dengan OR, rasio probabilitas $x = 1$ dengan $x = 0$, dan diberikan oleh Persamaan (5) sebagai berikut.

$$OR = \frac{odds_1}{odds_2} = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \frac{e^{\beta_0+\beta_i}}{e^{\beta_0}} = e^{\beta_i} \tag{5}$$

2.3 Adaptive Boosting (AdaBoost)

AdaBoost adalah algoritma yang fleksibel dan dapat diterapkan pada berbagai jenis data. *AdaBoost* dirancang untuk digunakan pada data yang sudah terlabeli, baik untuk klasifikasi biner maupun multiklas. *AdaBoost* memperoleh model dengan menggabungkan sekumpulan

pengklasifikasi yang lemah untuk membuat pengklasifikasi yang kuat (Freund & Schapire, 1997). Prosedur *AdaBoost* akan dibahas secara rinci sebagai berikut.

Diberikan himpunan sampel pelatihan:

$$D = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_i, Y_i)\}, i = 1,2, \dots, n, Y_i \in \{+1, -1\}$$

1. Inisialisasi distribusi bobot yang sama pada setiap sampel dalam data pelatihan untuk bobot awal dinyatakan oleh Persamaan (6) berikut

$$W_1(i) = \frac{1}{N}, i = 1,2, \dots, N \tag{6}$$

2. Lakukan iterasi dengan mengulangi langkah a – d berikut ini sampai M iterasi atau sampai diperoleh tingkat kesalahan yang rendah, dengan M adalah jumlah iterasi dan H adalah pengklasifikasi dasar.
   a. Mempelajari pengklasifikasi dasar H_m pada setiap iterasi m dengan data latih D dan distribusi bobot W_m untuk mendapatkan pengklasifikasi yang lemah sebagaimana pada Persamaan (7) berikut.

$$H_m : \mathcal{Y} \to \{+1, -1\} \tag{7}$$

   b. Hitung tingkat kesalahan klasifikasi ε_m pada Persamaan (8) berikut

$$\varepsilon_m = \sum_{i=1}^{N} W_m(i)\, I(H_m(x_i) \neq y_i) \tag{8}$$

   c. Hitung nilai α_m pada Persamaan 9 berikut

$$\alpha_m = \frac{1}{2} ln \frac{1 - \varepsilon_m}{\varepsilon_m} \tag{9}$$

   d. Perbarui distribusi bobot himpunan pelatihan untuk iterasi selanjutnya dinyatakan pada Persamaan (10) berikut

$$W_{m+1}(i) = \frac{W_m\, exp\big(-\alpha_m y_i H_m(x_i)\big)}{Z_m} \tag{10}$$

   dengan

$$Z_m = \sum_{i=1}^{N} W_m(i')\, exp\big(-\alpha_m y_i' H_m(x'_i)\big) \tag{11}$$

   dimana Z_m adalah faktor normalisasi untuk memastikan Persamaan (12) berikut

$$\sum_{i=1}^{N} W_{m+1}(i) = 1 \tag{12}$$

   Jika sampel diklasifikasikan salah, bobotnya akan meningkat pada iterasi berikutnya. Sebaliknya, jika sampel diklasifikasikan dengan benar, bobotnya akan berkurang.
3. Gabungkan pengklasifikasi yang lemah untuk menjadi pengklasifikasi yang kuat sebagaimana ditampilkan pada Persamaan (13) berikut

$$H(x) = \sum_{m=1}^{M} \alpha_m H_m(x) \tag{13}$$

2.4 Evaluasi Model

Mengevaluasi kinerja model klasifikasi dapat menggunakan *confusion matrix*. *Confusion matrix* adalah metode untuk mengevaluasi kinerja model klasifikasi dengan menghitung tingkat akurasi pada model yang digunakan. Berikut disajikan Tabel 1 *Confusion Matrix* menurut (Shah, *et al.*, 2020).

