

METODE *HYBRID DECISION TREE* – *ADAPTIVE BOOSTING*
(Studi Kasus: Klasifikasi *Credit Scoring*)

SKRIPSI

Diajukan untuk memenuhi sebagian syarat untuk memperoleh gelar
Sarjana Matematika



Oleh:

Nurul Aini

NIM 2003693

PROGRAM STUDI MATEMATIKA
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
2024

LEMBAR HAK CIPTA

METODE *HYBRID DECISION TREE – ADAPTIVE BOOSTING*
(Studi Kasus: Klasifikasi *Credit Scoring*)

Oleh:
Nurul Aini
2003693

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memperoleh
gelar Sarjana Matematika
pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

©Nurul Aini 2024
Universitas Pendidikan Indonesia
Agustus 2024

Hak Cipta dilindungi Undang-Undang
Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian dengan dicetak ulang,
difotokopi atau cara lainnya tanpa izin dari penulis

LEMBAR PENGESAHAN

NURUL AINI

METODE *HYBRID DECISION TREE* – *ADAPTIVE BOOSTING*

(Studi Kasus: Klasifikasi *Credit Scoring*)

Disetujui dan disahkan oleh:

Pembimbing I,



Dr. Hj. Entit Puspita, S.Pd., M.Si.

NIP. 196704081994032002

Pembimbing II,



Hj. Dewi Rachmatin, S.Si., M.Si.

NIP. 196909291994122001

Mengetahui,

Ketua Prodi Matematika



Dr. Kartika Yulianti, S.Pd., M.Si.

NIP. 198207282005012001

ABSTRAK

Bank Indonesia mengungkapkan bahwa adanya indikasi peningkatan dalam penyaluran kredit baru oleh lembaga perbankan pada bulan Agustus 2023. Semakin tinggi jumlah transaksi kredit, maka akan semakin tinggi juga 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 yaitu dengan melakukan *credit scoring*. *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. 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*. Berdasarkan hasil dari evaluasi model, 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* yaitu sebesar 70% yang menunjukkan bahwa model *Decision Tree – AdaBoost* sudah cukup baik dalam menentukan klasifikasi *credit scoring*.

Kata Kunci: Klasifikasi, *Credit Scoring*, *Decision Tree*, *Logistic Regression*, *AdaBoost*

ABSTRACT

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 through credit scoring. Credit scoring is a widely used risk assessment system that helps financial institutions or credit providers evaluate potential borrowers, individuals, and companies, who may fail to repay their loans. This study aims to identify the best model by calculating the accuracy levels of the Decision Tree – AdaBoost model and the Logistic Regression – AdaBoost model in classifying credit scoring for Home Credit. Based on the model evaluation results, the Decision Tree – AdaBoost model demonstrated the best performance with a good balance between 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: *Clasification, Credit Scoring, Decision Tree, Logistic Regression, AdaBoost*

DAFTAR ISI

LEMBAR HAK CIPTA.....	i
LEMBAR PENGESAHAN	ii
LEMBAR PERNYATAAN.....	iii
ABSTRAK.....	iiiv
<i>ABSTRACT</i>	v
KATA PENGANTAR	vi
UCAPAN TERIMA KASIH.....	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR	x
DAFTAR TABEL.....	xii
DAFTAR LAMPIRAN.....	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian	5
1.4 Manfaat Penelitian	5
BAB II KAJIAN PENELITIAN	7
2.1 Klasifikasi	7
2.2 Teori Probabilitas.....	7
2.3 <i>Decision Tree</i>	9
2.4 <i>Logistic Regression</i> (Regresi Logistik).....	14
2.5 <i>Boosting</i>	18
2.5.1 <i>Adaptive Boosting</i>	18
2.6 Evaluasi Model	20
2.7 Kredit	22
2.7.1 Unsur-Unsur Kredit	22
2.7.2 Tujuan dan Fungsi Kredit.....	23
2.7.3 Jenis-Jenis Kredit.....	24
2.7.4 Risiko Kredit.....	25
BAB III METODOLOGI PENELITIAN	28

