P-ISSN : 2622-1594
E-ISSN : 2685-449X

Credit Risk Analysis dengan Algoritma Extreme Gradient Boosting dan Adaptive Boosting

Rosa Delima Mendrofa[a], Maria Hosianna Siallagan[b], Diana Pebrianty Pakpahan[c], Junita Amalia[d]
[a,b,c,d] *Study Program of Information System, Institut Teknologi Del*,
E-mail: junita.amalia@del.ac.id

***Abstrak*—Credit Risk Analysis digunakan untuk mengenali resiko terhadap pinjaman untuk mencegah penunggakan pembayaran utang. Pemberian uji kelayakan pinjaman dapat di analisis menggunakan model klasifikasi. Untuk menghasilkan model credit risk analysis yang sesuai, penulis mengajukan Algoritma Extreme Gradient Boosting (XGBoost) dan Adaptive Boosting (AdaBoost). Data yang digunakan dalam penelitian ini adalah data pinjaman *platform* Peer to Peer (P2P) Lending. Penelitian ini menerapkan data preprocessing yang bertujuan untuk menghasilkan data yang lebih baik dan melakukan analisis terhadap data. Analisis dilakukan berdasarkan fitur yang dimiliki oleh peminjam menggunakan algoritma klasifikasi berdasarkan historical data pinjaman peminjam. Fitur yang digunakan seperti jumlah pinjaman yang diajukan, total pinjaman yang ditawarkan, jumlah pembayaran pinjaman, jangka waktu pembayaran, suku bungan pinjaman, jumlah angsuran dan lain lain. Jumlah fitur sebelum dilakukan data reduksi 136 dan setelah direduksi 34 fitur. Fitur tersebut digunakan pada penerapan algoritma XGBoost dan AdaBoost untuk menghasilkan klasifikasi *good borrower* dan *bad borrower*. Penulis menggunakan metode evaluasi kurva ROC dan nilai AUC untuk menilai performa dari kedua algoritma. Pada kurva ROC, nilai AUC dari algoritma XGBoost 0,92 dan nilai AUC dari algrithma AdaBoost adalah 0,89. Berdasarkan perbandingan nilai AUC tersebut dapat disimpulkan algoritma XGBoost menghasilkan klasifikasi yang lebih baik untuk model klasifikasi pemberian pinjaman.**

Kata Kunci— Data mining, XGBoost, AdaBoost, Peer to Peer (P2P) Lending

I. PENDAHULUAN

P2P *Lending* merupakan *platform* yang menyediakan layanan pinjaman uang dengan adanya perjanjian yang dilakukan secara online. P2P *Lending* menyediakan saluran pembiayaan baru untuk usaha kecil, usaha mikro dan peminjam individu [1]. Dalam mekanisme P2P *Lending*, pemberi dana (*investor*) tidak mengetahui stabilitas kondisi finansial peminjam. Kondisi finansial peminjam perlu diketahui oleh investor untuk analisis *credit assesment* yang bertujuan untuk menilai risiko gagal bayar. Oleh karena itu, *fintech* perlu meningkatkan kualitas aplikasi P2P *Lending* menggunakan *credit risk analysis* dengan penerapan ilmu statistika, *machine learning*, dan *data mining*.

Naskah Masuk : 05 Juli 2022
Naskah Direvisi : 06 Desember 2022
Naskah Diterima : 16 Januari 2023

Credit risk analysis akan menilai kemungkinan risiko gagal bayar yang ditimbulkan oleh peminjam kepada investor. Aliran uang yang dipinjamkan akan berpengaruh ketika bunga muncul dan jumlah pokok pinjaman tidak dibayarkan. Hal ini mengakibatkan biaya penagihan akan meningkat. Peminjam dengan karakteristik yang gagal bayar tersebut sulit untuk diidentifikasi secara manual berdasarkan pemeriksaan aplikasi pengajuan pinjaman. *Credit risk analysis* membantu fintech dalam mem-percepat proses persetujuan pinjaman karena dapat meng-identifkasi peminjam yang kemungkinan mengalami risiko kredit macet.

Penelitian mengenai *credit risk analysis* telah dilakukan sebelumnya, menggunakan data historis transaksi pinjaman untuk membangun model *credit risk* yang mampu menentukan kelayakan pemberian kredit atau tidak [2]. Penelitan tersebut melakukan perbandingan kinerja model yang dihasilkan dengan beberapa model klasifikasi yaitu *Logistic Regression*, XGBoost, *Linear Discriminant Analysis* (LDA) dan MXNET untuk memperoleh model dengan performansi terbaik berdasarkan metrik evaluasi yang digunakan. Dari penelitian yang dilakukan, diperoleh algoritma XGBOOST dan MXNET menghasilkan nilai akurasi tertinggi dibandingkan dengan *Logistic Regression* dan LDA. *Dataset* yang digunakan terdiri dari 354 variabel independen sehingga menerapkan *feature selection* berdasarkan plot hasil *feature importance* yang digunakan dalam *train* model. Namun, pemilihan fitur yang dilakukan tidak mempertimbangkan arti dari setiap fitur yang dihapus dalam train model [2]. Berdasarkan penelitian yang dilakukan, diperoleh oleh Morten Hansen Flood bahwa performansi model AdaBoost mampu mengidentifikasi pelanggan yang memiliki resiko penipuan terhadap kartu kredit [3].