Tabel 1. *Confusion Matrix*

		Prediction Class	
		Positive	*Negative*
Actual Class	*Positive*	TP *(True Positive)*	FN *(False Negative)*
	Negative	FP *(False Positive)*	TN *(True Negative)*

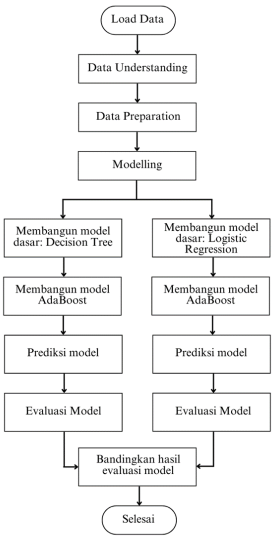
dimana:

1. TP (*True Positive*) adalah hasil kelas prediksi benar (positif) dan kelas sebenarnya juga benar (positif).
2. TN (*True Negative*) adalah hasil kelas prediksi salah (negatif) dan kelas sebenarnya juga salah (negatif).
3. FP (*False Positive*) adalah hasil kelas prediksi salah tetapi kelas sebenarnya benar.
4. FN (*False Negative*) adalah hasil kelas prediksi salah tetapi kelas sebenarnya positif.

Confusion matrix dapat menghitung nilai *accuracy*, *precision*, *recall*, *F1-Score* dan juga *ROC-AUC*.

2.5 Alur Penelitian

Alur penelitian metode *Decision Tree – AdaBoost* dan *Logistic Regression – AdaBoost* ditampilkan pada Gambar 1.



Gambar 1. Alur penelitian metode *Decision Tree* – AdaBoost dan *Logistic Regression – AdaBoost*

3. HASIL DAN PEMBAHASAN

Data yang digunakan pada penelitian ini yaitu data dari perusahaan *Home Credit* yang diperoleh dari situs web *Kaggle* (https://www.kaggle.com/). Penelitian ini menggunakan 356.255 data peminjam dan 122 variabel atau atribut riwayat kredit peminjam. Variabel yang digunakan sebagai variabel dependen adalah variabel "TARGET" dan 121 variabel lainnya merupakan variabel independen. Penelitian ini juga menerapkan metode CRISP-DM (*Cross Industry Standard Process for Data Mining*) yaitu sebuah metode yang digunakan dalam proses data *mining* (Brzozowska *et al*., 2023).

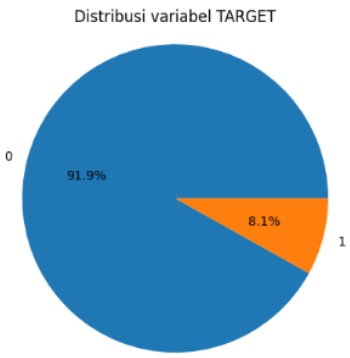
3.1 *Data Understanding*

Tahap ini merupakan tahap pemahaman data mulai dari mengumpulkan data awal, mendeskripsikan data, mengeksplorasi data, dan memverifikasi kualitas data. Pada Tabel 2 terdapat lima contoh kolom variabel riwayat peminjam yaitu "SK_ID_CURR" merupakan ID peminjam; "TARGET" merupakan karakteristik peminjam dalam membayar kredit di mana 1 = peminjam yang memiliki kesulitan pembayaran dan 0 = peminjam yang lancar membayar kredit; "NAME_CONTRACT_TYPE" merupakan kategori pinjaman apakah "*cash loans*" atau "*revolving loans*"; "CODE_GENDER" merupakan jenis kelamin peminjam di mana M = laki-laki dan F = perempuan; "FLAG_OWN_CAR" merupakan informasi peminjam mempunyai mobil (Y) atau tidak (N). dari 122 variabel, 16 variabel bertipe data kategori, 40 variabel bertipe data bilangan bulat, dan 66 variabel bertipe data kontinu.