3.1 Deskripsi Masalah.....	28
3.2 Data Penelitian	29
3.3 Langkah-Langkah Penelitian	29
3.4 Langkah-Langkah Pembentukan Model Gabungan.....	31
BAB IV HASIL DAN PEMBAHASAN	34
4.1 Analisis Data	34
4.2 Pemahaman Data (<i>Data Understanding</i>).....	35
4.2.1 Mendeskripsikan Data	35
4.2.2 Mengeksplorasi Data	37
4.2.3 <i>Exploratory Data Analysis (EDA)</i>	40
4.3 Persiapan Data (<i>Data Preparation</i>)	46
4.3.1 Pembersihan Data	47
4.3.2 Mentransformasi Data Kategorik (<i>Categorical Data Encoding</i>)	50
4.3.3 Pemilihan Fitur Data.....	51
4.4 <i>Modelling</i> dan Evaluasi.....	53
4.4.1 Menyeimbangkan Data dan Membagi Data	53
4.4.2 Membangun Model Dasar	54
4.4.3 Membangun Model <i>Hybrid</i>	55
4.4.4 Evaluasi Model	56
4.4.5 <i>Hyperparameter Tuning</i>	58
4.4.6 Pohon Keputusan dan Interpretasi	60
BAB V KESIMPULAN.....	65
5.1 Kesimpulan	65
5.2 Saran	65
DAFTAR PUSTAKA	67
LAMPIRAN.....	70

DAFTAR GAMBAR

Gambar 2. 1 Contoh General <i>Decision Tree</i> (Barros, et al., 2015).....	9
Gambar 2. 2 Kemungkinan partisi jika A bernilai diskrit.....	13
Gambar 2. 3 Contoh kemungkinan partisi jika A bernilai diskrit.....	13
Gambar 2. 4 Kemungkinan partisi jika A bernilai kontinu.....	13
Gambar 2. 5 Contoh kemungkinan partisi jika A bernilai kontinu.....	13
Gambar 2. 6 Kemungkinan partisi jika A bernilai diskrit dan membuat pohon biner	14
Gambar 2.7 Contoh kemungkinan partisi jika A bernilai diskrit dan membuat pohon biner	14
Gambar 3. 1 CRISP-DM.....	30
Gambar 3. 2 Diagram alir penelitian.....	33
Gambar 4. 1 Tipe Data.....	35
Gambar 4. 2 Distribusi Variabel TARGET	36
Gambar 4. 3 Contoh <i>Unique Value</i>	37
Gambar 4. 4 Visualisasi <i>missing value</i>	39
Gambar 4. 5 Diagram Batang Variabel “NAME_TYPE_CONTRACT”	40
Gambar 4. 6 Diagram Batang Variabel “CODE_GENDER”	41
Gambar 4. 7 Diagram Batang Variabel “FLAG_OWN_CAR”	41
Gambar 4. 8 Diagram Batang Variabel “FLAG_OWN_REALTY”	42
Gambar 4. 9 Diagram Batang Variabel “NAME_INCOME_TYPE”	42
Gambar 4. 10 Diagram Batang Variabel “NAME_EDUCATION_TYPE”	43
Gambar 4. 11 Diagram Batang Variabel “NAME_FAMILY_STATUS”	43
Gambar 4. 12 Diagram Batang Variabel “NAME_HOUSING_TYPE”	44
Gambar 4. 13 Diagram Batang Variabel “OCCUPATION_TYPE”	45
Gambar 4. 14 Diagram Batang Variabel “NAME_TYPE_SUITE”	45
Gambar 4. 15 Diagram Batang Variabel “ORGANIZATION_TYPE”	46
Gambar 4. 16 Menghapus kolom yang memiliki <i>missing value</i> > 30% dan variabel yang tidak relevan	47
Gambar 4. 17 Menghapus <i>Unique Value</i>	47
Gambar 4. 18 Tipe-Tipe Kategori Variabel “NAME_TYPE_SUITE”	48

