MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 8 | No. 1 | Halaman 42 - 51
DOI : https://doi.org/10.26760/mindjournal.v8i1.42-51 Juni 2023

RESULTANT: Data Preparation Techniques to Improve XGBoost Algorithm Performance

KURNIA RAMADHAN PUTRA, SOFIA UMAROH, NUR FITRIANTI, SATRIA NUGRAHA

Sistem Informasi, Institut Teknologi Nasional Bandung
Email: kurniaramadhan@itenas.ac.id

Received 30 November 201x **|** *Revised* 30 Desember 201x **|** *Accepted* 30 Januari 201x

ABSTRAK

Prediksi credit scoring saat ini banyak digunakan dalam layanan peer-to-peer lending oleh perusahaan teknologi finansial. Salah satu teknologi yang digunakan untuk credit scoring adalah data mining menggunakan algoritma machine learning XGBoost yang memiliki tingkat akurasi yang tinggi. RESULTANT diusulkan sebagai teknik yang digunakan untuk memaksimalkan hasil dari salah satu tahapan data mining yaitu preparasi data. Dataset yang digunakan adalah data Lending Club dengan total 2.260.701 record dan 151 variabel. Tahapan yang dilakukan pada RESULTANT adalah seleksi fitur, penanganan missing value, penanganan data outlier dan penanganan data ketidakseimbangan. Dari tahap RESULTANT, dihasilkan 44 variabel akhir yang siap digunakan untuk membangun model dengan menggunakan algoritma XGBoost. Hasil menunjukkan bahwa RESULTANT mampu meningkatkan performa algoritma XGBoost dengan akurasi 99,17%, presisi 99,28%, recall 99,05%, spesifisitas 99,29%, ROC/AUC 99,94%, dan skor f1 99,17%.

***Kata kunci**: XGBoost, Preparasi Data, Seleksi Fitur, Missing Value, Outlier*

ABSTRACT

Credit scoring predictions are currently widely used in peer-to-peer lending services by financial technology companies. One of the technologies used for credit scoring is data mining using the XGBoost machine learning algorithm which has a high degree of accuracy. We present RESULTANT as a technique used to maximize the results of one of the stages of data mining, namely data preparation. The dataset used is Lending Club data with a total of 2,260,701 records and 151 variables. The stages carried out in RESULTANT are feature selection, handling missing values, handling outlier data and handling imbalance data. From the RESULTANT stage, 44 final variables are produced which are ready to be used to build models using the XGBoost algorithm. The results showed that RESULTANT was able to improve the performance of the XGBoost algorithm with accuracy 99,17%, precision 99,28%, recall 99,05%, specificity 99,29%, ROC/AUC 99.94%, and f1-score 99,17%.

***Keywords**: XGBoost, Data Preparation, Feature Selection, Missing Value, Outlier*

1. PENDAHULUAN

Penilaian kredit adalah teknik yang digunakan beberapa lembaga keuangan khususnya teknologi finansial untuk mengevaluasi risiko kredit dan meningkatkan arus keuangan. Dalam prakteknya, penilaian kredit mengacu pada masalah klasifikasi dimana pemohon kredit baru harus dikategorikan ke dalam salah satu kelas yang ditentukan berdasarkan beberapa variabel atau atribut yang diamati yang menggambarkan karakteristik kondisi ekonomi nasabah tersebut. Setiap lembaga keuangan memiliki model penskoran untuk penilaian kreditnya sendiri yang biasanya menggunakan regresi logistik dan pohon keputusan, karena interpretasinya yang sederhana **(Abdykalykova, 2020)**.

Penggunaan credit scoring pada lembaga keuangan di Indonesia saat ini banyak digunakan oleh teknologi finansial yang lebih dikenal dengan *fintech* pada lembaga keuangan *peer-to-peer* lending. Keunggulan layanan *peer-to-peer lending* dibandingkan dengan layanan keuangan lainnya adalah tidak memerlukan jaminan dan proses waktu pencarian juga cepat. Penilaian kredit dalam pinjaman *peer-to-peer* harus akurat untuk mengurangi risiko kredit bahwa peminjam tidak akan membayar utangnya **(Risna, 2020)**.