Berdasarkan pemaparan diatas, mendorong peneliti untuk membandingkan performansi algoritma XGBoost dan AdaBoost menggunakan *dataset historical Lending Club* pada periode 2007 – 2017 di Amerika Serikat. *Lending club* merupakan perusahaan pemberi pinjaman *peer-to-peer* yang mencocokkan calon peminjam dengan investor atau pemberi pinjamannya melalui *platform online*. Kedua metode ini mampu menangani jumlah fitur yang cukup besar dalam mendapatkan kombinasi fitur terbaik dalam proses pemodelan. Penggunaan metode ini juga didasarkan pada kemampuan metode Boosting dalam mengurangi varians dan bias yang terdapat dalam dataset penelitian ini [3].

DOI : 10.37823/insight.v5i1.233

II. TINJAUAN PUSTAKA

A. Credit Risk Analysis

Credit Risk Analysis merupakan usaha untuk mengenali risiko terhadap kredit yang diberikan oleh lembaga keuangan yang bertujuan untuk mencegah kerugian. Risiko kredit timbul karena peminjam mengingkari pembayaran hutang yang dipengaruhi oleh kondisi stabilitas ekonomi peminjam. Hal tersebut berpengaruh pada penunggakan pembayaran hutang. Kondisi stabilitas tersebut berupa jumlah gaji maupun status pekerjaan dari seorang peminjam sehingga akan mempengaruhi kemampuan pembayaran hutang. Analisis risiko merupakan bagian penting dari keputusan kredit dan ketepatannya memiliki konsekuensi yang signifikan pada manajemen kredit. Ketidakmampuan mengidentifikasi risiko dengan benar dapat mempengaruhi keputusan kredit, yang dapat menyebabkan kegagalan aset usaha [4] . Salah satu tujuan dari *credit risk analysis* adalah untuk menemukan jawaban atas pertanyaan apakah kredit baru akan diperpanjang, atau apakah serangkaian fasilitas kredit saat ini yang diberikan kepada peminjam harus ditarik [4].

B. Peer to Peer (P2P) Lending

Peer to Peer Lending merupakan *platform* yang merujuk pada layanan pinjaman oleh pemberi dan penerima pinjaman yang dilakukan melalui *platform online*. Dalam *Peer to Peer (P2P) lending*, investor ingin mendapatkan informasi yang valid tentang peminjam, sementara peminjam cenderung menyembunyikan beberapa karakteristiknya untuk mendapatkan bunga yang rendah. Untuk memungkinkan pemberi pinjaman membuat keputusan berdasarkan informasi yang valid, *platform P2P lending* memaksa peminjam untuk memberikan informasi keuangan yang telah divalidasi oleh lembaga eskternal. Selain itu *P2P lending* menyarankan peminjam memberikan informasi demografis, seperti jenis kelamin, ras atau usia. Peminjam juga sering diberi kesempatan untuk memberikan informasi sosial, yang tidak dapat divalidasi, seperti hobi, latar belakang keluarga atau foto [5]. Karakteristik ini disebut sebagai penentu *P2P lending*, karena memiliki pengaruh besar pada keberhasilan pendanaan daftar pinjaman dan tingkat bunga yang diminta.

C. Knowledge Discovery in Database (KDD)

Penambangan data adalah proses menyelesaikan masalah dengan menganalisis data yang sudah ada dalam *database* dengan menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan terkait dari berbagai *database* besar [6]. Proses *Knowledge Discovery in Database* (KDD) adalah urutan yang berulang dari beberapa tahapan yaitu *data cleaning, data integration, data reduction, data transformation, data mining, pattern evaluation* dan *knowledge presentation* [6].

Data cleaning merupakan tahap menghilangkan *noisy* data, melengkapi data dan mengatasi data yang tidak relevan dan tidak konsisten. Pendekatan untuk melakukan pembersihan data dengan menggunakan *Use a measure of central tendency for the fitures*. Metode ini dilakukan dengan mengisi data yang hilang menggunakan *mean* atau median dari data. Sementara untuk data dengan distribusi data yang tidak normal atau tidak merata dapat menggunakan median dalam melakukan pengisian data yang hilang [6]. Data *reduction* adalah proses dalam menyeleksi fitur untuk tahap pemodelan dengan mengurangi jumlah dimensi [7]. Metode dalam melakukan data *reduction*, yaitu *dimensionality reduction atau* seleksi subset fitur. *Dimensionality reduction* adalah proses mengu-rangi jumlah variabel acak atau fitur yang dipertimbang-kan. Seleksi subset fitur adalah metode pengurangan dimensi di mana fitur atau dimensi yang tidak relevan, atau redundan dideteksi dan dihapus.