Gambar 4. 19 Mengisi <i>Missing Value</i> Variabel “NAME_TYPE_SUITE”	49
Gambar 4. 20 Gambar Distribusi Variabel Numerik	49
Gambar 4. 21 Tidak ada data yang hilang	50
Gambar 4. 22 Hasil Korelasi.....	51
Gambar 4. 23 Sepuluh Variabel Terbaik	53
Gambar 4. 24 Data sebelum SMOTE dan sesudah SMOTE	53
Gambar 4. 25 Model <i>Decision Tree</i>	54
Gambar 4. 26 Model <i>Logistic Regression</i>	55
Gambar 4. 27 Model <i>hybrid Decision Tree – AdaBoost</i>	55
Gambar 4. 28 Parameter Terbaik dari model <i>Decision Tree – AdaBoost</i>	58
Gambar 4. 29 Pohon Keputusan	61
Gambar 4. 30 Gambar Pohon Keputusan Bagian 1	62
Gambar 4. 31 Gambar Pohon Keputusan Bagian 2	62
Gambar 4. 32 Gambar Pohon Keputusan Bagian 3	63
Gambar 4. 33 Gambar Pohon Keputusan Bagian 4	63

DAFTAR TABEL

Tabel 2. 1 <i>Confusion Matrix</i> (Han, et al., 2012)	20
Tabel 4. 1 Deskripsi Data Penelitian.....	34
Tabel 4. 2 Contoh Data yang akan digunakan	35
Tabel 4. 3 Variabel Numerik.....	36
Tabel 4. 4 Variabel Kategori.....	37
Tabel 4. 5 Data yang memiliki <i>missing value</i>	38
Tabel 4. 6 Variabel yang Hilang	48
Tabel 4. 7 Transformasi variabel “NAME_EDUCATION_TYPE”	50
Tabel 4. 8 Hasil Transformasi Data	51
Tabel 4. 9 Hasil Evaluasi Model.....	56

DAFTAR LAMPIRAN

Lampiran 1. Deskripsi Kolom Data Penelitian	70
Lampiran 2. Variabel Numerik	77
Lampiran 3. Variabel Kategori	79
Lampiran 4. Jumlah <i>Missing Value</i> Setiap Kolom	80
Lampiran 5. Korelasi	83
Lampiran 6. <i>Coding</i> proses klasifikasi.....	85

DAFTAR PUSTAKA

- Abdullah, T., & Wahjusaputri, S. (2018). *Bank dan Lembaga Keuangan*. Jakarta: Mitra Wacana Media.
- Aggarwal, C. C. (2015). *Data Classification Algorithm and Applications*. New York: CRC Press.
- 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). doi:<https://doi.org/10.14569/ijacsa.2020.0111265>
- Ambarsita, L. (2013). Analisis Penanganan Kredit Macet. *Manajemen Bisnis*, 3(1).
- Amin, R. K., Indwiarti, I., & Sibaroni, Y. (2015). Implementasi Klasifikasi Decision Tree Dengan Algoritma C4. 5 Dalam Pengambilan Keputusan Permohonan Kredit Oleh Debitur (Studi Kasus: Bank Pasar Daerah Istimewa Yogyakarta). *eProceeding of Engineering*, 2(1).
- Barros, R. C., De Carvalho, A. C., & Freitas, A. A. (2015). *Automatic design of decision-tree induction algorithms*. Cham, Switzerland: Springer.
- Bastos, J. A. (2022). Predicting Credit Scores with Boosted Decision Trees. *Forecasting*, 4, 925-935. doi:<https://doi.org/10.3390/forecast4040050>
- 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. doi: 10.35784/acs-2023-26
- Budianto, E. W. (2023). Pemetaan Penelitian Seputar Risiko Kredit pada Perbankan Syariah dan Konvensional: Studi Bibliometrik VOSviewer dan Literature Review. *BANCO: Jurnal Manajemen dan Perbankan Syariah*, 5(1), 20-34. doi:<https://doi.org/10.35905/banco.v5i1.4987>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16,321-357. doi:<https://doi.org/10.1613/jair.953>
- Chopra, A., & Bhilare, P. (2018). Application of Ensemble Models in Credit Scoring Models. *Business Perspectives and Research*, 6(2), 129-141. doi:DOI: 10.1177/2278533718765531