Solusi yang yang dapat digunakan adalah data mining menggunakan salah satu algoritma machine learning yaitu XGBoost yang memiliki tingkat akurasi yang tinggi berdasarkan penelitian credit scoring sebelumnya yang dilakukan oleh Abdylalykova dalam penelitian credit scoring menggunakan beberapa algoritma machine learning **(Abdykalykova, 2020)**. Teknologi seperti pembelajaran mesin dalam penambangan data dapat membantu penilaian kredit menjadi lebih akurat dan efisien.

Tahapan dalam data mining terdiri dari pemahaman data, penyiapan data, pemodelan dan evaluasi. Preparasi data adalah proses pengumpulan, pembersihan, dan penyatuan data ke dalam file atau tabel data untuk tujuan analisis. Preparasi data melibatkan manipulasi data mentah yang tidak terstruktur menjadi bentuk yang lebih terstruktur yang siap untuk analisis lebih lanjut **(Kadav, A., dkk, 2013)**.

Pada penelitian diiusulkan kombinasi tahapan preparasi data yang dikenal dengan RESULTANT untuk menangani nilai data yang hilang, data pencilan, dan ketidakseimbangan data untuk meningkatkan kinerja dari algoritma XGBoost sehingga dapat memberikan hasil prediksi yang lebih baik terhadap penskoran kredit pada studi kasus *peer-to peer landing*.

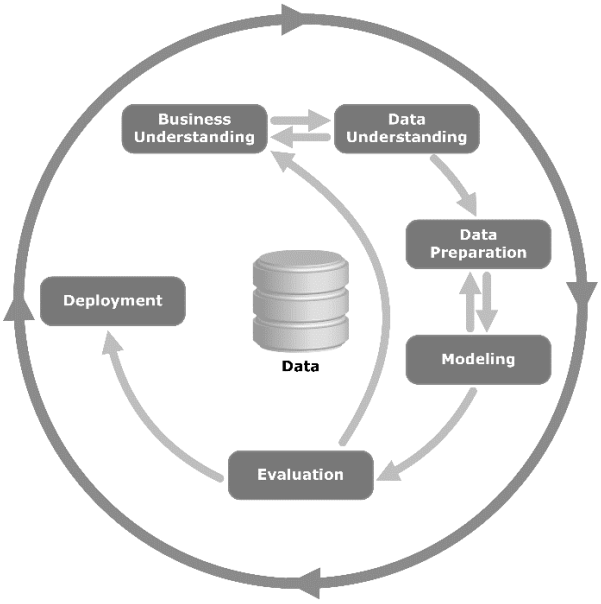
2. METODE PENELITIAN

2.1. CRISP-DM

CRISP-DM adalah singkatan dari *Cross-Industry Standard Process for Data Mining* yang telah terbukti digunakan oleh industri sebagai panduan dalam *data mining* **(Wirth, R., dkk, 2000)**.

Pada Gambar 1 menggambarkan siklus hidup dari data mining menggunakan standar CRISP-DM yang mana ada enam proses utama di dalamnya yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Tahapan dari CRISP-DM dijelaskan sebagai berikut:

a. *Business understanding*, yaitu pemahaman tujuan dan kebutuhan dari perspektif bisnis, kemudian mengubah pengetahuan dari pemahaman tersebut menjadi definisi masalah data mining dan rencana awal yang dirancang untuk mencapai tujuan tersebut.



Gambar 1. Siklus Hidup *Data Mining* (Sumber: IBM SPSS Modeler CRISP-DM Guide, 2020)