Data *transformatio*n adalah proses untuk mengubah data atau menggabungkan data ke dalam format yang sesuai untuk diproses dalam data *mining*. Beberapa metode data *mining* membutuhkan format data yang khusus sebelum diaplikasikan [6]. Beberapa cara yang dilakukan untuk melakukan data *transformation*, yaitu:

1) *Feature construction*
   Tahap ini digunakan untuk menambahkan fitur baru pada dataset untuk membantu proses data *mining*.
2) *Normalization*
   Tahap ini digunakan untuk melakukan *normalization* data pada dataset. Salah satu metode yang dilakukan dalam *normalization* adalah *min-max normalization* [3].
3) *Discretization*
   Tahap ini digunakan untuk memastikan bahwa algoritma pembelajaran mesin menginterpretasikan fitur ordinal dengan benar, maka fitur dengan tipe data string atau kategorikal diubah menjadi integer dengan menggunakan teknik *discretization* [7].

D. Adaptive Boosting (AdaBoost)

Algoritma AdaBoost adalah salah satu algoritma *boosting* yang meningkatkan model klasifikasi secara iteratif menggunakan *base learner* [8]. Algoritma AdaBoost dapat dilihat pada Pseudocode 1.

*Pseudocode.1. Algoritma AdaBoost (*Sumber: Boosting Foundations and Algorithms, 2012)

Given : $(x_1, y_1), \dots, (x_m, y_m)$ where $\mathrm{x_i} \in \mathcal{X}, \mathrm{y_i} \in \{-1, +1\}$

Initialize : $D_1(i) = \frac{1}{m}$ for $\mathrm{i} = 1, \dots, \mathrm{m}$

For $\mathrm{t} = 1, \dots, \mathrm{T}$:

- Train weak learner using distribution D_t
- Get weak hypothesis $h_t: \mathcal{X} \to \{-1, +1\}$
- Aim : Select h_t to minimalize the weighted error:
$$\epsilon_t = \boldsymbol{Pr}_{\mathrm{i}\sim \mathrm{D}_t}[h_t(x_i) \neq y_i]$$
- Choose:
$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$
- Update, for $1 = 1, \dots, \mathrm{m}$:
$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } \mathrm{h_t(x_i} = \mathrm{y_i)} \\ e^{\alpha_t} & \text{if } \mathrm{h_t(x_i} \neq \mathrm{y_i)} \end{cases} = \frac{D_t(i)\exp(-\alpha_t\, y_i h_t(x_i))}{Z_t}$$

Where Z_t is a normalization factor (choosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t\, h_t\,(x)\right)$$

 DOI : 10.37823/insight.v5i1.233

E. Data

Data merupakan sekumpulan fakta mentah yang belum memiliki nilai fungsionalitas atau arti [6]. Dalam melakukan penambangan data, pertama sekali yang harus dilakukan adalah mempersiapkan pengolahan data. Data yang diolah oleh penulis dalam penelitan ini adalah data dari *platform* Kaggle.com yang merupakan data *Lending Club Peer to Peer* AS dari periode 2007 – 2017.

F. Extreme Gradient Boosting (XGBoost)

XGBoost merupakan implementasi dari *gradient boosting* yang menggunakan formalisasi model yang lebih teratur untuk mengontrol *overfitting*. XGBoost dikembangkan oleh T. Chen dan C. Guestrin pada tahun 2016 yang merupakan *library gradient boosting* terdistribusi yang telah dioptimasi untuk desain yang lebih efisien, fleksibel dan *partable*. XGBoost bekerja untuk melakukan klasifikasi sehingga memperoleh akurasi yang tinggi dengan kinerja model yang lebih baik. XGBoost digunakan untuk masalah *supervised learning* dengan menggunakan beberapa fitur dari data *train* (x_i) untuk memprediksi variabel target (y_i). Diberikan sekumpulan dataset dengan n sampel dan m fitur $= \{(x_i, y_i)\}(|\mathcal{D}| = n, x_i \in \mathbb{R})$, sebuah *tree ensemble* model menggunakan fungsi penjumlahan K untuk memprediksi *output*. Untuk mencari nilai prediksi $\hat{y}_i$ pada setiap iterasi maka digunakan rumus berikut

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k\,(x_i), f_k \in \mathcal{F} \quad (1)$$

Dimana

$$\mathcal{F}: \{f(x) = wq_{(x)}\}(q: \mathbb{R}^m \longrightarrow T, \in\ \mathbb{R}^T) \quad (2)$$

adalah space dari *regression tree* (CART). Disini q merepresentasikan struktur dari setiap *tree* yang memetakan sebuah sampel indeks *leaf* yang sesuai. T merupakan jumlah daun dalam *tree*. Setiap f_k sesuai dengan struktur *tree independent* q dan bobot *leaf* w_i yang merepresentasikan skor pada *leaf* ke-i.

Model XGBoost mengoptimalkan kumpulan model *ensemble* prediksi yang lemah untuk membangun prediktor yang akurat, sehingga memungkinkan output selanjutnya dari model yang berbeda untuk ditambahkan. Melalui pendekatan ini memungkinkan peningkatan residu dalam prediksi yang mengarah ke prediksi yang lebih tepat.

Secara teoritis berikut perbandingan Algoritma XGBoost dan AdaBoost.