Tabel 2. Contoh Data Penelitian

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR
0	100002	1	Cash loans	M	N
1	100003	0	Cash loans	F	N
2	100004	0	Revolving loans	M	Y
3	100006	0	Cash loans	F	N
4	100007	0	Cash loans	M	N

Terlihat pada Gambar 2 bahwa distribusi data pada variabel TARGET tidak seimbang. Sebanyak 91,9% peminjam lancar membayar pinjaman dan sebanyak 8,1% peminjam kesulitan membayar pinjaman.



Gambar 2. Distribusi Variabel TARGET

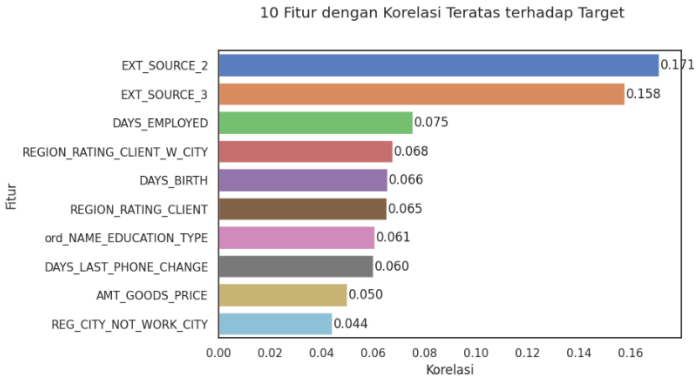
Dalam proses *data mining*, penting untuk mengenali dan memahami *unique values* yaitu nilai-nilai yang unik atau kategori-kategori dan berbeda dalam suatu variabel. Pada penelitian ini terdapat *unique value* "XNA" pada variabel "CODE_GENDER", *unique value* "Unknown" pada variabel "NAME_FAMILY_STATUS", *unique value* "XNA" pada variabel "ORGANIZATION_TYPE", dan *unique value* "365243 (nilai positif)" pada variabel "DAYS_EMPLOYED". Selain itu, terdapat data yang mempunyai *missing value* (nilai yang hilang) sebanyak 62 variabel untuk variabel numerik dan 6 variabel untuk variabel kategori.

3.2 *Data Preparation*

Data preparation (persiapan data) merupakan langkah prapemrosesan data yang mencakup pembersihan data, transformasi data, dan mengintegrasikan data. Berdasarkan

Data Understanding tersebut, penanganan yang tepat untuk variabel yang memiliki *missing value* > 30% yaitu dengan menghapus kolom variabel tersebut. Setelah itu, variabel kategorik akan ditransformasi menjadi variabel numerik. Hal ini dilakukan untuk memudahkan dalam pengolahan data. Pada saat mentransformasi data terdapat perbedaan. Untuk variabel yang bertipe data ordinal akan ditransformasi dengan diberi peringkat, sedangkan variabel yang bertipe data nominal akan ditransformasi berdasarkan dengan kategorinya.

Peneliti menggunakan analisis korelasi untuk memilih fitur terbaik karena korelasi dapat mengevaluasi hubungan antara variabel dependen dan variabel independen. Berikut ini sepuluh variabel terbaik dengan korelasi tertinggi terhadap variabel "TARGET" ditampilkan pada Gambar 3.



Gambar 3. Sepuluh Variabel Terbaik

3.3 *Modelling* dan Evaluasi

Pada tahap pemodelan dimulai dari memilih teknik atau metode pemodelan yang akan digunakan lalu membangun model. Penelitian ini memilih untuk menggunakan algoritma *Decision Tree*, *Logistic Regression*, dan dilanjutkan dengan algoritma *AdaBoost*. Proses pemodelan dalam penelitian ini menggunakan bantuan *Google Collab* dalam bahasa pemrograman *Python*. Berikut ini merupakan hasil dari evaluasi model *Decision Tree – AdaBoost* dan *Logistic Regression – AdaBoost* ditampilkan pada Tabel 3.