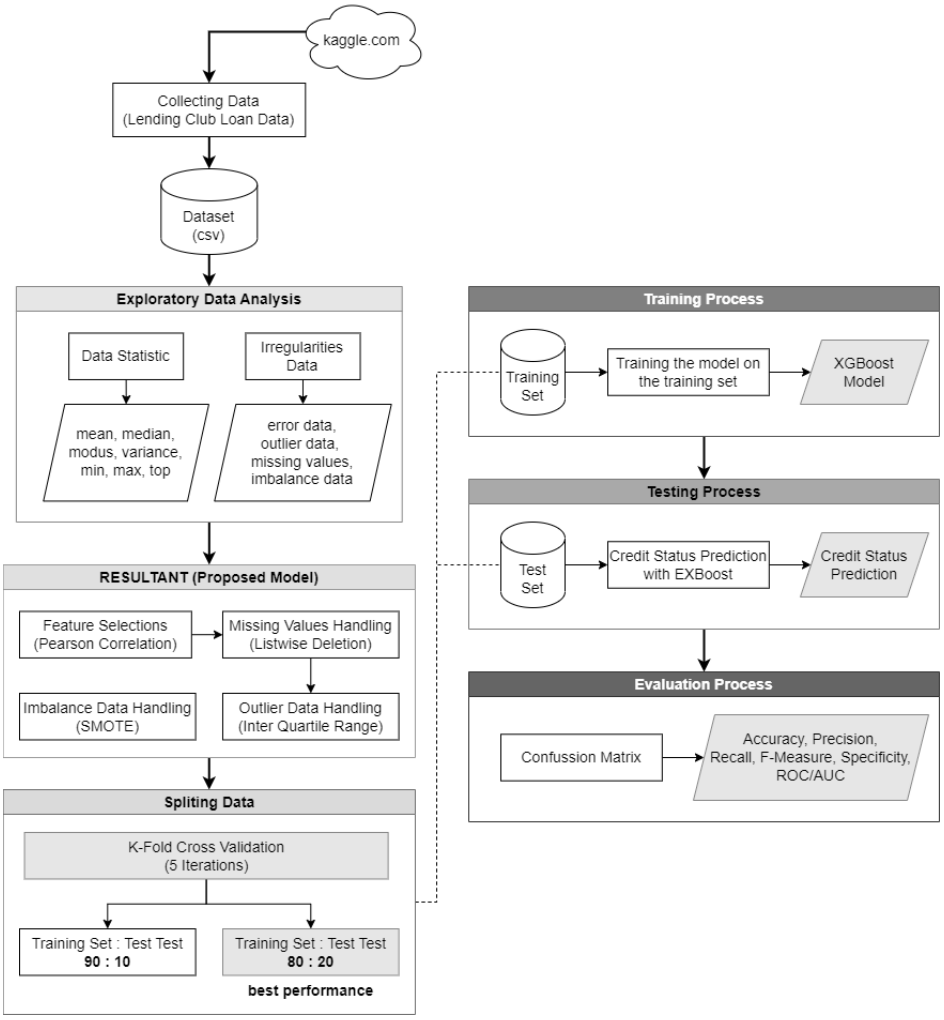
b. *Data understanding*, yaitu pengumpulan data mentah kemudian dideskripsikan dan dieksplorasi, mencari atribut target dan data tersebut diverifikasi.
c. *Data preparation*, yaitu mengubah data mentah yang masih memiliki nilai yang hilang, data pencilan, dan ketidakseimbangan data menjadi data yang bersih dan siap untuk digunakan dalam pemodelan.
d. *Modeling*, yaitu memilih teknik pemodelan dan menguji dengan data yang sudah diperoleh dari tahap preparasi sebelumnya.
e. *Evaluations*, yaitu mengukur pemodelan yang telah dipilih untuk memastikan model yang dibuat menghasilkan keluaran dengan performa yang baik.
f. *Deployment*, yaitu mempublikasikan model yang telah dievaluasi agari bisa digunakan pada data warehouse agar mendapatkan lebih banyak wawasan tentang data.

2.2. Pemahaman Data

Tahapan ini dilakukan dengan cara studi pustaka untuk memahami arti dari fitur-fitur yang ada pada dataset pinjaman di LendingClub kemudian menyeleksi fitur-fitur tersebut menjadi target berdasarkan penelitian **(Schröer, C., dkk, 2019)**. Dataset pinjaman pada LendingClub merupakan dataset publik yang bertujuan untuk menguji metode penelitian yang dikembangkan oleh peneliti, dataset pinjaman di Lending Club memiliki 151 kolom dengan 2.260.701 baris. Kelebihan dari dataset ini adalah memiliki banyak variabel dan atribut sehingga tingkat kebenarannya sangat baik, sedangkan kekurangan dari dataset ini adalah memiliki nama kolom yang asing dan harus mencari arti dari nama kolomnya tersebut terlebih dahulu. Ada 2 jenis data dalam satu kolom sehingga belum siap secara langsung digunakan untuk prediksi data.