TABEL I
PERBANDINGAN ALGORITMA XGBOOST DAN ADABOOST

Parameter	XGboost	AdaBoost
Dapat menghasilkan *decision rules* untuk analisis atribut	Ya	Ya
Dapat menangani kasus *imbalanced data*	Ya	Ya
Dapat menangani jumlah data yang besar	Ya	Ya
Penanganan tipe data yang berbeda	Ya	Ya
Penanganan *missing value*	Ya	Ya
Kemampuan untuk menangani *outliers*	Ya	Ya
Sensitif terhadap transformasi input yang monoton	Ya	Ya
Kemampuan untuk menangani input yang tidak relevan	Tidak	Tidak
Kemampuan mengekstrak kombinasi fitur yang linear	Tidak	Ya

G. Metrik Evaluasi Klasifikasi

Metrik evaluasi yang dapat digunakan dalam klasifikasi data mining antara lain:

1) *Hold-out Method*

Hold-out adalah metode evaluasi klasifikasi dengan membagi dataset secara acak menjadi *data train* dan *data test* [9]. *Data train* adalah data yang digunakan untuk model yang dilatih, dan *data test* untuk mengukur seberapa baik kinerja model pada data yang baru. Pembagian data (*splitting data*) pada penelitian ini sebesar 70 persen data untuk *data train* dan 30 persen sebagai *data test* untuk pengujian model.

2) *Confusion Matrix*

Confusion matrix adalah suatu metode *supervised learning* yang digunakan untuk melakukan perhitungan kinerja dalam proses klasifikasi *accuracy* dalam konsep data mining. Evaluasi dengan *confusion matrix* akan menghasilkan nilai *accuray, precision, recall* dan *f-measure*. *Confusion matrix* berbentuk tabel matriks yang menggambar-kan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Tabel *confusion matrix* dapat dilihat pada Tabel I.

TABEL II
CONFUSION MATRIX

		Predicted	
		1	**0**
Actual	**1**	True Positif	False Negarif
	0	False Positif	True Negarif

3) *Receiver Operating Characteristic* (ROC) dan *Area Under Curve* (AUC)

ROC adalah kurva probabilitas yang merupakan plot dari nilai *true positive rate* (TPR) dan *false positive rate* (FPR) berdasarkan nilai elemen-elemen *confusion matrix* yang diperoleh. Kurva ROC menunjukkan *trade-off* antara TPR dan FPR. Sumbu Y merepresentasikan TFR (*sensitivity*), sumbu X merepresentasikan FPR (1-*specifity*). Pada proses pengujian model, TPR merupakan proporsi *row positif* dengan label benar yang diklasifikasikan oleh model. Sedangkan FPR adalah proporsi dari *row negatif* dengan label salah sebagai positif [10].

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

$$Sensitivity\ (TPR) = \frac{TP}{TP+FN} \quad (4)$$

AUC merupakan luas area di bawah kurva ROC atau integral dari fungsi ROC yang digunakan untuk membandingkan kinerja antara dua algoritma [11]. AUC digunakan sebagai ukuran performansi model dalam memisahkan kelas. Berikut kategori performansi model berdasarkan rentang nilai AUC.

TABEL III
KATEGORI MODEL BERDASARKAN RENTANG NILAI AUC

No.	Rentang Nilai AUC	Kategori
1	0,9 – 1,0	Very Good
2	0,8 – 0,9	Good
3	0,7 – 0,8	Fair
4	0,6 – 0,7	Poor
5	0,5 – 0,6	Fail

 DOI : 10.37823/insight.v5i1.233

Secara statistik, interpretasi nilai AUC memiliki beberapa aturan dalam menentukan kinerja algoritma yang dapat dilihat pada Tabel II. Semakin tinggi nilai AUC maka performansi model semakin bagus dalam memisahkan kelas hasil klasifikasi [11].

III. METODE PENELITIAN

Penelitian ini menggunakan dataset yang diperoleh dari *website Kaggle* yang merupakan data *lending club* pada periode 2007 – 2017 di Amerika Serikat. *Lending club* merupakan perusahaan pemberi pinjaman *peer to peer* yang mencocokan calon peminjam dengan investor atau pemberi pinjamannya melalui *platform online*. Dataset ini merupakan data pinjaman yang terdiri dari 3 file yaitu *lc_2016_2017, Accepted_2007_to_2016* dan *us_state_code* dengan jumlah baris 1.581.185 dan 112 fitur.

Tahapan penelitian dilakukan berdasarkan proses *Knowledge Discovery and Data Mining*. Berdasarkan data *lending club* yang merupakan *input* pada proses pemodelan uji kelayakan pemberian kepada peminjam. Data *preprocessing* terdiri dari *data reduction, data cleaning* dan *data transformation.*

A. Data Reduction

Penelitian ini perlu melakukan proses data reduction karena dataset dalam penelitian ini merupakan fitur yang dimiliki peminjam dan pemberi pinjaman, sehingga perlu dilakukan data reduction dengan menghapus fitur yang tidak berhubungan dengan peminjam atau pihak yang akan mengajukan pinjaman. Metode *data reduction* yang digunakan adalah *dimensionality reduction* dengan mengidentifikasi nilai korelasi antar fitur dengan target variabel, yaitu *loan_status*. fitur yang tidak memiliki korelasi dengan target variabel akan dihapus dari dataset dan fitur yang berbeda tetapi memiliki nilai korelasi yang sama dengan target variabel akan dihapus dari dataset.