Tabel 3. Hasil Evaluasi Model

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree – AdaBoost	0,69	0,67	0,71	0,69	0,69
Logistic Regression – AdaBoost	0,54	0,55	0,30	0,39	0,54

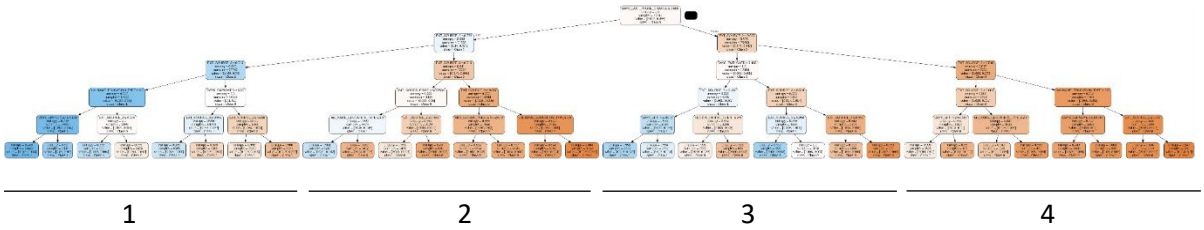
Berdasarkan hasil evaluasi model yang telah dilakukan, *Decision Tree – AdaBoost* menunjukkan performa terbaik dengan keseimbangan yang baik antara *precision*, *recall*, *F1-Score*, dan *ROC-AUC*. Model ini berhasil mengungguli model *Logistic Regression – AdaBoost.* Sebaliknya, *Logistic Regression* menunjukkan performa yang kurang memuaskan, dengan *precision*, *recall*, dan *F1-Score* yang rendah.

Untuk lebih meningkatkan performa model *Decision Tree – AdaBoost*, langkah selanjutnya adalah melakukan *hyperparameter tuning*. *Hyperparameter tuning* adalah proses mengoptimalkan parameter-parameter model yang tidak dapat dipelajari dari data selama pelatihan, dengan tujuan untuk menemukan kombinasi parameter yang menghasilkan performa terbaik.

Berikut hasil dari evaluasi model *Decision Tree – AdaBoost* setelah dilakukan *hyperparameter tuning*:

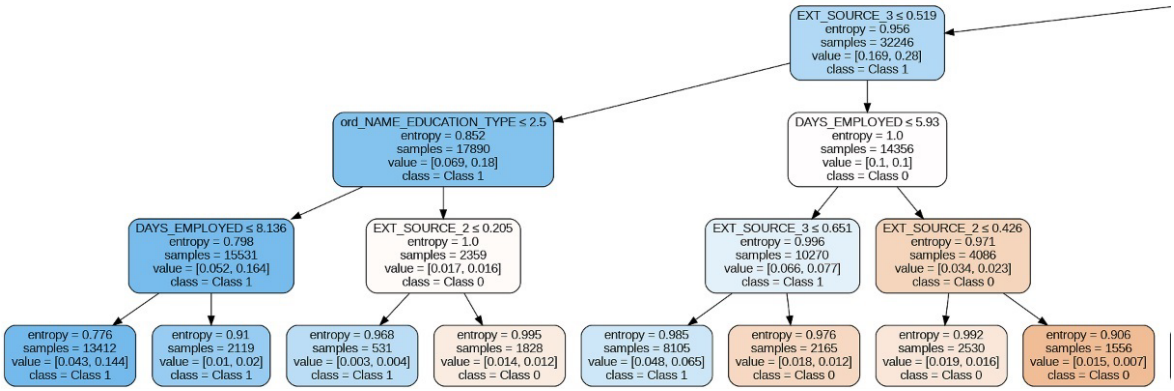
- *Accuracy*: 0,70, artinya model dapat mengklasifikasikan 70% dari data dengan benar.
- *Precision*: 0,68, artinya semua prediksi positif yang dibuat oleh model, 68% diantaranya benar.
- *Recall*: 0,71, artinya model ini dapat mendeteksi 71% dari semua data yang benar-benar positif.
- *F1-Score*: 0,69, merupakan rata-rata harmonis dari *precision* dan *recall*, yang menunjukkan keseimbangan antara keduanya.
- *ROC-AUC*: 0,70, artinya model memiliki kemampuan untuk membedakan antara kelas positif dan negatif dengan tingkat keberhasilan 70%.