- Daniel, T. L., & Chantal, D. L. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. United States of America: John Wiley & Sons, Inc.
- Djuarni, W., & Ratnasari, R. (2022). Implementasi Prinsip 5C Dalam Menentukan Kelayakan Pemberian Kredit Pada Nasabah. *Ar-Rihlah: Jurnal Keuangan dan Perbankan Syariah*, 2(2), 99-113.
- Eprianti, N. (2019). Penerapan Prinsip 5C Terhadap Tingkat Non Performing Finance (NPF). *Amwaluna: Jurnal Ekonomi dan Keuangan Syariah*, 3(2), 252-266. doi:<https://doi.org/10.29313/amwaluna.v3i2.4645>
- Firmanto, F. (2019). Penyelesaian Kredit Macet di Indonesia. *Jurnal Pahlawan*, 2(2). doi:<https://doi.org/10.31004/jp.v2i2.577>
- 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. doi:<https://doi.org/10.1006/jcss.1997.1504>
- Garson, G. D. (2014). *Logistic Regression: Binary & Multinomial*. Asheboro: Statistical Associates Publishing.
- Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics*. New York: McGraw-Hill/Irwin.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third Edition*. Waltham, MA: Morgan Kaufmann.
- Herrhyanto, N., & Gantini, T. (2009). *Pengantar Statistika Matematis*. Bandung, Indonesia: Yrama Widya.
- Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). *Introduction to Mathematical Statistics*. Boston MA: Pearson.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken, New Jersey: John Wiley & Sons.
- 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). doi:<https://doi.org/10.21275/v5i1.nov153131>

- Naufal, M. F., Subrata, Susanto, A. F., Kansil, C. N., & Huda, S. (2023). Analisis Perbandingan Algoritma Machine Learning untuk. *Techno.COM*, 22(1), 1-11. doi:<http://publikasi.dinus.ac.id/index.php/technoc/article/view/7302>
- 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(12). doi:<https://doi.org/10.1007/s41133-020-00032-0>
- Silva, E. C., Lopes, I. C., Correia, A., & Faria, S. (2020). A Logistic Regression Model for Consumer Default Risk. *Journal of Applied Statistics*. doi:<https://doi.org/10.1080/02664763.2020.1759030>
- Sriwati, N. K. (2017). Analisis Tingkat Risiko Kredit Pada Bank Rakyat Indonesia Cabang Poso. *Jurnal EKOMEN*, 17(1).
- Susilawati, & Putri, D. (2019). Analisis Pengaruh Transaksi Non Tunai Dan Suku Bunga Bi Terhadap Pertumbuhan Ekonomi Di Indonesia. *Jurnal Kajian Ekonomi dan Pembangunan*, 1(2), 667-678. doi:<http://dx.doi.org/10.24036/jkep.v1i2.6294>
- Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procdia Computer Science*, 174, 150-160. doi:<https://doi.org/10.1016/j.procs.2020.06.070>
- 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. doi:<https://doi.org/10.1016/j.knosys.2011.06.020>
- 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. doi:<https://doi.org/10.1016/j.ins.2021.05.029>
- 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). doi: <https://doi.org/10.1088/1742-6596/1982/1/012049>