B. Data Cleaning

Penelitian ini perlu dilakukan proses *data cleaning* karena dataset penelitian ini memiliki *missing value* dan *outlier*. Teknik yang digunakan untuk mengatasi *missing value* adalah *use a measure of central tendency for the fitures*. Metode ini dilakukan dengan mengisi data yang hilang menggunakan *mean* atau *mode* dari data. Nilai *mean* digunakan untuk mengisi *missing value* yang terdapat pada fitur numerik dan nilai *mode* digunakan untuk mengisi *missing value* yang terdapat pada fitur kategorikal. Tupel yang memiliki nilai *outlier* pada atrbiut numerik akan dihapus dari dataset penelitian ini.

C. Data Transformation

Penelitian perlu melakukan *data transformation* untuk mentransformasikan data dalam bentuk yang sesuai sehingga dapat digunakan dalam pemodelan. Teknik *data transforma-tion* dilakukan dalam penelitian ini adalah *feature construction, normalization* dan *discreatization*. *Feature construction* dilakukan untuk menghasilkan fitur baru berdasarkan fitur yang sudah ada sebelumnya. *Normalization* yang digunakan adalah *min-max normaliza-tion* untuk merepresentasikan nilai fitur yang berada pada skala sama. *Discreatization* dilakukan pada fitur kategorikal dalam dataset penelitian yang bertujuan untuk merepresentasikan nilainya kedalam bentuk *numerical*.

Kemudian dataset yang telah melalui tahapan *preprocessing* digunakan untuk membangun model klasifikasi pinjaman menggunakan kedua algoritma yaitu XGBoost dan AdaBoost. Selain itu experiment juga dilakukan pada data tanpa *data reduction.* Tujuannya peneliti ingin melihat performa model yang dibangun melalui eksperimen tanpa *data reduction* dan dengan *data reduction.* Setelah melakukan pembangunan model, maka dilakukan perbandingan kinerja model menggunakan metrik kurva ROC dan nilai AUC.

IV. HASIL DAN PEMBAHASAN

Pada bagian ini dijelaskan hasil dan pembahasan dari penelitian yang telah dilakukan.

A. Hasil

Berikut hasil yang diperoleh dalam penelitian.

1) Hasil Eksperimen *Train Model* Tanpa *Data Reduction*

Pembangunan model tanpa menerapkan *data reduction* menggunakan 136 fitur untuk *train model*. Model XGBoost yang dihasilkan menggunakan 100 *tree* dengan *max_depth* sebesar 3 berdasarkan parameter XGBoost. Berdasarkan model XGBoost yang dihasilkan, terdapat 10 fitur utama (*feature importance*) yang memiliki pengaruh dalam *train model*. Tabel IV menunjukkan fitur penting yang berpengaruh pada klasifikasi pinjaman melalui proses tanpa *data reduction*.

TABEL IV
IMPORTANT FEATURES XBOOST TANPA DATA REDUCTION

No.	Fitur
1.	recoveries
2.	funded_amnt
3.	id
4.	last_pymnt_d
5.	funded_amnt_inv
6.	sub_grade
7.	last_credit_pull_d
8.	total_rec_int
9.	total_rec_prncp
10.	out_prncp

Fitur yang terdapat pada Tabel IV merupakan 10 fitur utama yang relevan dalam pembangunan model XGBoost tanpa *data reduction.*

TABEL V
IMPORTANT FEATURES ADABOOST TANPA DATA REDUCTION

No.	Fitur
1.	total_rec_prncp
2.	installment
3.	last_pymnt_d
4.	member_id
5.	id
6.	out_prncp_inv
7.	sub_grade
8.	total_rec_int
9.	last_credit_pull_d
10.	int_rate

Model AdaBoost yang dihasilkan menggunakan 50 *decision stumps* dengan *max_depth* sebesar 1. *Decision stumps* yang digunakan model AdaBoost merupakan *weak learner* yang akan dilatih menjadi *strong classifer*.

 DOI : 10.37823/insight.v5i1.233

Berdasarkan model AdaBoost yang dihasilkan, terdapat 10 fitur utama (*feature importance*) yang memiliki pengaruh dalam *train model*. Tabel V menunjukkan fitur penting dalam proses tanpa data reduction.

Fitur yang terdapat pada Tabel V merupakan 10 fitur utama yang relevan dalam pembangunan model AdaBoost tanpa data reduction.

2) Hasil Eksperimen *Train Model* Melalui *Data Reduction*

Pembangunan model XGBoost dengan menerapkan *data reduction* menggunakan 34 fitur untuk train model. Model XGBoost yang dihasilkan menggunakan 100 tree dengan parameter max_depth sebesar 3. Berdasarkan model XGBoost yang dihasilkan, terdapat 8 fitur utama (*feature importance*) yang memiliki pengaruh dalam train model. Tabel VI menunjukkan fitur penting yang berpengaruh pada klasifikasi pinjaman tanpa data reduction.