2.3. RESULTANT: Teknik-Teknik Persiapan Data

Persiapan data adalah proses pengumpulan, pembersihan, dan konsolidasi data ke dalam file atau tabel data untuk keperluan analisis **(Abdallah, dkk, 2017)**. Persiapan data harus dilakukan karena data tidak konsisten, tidak terstruktur, adanya data pencilan, nilai yang hilang, dan ketidakseimbangan data. RESULTANT menggabungkan beberapa teknik dari persiapan data untuk menangani permasalahan tersebut.



Gambar 2. Teknik Preparasi Data RESULTANT

Listwise Deletion

Teknik yang digunakan dalam mengatasi data yang hilang dalam penelitian ini adalah *Complete-Case Analysis* (*Listwise Deletion*). Dalam teknik ini, kumpulan data dengan >50% data yang hilang akan dibuang. Dataset dengan data yang hilang di bawah 50% akan dikoreksi menggunakan teknik imputasi rata-rata atau median **(Hartini, 2016)**. Perbaikan data untuk tipe kategori adalah mengisi nilai kosong dengan teknik imputasi rata-rata fitur. Dalam data numerik, nilai kosong dapat diisi dengan rata-rata atau median.

Pearson Correlation

Teknik yang digunakan dalam menentukan korelasi fitur adalah *Pearson Correlation* dimana rentang nilai yang dihasilkan berada pada rentang -1 sampai dengan 1 yang menunjukkan sejauh mana dua variabel berhubungan secara linier. Fitur dengan korelasi tinggi lebih bergantung secara linier dan karenanya memiliki efek yang hampir sama pada variabel dependen. Jadi, ketika dua fitur memiliki korelasi yang tinggi, salah satu dari dua fitur tersebut dapat dihapus **(Samuels, 2014)**. Korelasi fitur menggunakan *Pearson Correlation*_dapat dihitung menggunakan Persamaan (1).

$$r_{xy} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i) \sqrt{n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)}}} \quad (1)$$

P-Value

Pengujian hipotesis menggunakan *P-Value* untuk menentukan apakah sebuah hipotesis benar atau salah. *P-Value* berada pada rentang nilai antara 0 dan 1. Tingkat signifikansi merupakan ambang batas yang telah ditentukan sebelumnya, umumnya sebesar 0.05 **(Sedgwick, 2014)**. Sedangkan *p-value* dapat dihitung menggunakan *z-score* terlebih dahulu menggunakan Persamaan (2), kemudian hasil dari *z-score* tersebut dibandingkan ke dalam tabel z-score untuk melihat *p-value* nya.

$$Z = \frac{x-\mu}{\sigma} \quad (2)$$

***Inter Quartile Range* (IQR)**

Data pencilan adalah pengamatan dalam kumpulan data yang memiliki pola atau nilai yang berbeda dari pengamatan lain dalam kumpulan data tersebut **(Cousineau, 2010)**. Titik ekstrim dalam pengamatan adalah nilai yang jauh atau sama sekali berbeda dengan kebanyakan nilai lain dalam kelompok tersebut, misalnya nilai terlalu kecil atau terlalu besar. Untuk mengetahui data pencilan dapat menggunakan teknik *boxplot*, *scatter plot* atau *Inter Quartile Range* (IQR). Teknik yang digunakan pada penelitian ini untuk menemukan nilai pencilan yaitu IQR dengan cara mencari selisih antara kuartil ketiga dengan kuartil pertama (IQR = Q3 -Q1).

Undersampling

Ketidakseimbangan data adalah kondisi dimana ada satu atau lebih kelas yang mendominasi keseluruhan kelas lainnya. Situasi ini akan membuat kinerja pembelajaran mesin menjadi lebih buruk karena akan cenderung memberi label dengan label mayoritas sehingga hasil evaluasinya akan terlihat bagus. Untuk mengatasi kondisi tersebut ada beberapa cara yang dapat dilakukan yaitu *undersampling* dan *oversampling* **(Kotsiantis, 2011)**. Teknik yang digunakan untuk menangani ketidakseimbangan data pada penelitian ini adalah *undersampling*.