Setelah diperoleh hasil model terbaik dari *Decision Tree – AdaBoost*, langkah berikutnya adalah menganalisis struktur pohon keputusan yang dihasilkan. Hal ini bertujuan untuk memahami variabel-variabel yang paling penting dalam mempengaruhi klasifikasi *credit scoring*. Variabel-variabel penting ini tidak hanya membantu dalam memahami kinerja model, tetapi juga dalam membuat keputusan bisnis yang lebih informatif dan strategis terkait penilaian kredit. Berikut Gambar 4 menampilkan pohon keputusan dari model terbaik *Decision Tree – AdaBoost*.

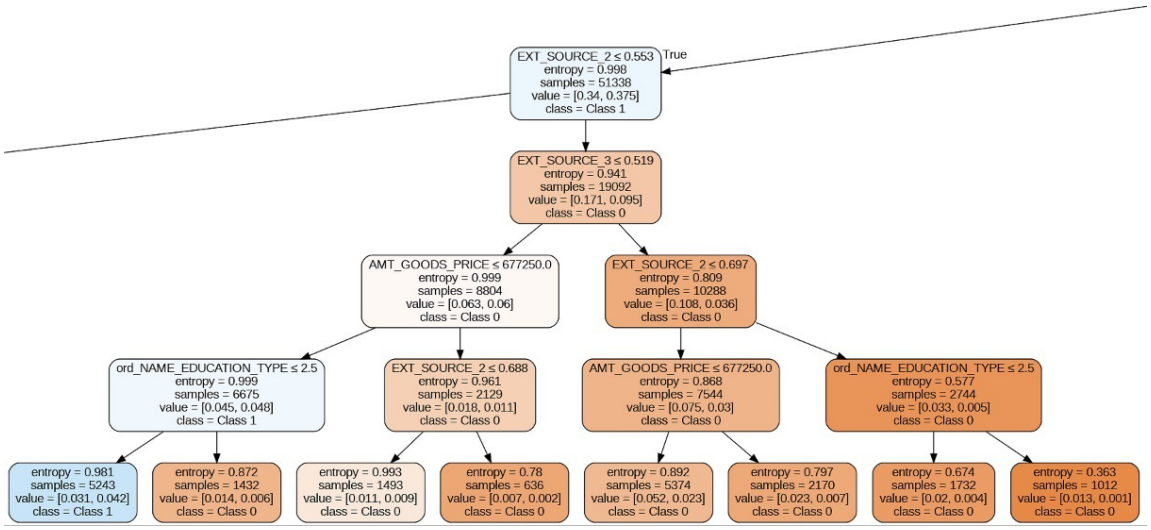


Gambar 4. Pohon Keputusan

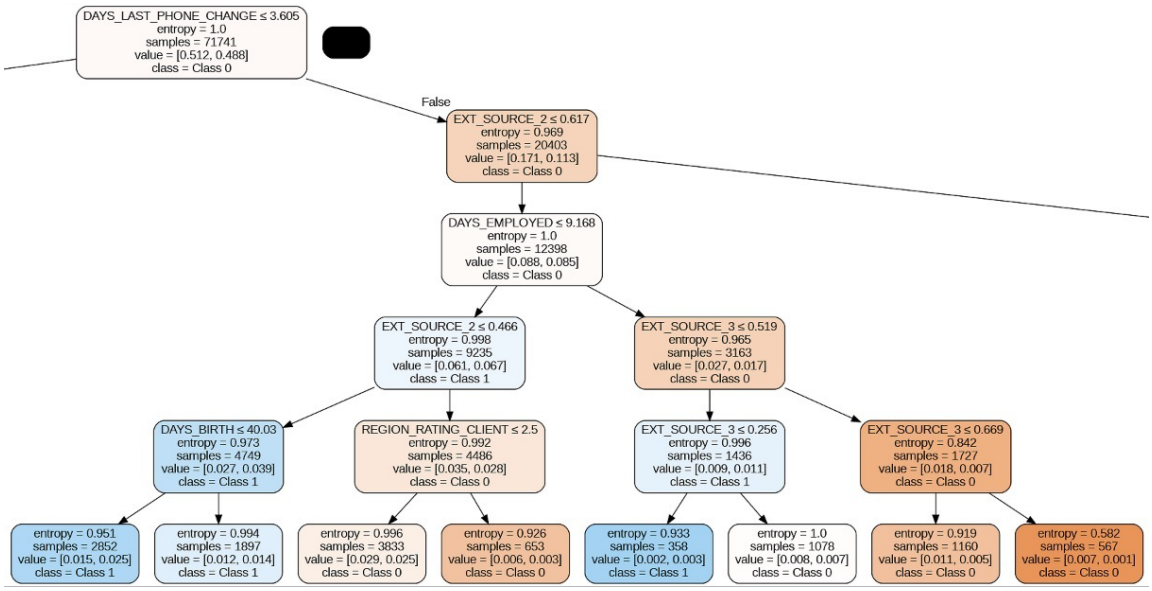
Dari gambar 4, untuk memperjelas gambar pohon keputusan tersebut, akan dibagi menjadi empat bagian sebagai ditampilkan pada Gambar 5 - 8 berikut.