TABEL VI
IMPORTANT FEATURES XSBOOST DENGAN DATA REDUCTION

No.	Fitur
1.	int_rate
2.	installment
3.	year_issue_d
4.	total_pymnt
5.	month_issue_d
6.	term_0
7.	loan_amnt
8.	pymnt_plan

Fitur yang terdapat pada Tabel VI merupakan fitur yang relevan dalam pembangunan model dengan *data reduction* sehingga kinerja model lebih baik. Terdapat 8 fitur yang berpengaruh dalam pembangunan model XGBoost dengan menerapkan data reduction.

Model AdaBoost yang dihasilkan menggunakan 50 *decision stumps* dengan parameter *max_depth 1*. *Decision stump* yang digunakan model AdaBoost merupakan *weak learner* yang akan dilatih menjadi *strong classifer*. Berdasarkan model Adaboost yang dihasilkan, terdapat 10 fitur utama (*feature importance*) yang memiliki pengaruh dalam *train model*. Tabel VII menunjukkan fitur penting yang berpengaruh pada proses klasifikasi pinjaman dengan menerapkan data reduction.

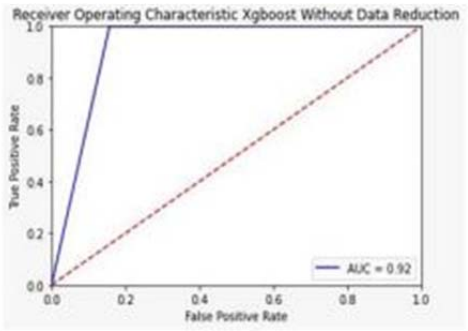
TABEL VII
IMPORTANT FEATURES ADABOOST DENGAN DATA REDUCTION

No.	Fitur
1.	total_pymnt
2.	year_issue_d
3.	installment
4.	int_rate
5.	month_issue_d
6.	loan_amnt
7.	pymnt_plan
8.	dti
9.	application_type_1
10.	home_ownership_1

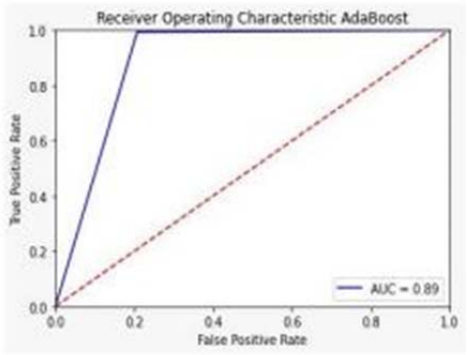
Fitur yang terdapat pada Tabel VII merupakan fitur yang relevan dalam pembangunan model dengan data reduction sehingga kinerja model lebih baik. Terdapat 10 fitur yang berpengaruh dalam pembangunan model AdaBoost dengan menerapkan data reduction.

3) Hasil Metode Evaluasi

Metode evaluasi yang digunakan untuk menguji kinerja model adalah metode *hold out* dengan membagi dataset menjadi *data train* dan *data test*. Metrik yang digunakan untuk menguji perfoma model menggunakan kurva ROC dan nilai AUC. Gambar 1 dan 2 merupakan perbandingan hasil evaluasi model Model XGBoost dan AdaBoost tanpa *data reduction.*



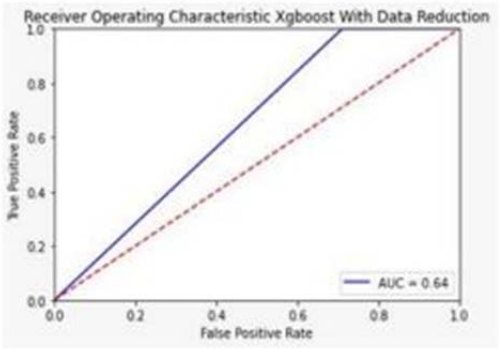
Gambar 1. Kurva ROC Model XGBoost Tanpa Data Reduction



Gambar 2. Kurva ROC Model AdaBoost Tanpa Data Reduction

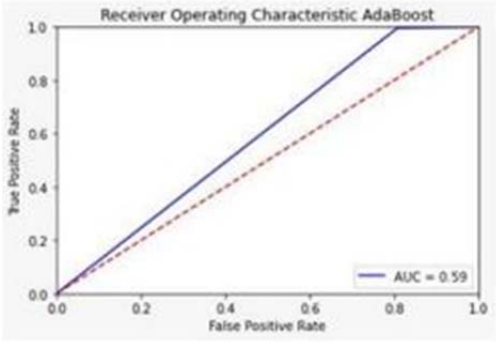
Hasil evaluasi model terbaik berdasarkan kurva ROC, diperoleh nilai AUC model XGBoost tanpa *data reduction* adalah 0,92. Model ini termasuk dalam kategori "*very good"* dalam memisahkan kedua kelas dengan tepat. Artinya model yang dihasilkan dapat digunakan untuk menguji kelayakan pemberian pinjaman. Dari kurva ROC, nilai AUC model AdaBoost tanpa data reduction adalah 0,89. Berdasarkan nilai AUC yang diperoleh model ini termasuk dalam kategori *"good"* dalam memisahkan kedua kelas dengan tepat. Artinya model ini layak digunakan untuk menguji kelayakan pemberian pinjaman.