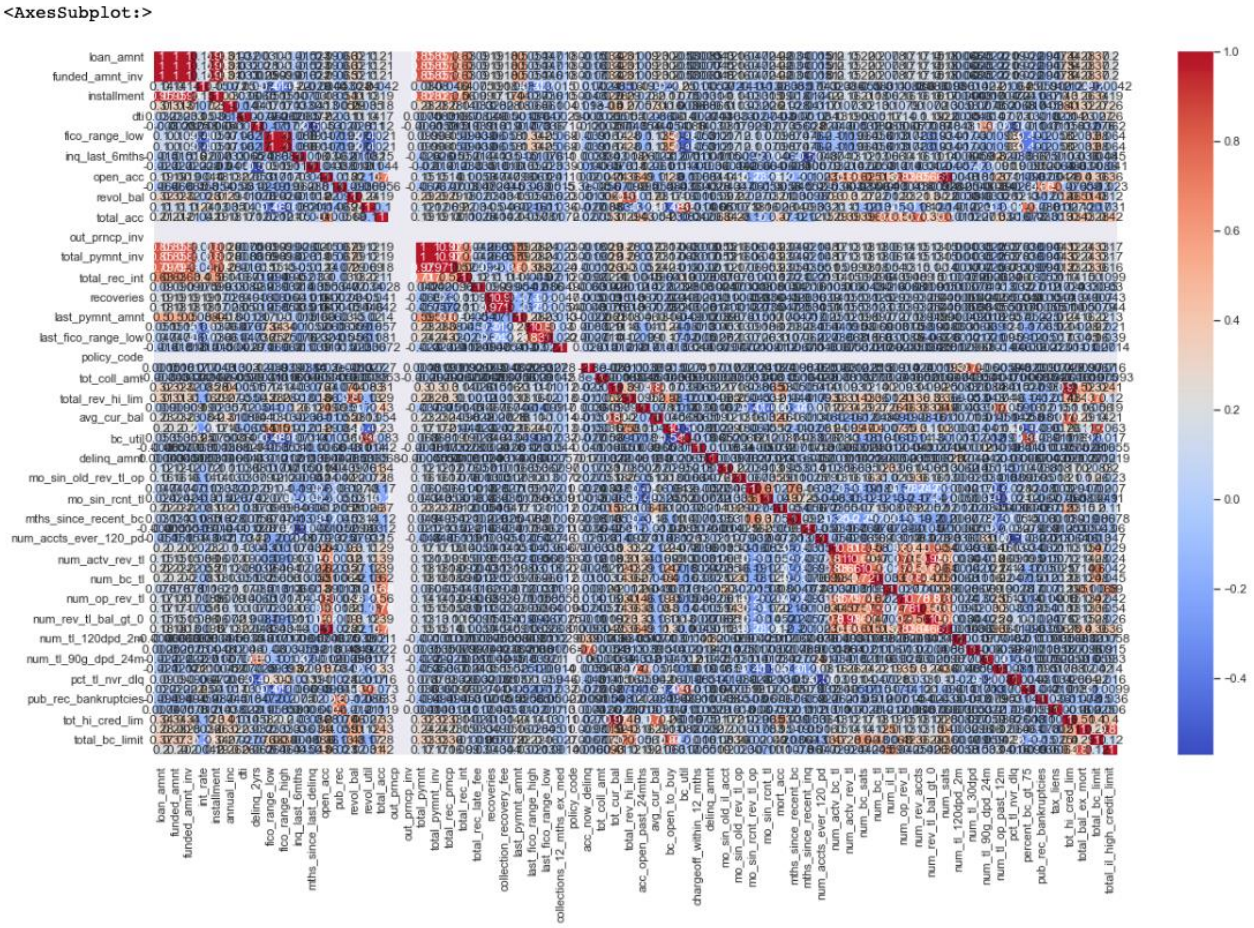
3. HASIL DAN PEMBAHASAN

3.1 Penanganan Data Hilang

Pada tahap ini data yang hilang ditemukan terlalu besar sehingga membuat penanganannya sulit. Oleh karena itu diperlukan suatu cara untuk mencari data yang hilang tersebut menggunakan menggunakan library pandas dengan fungsi df.isna(). Dari fungsi tersebut didapatkan ada 57 fitur mengalami data yang hilang lebih besar dari 50% dari total keseluruhan 151 fitur. Berdasarkan studi literatur yang dilakukan bahwa teknik *Listwise Deletion* dapat digunakan untuk mengeliminasi fitur dengan data yang hilang di atas 50%, yang mana dari 151 fitur dikurangi 57 fitur sehingga tersisa menjadi 94 fitur.