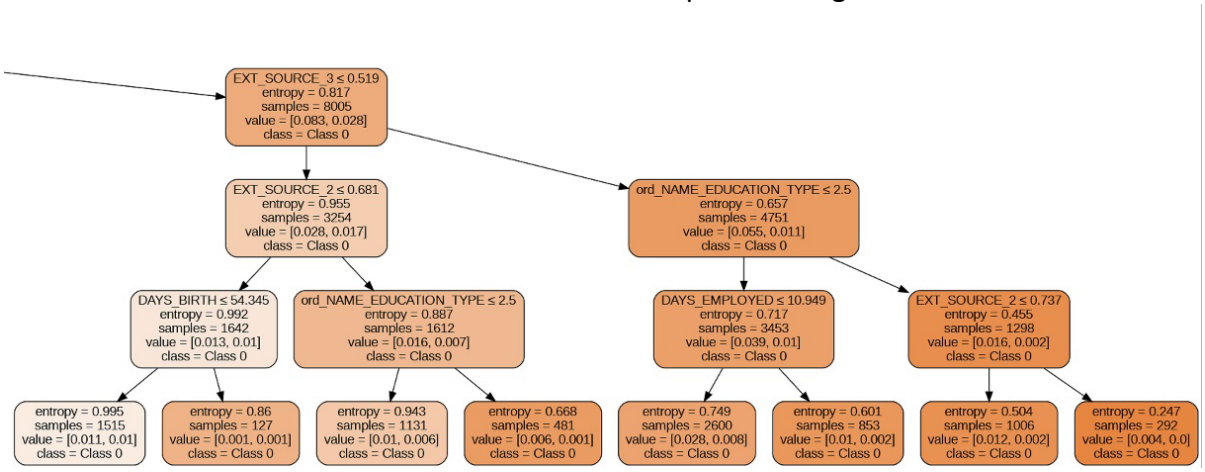
Gambar 5. Gambar Pohon Keputusan Bagian 1



Gambar 6. Gambar Pohon Keputusan Bagian 2



Gambar 7. Gambar Pohon Keputusan Bagian 3



Gambar 8. Gambar Pohon Keputusan Bagian 4

Dari sepuluh variabel terbaik yang dipilih untuk pemodelan, dapat dilihat dari hasil gambar pohon keputusan tersebut, terdapat lima variabel yang mempengaruhi peminjam gagal melakukan pembayaran kredit, di mana kelima variabel tersebut diklasifikasikan menjadi kelas 1.