Gambar 3 dan 4 merupakan perbandingan hasil evaluasi model Model XGBoost dan AdaBoost dengan *data reduction*.



Gambar 3. Kurva ROC Model XGBoost dengan Data Reduction

 DOI : 10.37823/insight.v5i1.233



Gambar 4. Kurva ROC Model AdaBoost dengan Data Reduction

Berdasarkan kurva ROC pada Gambar 3 dan 4, diperoleh nilai AUC model XGBoost dengan *data reduction* adalah 0,64. Model ini termasuk dalam kategori "*poor*" dalam memisahkan kedua kelas. Artinya model yang dihasilkan memiliki performansi yang buruk atau hampir tidak mampu memisahkan kedua kelas dengan tepat. Kurva ROC pada model AdaBoost dengan *data reduction* adalah 0,59. Model ini termasuk dalam kategori "fail" dalam memisahkan kedua kelas. Artinya model yang dihasilkan gagal dalam memisahkan kedua kelas atau tidak mampu memisahkan kedua kelas, sehingga model tidak dapat digunakan untuk menguji kelayakan pinjaman.

B. Pembahasan

Berikut dijelaskan pembahasan hasil yang diperoleh penulis dalam penelitian.

1) Pembangunan Model

Pembangunan model XGBoost dan AdaBoost dilakukan dengan *data reduction* dan tanpa *data reduction*. Dengan rincian data yang digunakan seperti berikut.

TABEL VIII
PERBANDINGAN DATA DENGAN REDUCTION DAN TANPA REDUCTION

	Data Reduction	**Tanpa *Data Reduction***
Jumlah tuple untuk membangun model	796.882 rows dengan 34 fitur	796.882 rows dengan 136 fitur
Jumlah tuple untuk menguji model (data test)	341.521 rows dengan 34 fitur	341.521 rows dengan 135 fitur

Adapun target variabel adalah *loan_status* yang telah direpresentasikan menjadi numerik. Nilai 1 mereperesentasikan kategori pinjaman *Fully Paid* dan 0 merepresentasikan kategori pinjaman *Charged Off.*

2) Pembahasan Hasil Eksperimen

Pada bagian ini dijelaskan pembahasan dari hasil eksperimen dengan menerapkan *data reduction* dan tanpa *data reduction* dalam pembangunan model. Berikut merupakan tabel *confusion matrix* pada model XGBoost dan AdaBoost melalui *data reduction* dan tanpa penerapan *data reduction.*

TABEL IX
CONFUSION MATRIX MODEL XGBOOST DENGAN DATA REDUCTION

		Predicted	
		1	**0**
Actual	**1**	301.147	62
	0	28.653	11.659

Dimana:
TP = 301.147 FN = 62
FP = 28.653 TN = 11.659

TABEL X
CONFUSION MATRIX MODEL ADABOOST DENGAN DATA REDUCTION

		Predicted	
		1	**0**
Actual	**1**	300.070	1.139
	0	32.584	7.728

Dimana:
TP = 300.070 FN = 1.139
FP = 32.584 TN = 7.728

TABEL XI
CONFUSION MATRIX MODEL XGBOOST TANPA DATA REDUCTION

		Predicted	
		1	**0**
Actual	**1**	300.888	321
	0	6.310	34.002

Dimana:
TP = 300.888 FN = 321
FP = 6.310 TN = 34.001

TABEL XII
CONFUSION MATRIX MODEL ADABOOST TANPA DATA REDUCTION

		Predicted	
		1	**0**
Actual	**1**	299.331	1.878
	0	8.362	31.950

Dimana:
TP = 299.331 FN = 1.878
FP = 8.362 TN = 31.950

Berdasarkan nilai *confusion matrix* pada Tabel IX sampai Tabel XII ,maka diperoleh kalkulasi nilai *sensitivity* (TPR) dan *specificity*. Nilai tersebut dapat dilihat pada Tabel XIII.

TABEL XIII
PERFORMANSI MODEL XGBOOST DAN ADABOOST

Algoritma	**Dengan *Data Reduction***		**Tanpa *Data Reduction***	
	Sensitivity	***Specificity***	***Sensitivity***	***Specificity***
XGBoost	0,99	0,29	0,99	0,84
AdaBoost	0,99	0,19	0,99	0,79

Berdasarkan Tabel XIII, nilai *sensitivity* model XGBoost tanpa *data reduction* dan dengan data *reduction sama* yaitu sebesar 0,99. Hal ini mengindikasikan bahwa rasio kebenaran model dalam memprediksi pinjaman kelas 1 (*Fully Paid*) dengan keseluruhan kelas 1 (*Fully Paid*) adalah 99%. Nilai *specificity* pada model XGBoost dengan menerapkan *data reduction* adalah 0,29. Hal ini mengindikasikan bahwa rasio kebenaran model dalam meprediksi kelas 0 (*Charged Off*) dengan keseluruhan kelas 0 (*Charged Off*) adalah 29%. Nilai *specificity* model XGBoost tanpa *data reduction* adalah 0,84. Hal ini mengindikasikan bahwa kebenaran model XGBoost dalam memprediksi kelas 0 (*Charged Off*) dengan total kelas 0 (*Charged off*) adalah 84%.