3.2 Seleksi Fitur

Pada tahap ini dilakukan pemetaan korelasi antar fitur dataset. Fitur yang memiliki korelasi tinggi satu sama lain akan mempengaruhi kinerja model. Fitur dengan korelasi tinggi lebih bergantung secara linier dan memiliki efek yang hampir sama pada variabel dependen. Sehingga ketika dua fitur memiliki korelasi yang tinggi, maka dapat menghilangkan salah satu dari dua fitur tersebut **(Doshi, 2014)**. Teknik yang digunakan untuk mencari korelasi fitur pada dataset tersebut adalah *Pearson Correlation*.

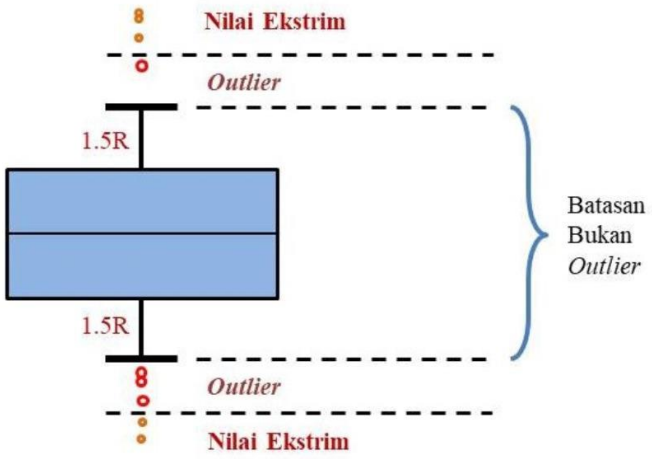


Gambar 3. *Pearson Correlation*

Fitur dengan korelasi mendekati nilai 1 berwarna merah memiliki arti berkorelasi tinggi sedangkan fitur korelasi mendekati nilai -1 berwarna biru memiliki arti berkorelasi rendah. Fitur dengan nilai koreasi di atas 0,7 akan dihapus karena memiliki nilai data yang sama atau mewakili hal yang sama sehingga membuat data menjadi redundan. Setelah menggunakan teknik *Pearson Correlation* dari 94 fitur dikurangi 27 fitur yang dihilangkan tersisa 67 fitur.

3.3 Penanganan Data Pencilan

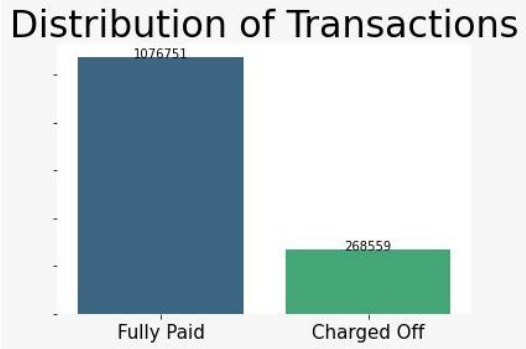
Penanganan data pencilan dilakukan untuk membuat kumpulan data yang ideal sehingga model yang dibuat menjadi lebih baik. Adanya data pencilan membuat interpretasi statistik model menjadi tepat. Data outlier dapat dideteksi menggunakan teknik *Inter Quartile Range* (IQR). Data terpencil dapat diganti dengan nilai representatif lainnya seperti rata-rata atau median dari setiap *cluster* atau dapat dihapus dengan menghapus seluruh *record*. Konsep Interquartile Range (IQR) digunakan untuk membuat grafik boxplot.