1. EXT_SOURCE_2 dan EXT_SOURCE_3 (Informasi Keuangan Peminjam dari Sumber Eksternal 2 dan 3):
   Ini merupakan skor risiko dari sumber luar seperti informasi mengenai pengelolaan keuangan peminjam seperti riwayat pembayaran utang, jumlah utang yang dimiliki, riwayat kredit dan lain sebagainya. Skor yang rendah menunjukkan bahwa peminjam mungkin berisiko lebih tinggi untuk gagal bayar.
2. DAYS_EMPLOYED (Lama Bekerja):
   Jika peminjam baru saja mulai bekerja atau memiliki riwayat pekerjaan yang tidak stabil, mereka cenderung berisiko lebih tinggi untuk gagal bayar.
3. DAYS_BIRTH (Usia):
   Usia peminjam juga menjadi salah satu faktor penting. Peminjam yang berusia sangat muda atau sangat tua cenderung memiliki risiko lebih tinggi untuk gagal bayar.
4. NAME_EDUCATION_TYPE (Tingkat Pendidikan):
   Peminjam dengan tingkat pendidikan yang lebih rendah berisiko lebih tinggi untuk gagal bayar.

4. KESIMPULAN

Model *Decision Tree – AdaBoost* menunjukkan performa terbaik dengan keseimbangan yang baik antara akurasi, *precision*, *recall*, *F1-Score*, dan *ROC-AUC*. Tingkat akurasi model terbaik dari *Decision Tree – AdaBoost* dalam menentukan klasifikasi *credit scoring* pada perusahaan *Home Credit* yaitu sebesar 70% yang menunjukkan bahwa model *Decision Tree – AdaBoost* sudah cukup baik dalam menentukan klasifikasi *credit scoring*. Pohon keputusan dari model terbaik menunjukkan lima variabel yang signifikan dalam mempengaruhi klasifikasi *credit scoring*, yaitu EXT_SOURCE_2, EXT_SOURCE_3, DAYS_EMPLOYED, NAME_EDUCATION_TYPE, dan DAYS_BIRTH. Hal ini menunjukkan bahwa penilaian risiko kredit sangat dipengaruhi oleh skor eksternal yang berasal dari sumber-sumber terpercaya, stabilitas pekerjaan yang mencerminkan kemampuan pembayaran yang konsisten, tingkat pendidikan yang berkaitan dengan potensi penghasilan, dan usia peminjam yang mempengaruhi profil risiko mereka.

5. DAFTAR PUSTAKA

Alenzi, H. Z., & Aljehane, N. O. (2020). Fraud detection in credit cards using logistic regression. *International Journal Of Advanced Computer Science And Applications (Ijacsa), 11*(12), 540-551.

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal, 3*(100071), 1-21.

Bastos, J. A. (2022). Predicting credit scores with boosted decision trees. *Forecasting, 4*, 925-935.

Brzozowska, J., Pizon, J., Baytikenova, G., Gola, A., Zakimova, A., & Piotrowska, K. (2023). Data

engineering in crisp-dm process production data – case study. *Applied Computer Science, 19*(3), 83-95.

Chopra, A., & Bhilare, P. (2018). Application of ensemble models in credit scoring models. *Business Perspectives and Research, 6*(2), 129-141.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications, 4*(2), 13-19.

Jadhav, S. D., & Chane, H. P. (2016). Comparative Study of K-NN, naive bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR), 5*(1), 1842-1845.

Lee, C. S., & Cheang, P. Y. (2021). Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences, 26*(1), 1-30.

Muflihah, I. Z. (2017). Analisis financial distress perusahaan manufaktur di Indonesia dengan regresi logistik. *Majalah Ekonomi*, *22*(2), 254-269.

Naufal, M. F., Subrata, Susanto, A. F., Kansil, C. N., & Huda, S. (2023). Analisis perbandingan algoritma machine learning untuk prediksi potensi hilangnya nasabah bank. *Techno.com, 22*(1), 1-11.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research, 5*(1), 12.

Silva, E. C., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics, 47*(13), 2879-2894.

Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems, 26*, 61-68.

Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. *Information Sciences, 569*, 508-526.

Zhang, X., & Chen, X. (2021). Research on breach prediction for big data through hybrid ensemble learning and logistic regression. *Journal of Physics: Conference Series, 1982*(2021).