Berdasarkan nilai AUC yang diperoleh model XGBoost tanpa penerapan *data reduction* termasuk kategori model yang sangat bagus dalam memisahkan kedua kelas.

 DOI : 10.37823/insight.v5i1.233

Sehingga model ini layak digunakan untuk mengklasifikasikan status pinjaman. Model XGBoost yang dihasilkan melalui data reduction termasuk kategori model *poor*. Artinya model XGBoost melalui data reduction memiliki kinerja yang buruk dalam memisahkan setiap kelas.

Berdasarkan Tabel XIII, nilai *sensitivity* model Adaboost tanpa *data reduction* dan melalui data *reduction sama* yaitu 0,99. Hal ini mengindikasikan bahwa rasio kebenaran model AdaBoost dalam memprediksi pinjaman kelas 1 (Fully Paid) dengan keseluruhan kelas 1 (Fully Paid) adalah 99%. Nilai *specifcity* pada model AdaBoost dengan menerapkan *data reduction* adalah 0,19. Hal ini mengindikasikan bahwa rasio kebenaran model dalam memprediksi kelas 0 (Charged Off) dengan keseluruhan kelas 0 (Charged Off) adalah 19%. Nilai *specificity model* AdaBoost tanpa *data reduction* adalah 0,79. Hal ini mengindikasikan bahwa kebenaran model AdaBoost dalam memprediksi kelas 0 (Charged Off) dengan total kelas 0 (Charged Off) adalah 79%.

Berdasarkan nilai AUC yang diperoleh model AdaBoost tanpa penerapan *data reduction* termasuk model yang bagus. Sehingga model AdaBoost tanpa penerapan *data reduction* dapat digunakan untuk memprediksi status pinjaman. Model AdaBoost melalui *data reduction* termasuk kategori *fail*. Artinya model AdaBoost melalui data reduction gagal dalam memisahkan kedua kelas. Sehingga model ini tidak dapat digunakan untuk memprediksi status pinjaman.

Perbedaan yang signifikan terjadi antara penerapan *data reduction* dengan tanpa *data reduction,* salah satu penyebabnya adalah jumlah data dengan penerapan *data reduction* lebih sedikit. Saat melakukan *data reduction* lebih dari 20% data dihapus dari data set. Hal ini menyebabkan data train lebih sedikit sehingga tidak cukup baik untuk diterapkan data test. Namun hal ini masih perlu untuk dikaji lebih lanjut.

V. KESIMPULAN

Adapun kesimpulan dari hasil penelitian yang sudah dilakukan adalah terdapat perbedaan performa model yang dibangun melalui eksperimen tanpa *data reduction* dan dengan *data reduction.* Eksperimen tanpa *data reduction* menghasilkan model yang lebih baik daripada eksperimen dengan menerapkan *data reduction*.

Berdasarkan nilai AUC dan bentuk kurva ROC pada model XGBoost memiliki nilai AUC 0,92 yang termasuk kategori model "*very good*". Sedangkan model AdaBoost memiliki nilai AUC 0,89 yang termasuk kategori model "*good*". Artinya model AdaBoost dan XGBoost tanpa data reduction mampu memisahkan kedua kelas dengan baik. Oleh karena itu, kedua model dapat digunakan untuk menguji kelayakan pemberian pinjaman.

DAFTAR PUSTAKA

[1] M. M. Bokhari, "Credit Risk Analysis in Peer to Peer Lending Dataset: Lending Club,"Senior Projects Spring, p. 1- 49, 2019.
[2] V. S. E. S. A. Petropoulus, "A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting," Irving Fisher Committee on Central Bank Statistics, 2018.
[3] H. M. Flood, "Early identification of high-risk credit card customers based on behavioral data," NTNU, 2017.
[4] E. R. Schapire and Y. Freund, "Boosting Foundations and Algorithms," London: Massachusetts Institute of Technology, 2012.
[5] D. G. d. B. H. P. . M. Addo, "Credit Risk Analysis Using Machine and Deep Learning Models," no. Risks, 2018. .
[6] C. Joseph, Advanced Credit Risk Analysis and Management, Published 2013 by John Wiley & Sons, Ltd., 2013.
[7] M. K. d. J. P. J. Han, Data Mining Concepts and Techniques, USA: Elsevier, 2011.
[8] S. Raschha and V. mirjalili, "Performing one-hot encoding on nominal features," dalam python machine learning, pp. 116-118, 2017.
[9] T. Chen dan C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 10 Jun 2016.
[10] D. Berrar, "Cross-validation," https://www.researchgate.net/publication/324701535, vol. 1, p. 9, 2018.
[11] E. F. M. A. H. Ian H. Witten, "Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers is an imprint of Elsevier," 2011.

 DOI : 10.37823/insight.v5i1.233