Gambar 4. Grafik Boxplot

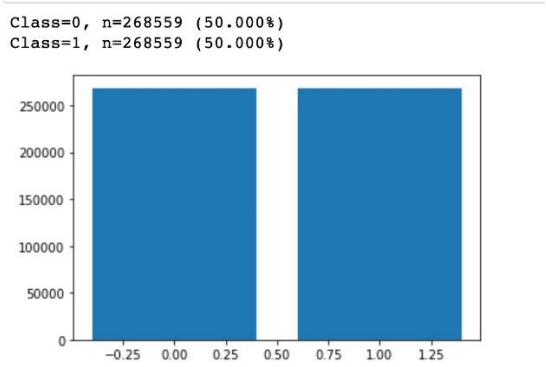
3.4 Penanganan Ketidakseimbangan Data

Data yang labelnya tidak seimbang akan sangat mempengaruhi kinerja algoritma, karena hasil prediksi akan lebih condong pada label yang dominan nilainya. Untuk itu, penanganannya akan menggunakan teknik *undersampling*. Gambar 5 menunjukkan ketidakseimbangan distribusi data sebelum ditangani menggunakan teknik *undersampling*.



Gambar 5. Ketidakseimbangan Data Sebelum Teknik *Undersampling*

Setelah menggunakan teknik *undersampling* yaitu dengan mereduksi data pada kelas dengan label mayoritas (*Fully Paid*) sehingga datanya lebih seimbang setara dengan label minoritas (*Charged Off*) seperti pada Gambar 6.



Gambar 6. Keseimbangan Data Setelah Teknik *Undersampling*

3.5 Pengujian

K-Fold Cross Validation

Teknik yang digunakan untuk melakukan evaluasi terhadap model adalah *K-Fold Cross Validation*. Pengujian dilakukan dengan nilai k=5 yaitu 5 kali iterasi dengan perbandingan 90% data latih dan 10% data uji, 80% data latih dan 20% data uji. Pada Tabel 1 dipaparkan hasil pengujian *k-fold cross validation* dengan 5 kali iterasi untuk perbandingan 90% data latih dan 10% data uji serta 80% data latih dan 20% data uji.

Tabel 1. Pengujian Menggunakan Teknik *K-Fold Cross Validation*

Iteration	***Train (90%)***	***Test (10%)***	***Accuracy***	***Train (80%)***	***Test (20%)***	***Accuracy***
1	483406	53712	97,19	429694	107424	97,13
2	483406	53712	97,15	429694	107424	97,28
3	483406	53712	97,17	429694	107424	97,08
4	483407	53711	97,14	429695	107423	97,71
5	483407	53711	97,18	429695	107423	97,16

Confusion Matrix

Tabel 2 menunjukkan hasil dari *confusion matrix*, dimana ada fitur target yang digunakan yaitu *fully paid* (dibayar penuh) dan *charged*-off (dibebankan). Dapat dilihat bahwa 53059 data *fully paid* dan diklarifikasi oleh model juga *fully paid*. 449 data *charge off* dan diklarifikasi *fully paid* oleh model. 399 data *fully paid* diklarifikasi *charge off* model dan 53067 data *charge off* diklarifikasi *fully paid* oleh model.

Tabel 2. Hasil Confusion Matrix

		Prediction	
		Fully Paid (0)	***Charged-off (1)***
Actual	***Fully Paid (0)***	53509 (TP)	399 (FP)
	Charged-off (1)	449 (FN)	53067 (TN)

Kinerja Algoritma XGBoost

Tabel 3 menunjukkan hasil kinerja algoritma XGBoost dengan akurasi sebesar 99,17% menyatakan akurasi model dalam memprediksi data dengan perbandingan data sebenarnya, presisi sebesar 99,28% menyatakan model memprediksi 99% kelas diklasifikasikan *Fully Paid*, *recall* sebesar 99,05% menyatakan model memprediksi 99% kelas diklasifikasikan *Charged Off*, spesifisitas sebesar 99,29% menyatakan model memprediksi 99% kelas *Fully Paid*, ROC/AUC sebesar 99,94% menyatakan grafik ROC data sampel dengan model XGBoost maka model yang dievaluasi memiliki tingkat akurasi yang sangat tinggi dan menunjukkan bahwa model ini baik untuk digunakan dan nilai F1 sebesar 99,17% menunjukkan bahwa model klasifikasi memiliki presisi dan *recall* yang baik.

Tabel 3. Kinerja Algoritma XGBoost

Accuracy	0.9917	***Specitifity***	0.9929
Precision	0.9928	***ROC/AUC***	0.9994
Recall	0.9905	***F1 Score***	0.9917
Training Time	350.126 s	***Prediction Time***	1.07465 s

4. KESIMPULAN

RESULTANT dengan menggabungkan beberapa teknik preparasi data seperti pemilihan fitur, penanganan *missing value*, penanganan data *outlier*, penanganan data ketidakseimbangan dan EDA telah berhasil meningkatkan kinerja dari algoritma XGBoost dengan akruasi sebesar 99,17%, presisi sebesar 99,28%, *recall* sebesar 99,05%, spesifisitas sebesar 99,29%, ROC/AUC sebesar 99,94%, dan skor f1 sebesar 99,17%.

Analisis data kredit dengan teknik preparasi data berhasil menghasilkan model algoritma XGBoost dengan skor tinggi pada dataset Lending Club. Algoritma ini unggul dalam akurasi, presisi, dan sensitivitas sehingga sangat memungkinkan untuk digunakan dalam kasus nyata.

UCAPAN TERIMA KASIH

Penelitian ini berjalan dengan baik dan lancar berkat kerja sama dan dukungan dari banyak pihak. Kami mengucapkan terima kasih kepada Itenas, khususnya FTI yang telah memberikan dukungan pendanaan registrasi pada konferensi ICGTD 2022.

DAFTAR RUJUKAN

Abdykalykova. (2020). Credit Scoring Using Machine Learning, *Information Technologies and Management*, 45 - 46.

Kartika, Risna. (2020). Analisis Peer To Peer Lending Di Indonesia. *Ilmu-Ilmu Ekon*, vol. 12, no. 2, pp. 75–86, 2020, doi: 10.35457/akuntabilitas.v12i2.902.

A. Kadav, J. Kawale., & Mitra P. (2013). Data Mining Standards. Online: http://www.idmarch.org/document_cache/0e9fd64335b3c9f01f6b39320f99c190.pdf.

Wirth, R., & Hipp, J. (2000). CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39. *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, 29–39. Online: https://www.researchgate.net/publication/239585378_CRISPDM_Towards_a_standard_process_model_for_data_mining.

Schröer, C., Kruse, F., & Gómez, J. M. (2019). A systematic literature review on applying CRISP-DM process model. *Procedia Comput. Sci.*, vol. 181, 526 - 534. doi:

10.1016/j.procs.2021.01.199.

Abdallah, Z. S., and Webb, G. (2017). Encyclopedia of Machine Learning and Data Mining. *Encycl. Mach. Learn. Data Min.*, no. September 2018. doi: 10.1007/978-1-4899-7687-1.

Hartini, E. (2016). Efficiency Comparison of Method of Handling Missing Value in Data Evaluation System or Component. *Pros. Semin. Nas. Teknol. Energi Nukl*, 4–5.

Samuels, P. (2014). Pearson Corelation," no. April 2014, 1 - 5. [Online]. Available: https://www.researchgate.net/publication/274635640.

Sedgwick, P. (2014). Understanding P values. *BMJ*, vol. 349, no. July 2014, pp. 10–12, 2014, doi: 10.1136/bmj.g4550.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *Int. J. Psychol. Res.*, vol. 3, no. 1, 58–67. doi: 10.21500/20112084.844.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2011). Handling imbalanced datasets : A review. *Science (80-.).*, vol. 30, no. 1, 25 - 36, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.9248&rep=rep1&type=pdf.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13 - 17-Augu, 785–794. doi: 10.1145/2939672.2939785.

Doshi, M., & Chaturvedi, S. K. (2014). Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance. *Int. J. Comput. Networks Commun.*, vol. 6, no. 3, pp. 197–206, 2014, doi: 10.5121/ijcnc.2014.6315.

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika J. Sci. Technol.*, vol. 26, no. 1, 329 - 340.

Jassim, A. M., Abdul Wahid, S. N. (2020). Data Mining Preparation: Process, Techniques, and Major Issues in Data Analysis. *ICEST*, 1090 (2021) 012053.