296-Bistek-Turnitin

by Ama Febriyanti

Submission date: 03-Jun-2023 06:07PM (UTC+0500)
Submission ID: 2108041411
File name: 296-Bistek-Turnitin.docx (538.01K)
Word count: 1908
Character count: 12567

MEMPREDIKSI KEMAMPUAN PEMBAYARAN KREDIT NASABAH DENGAN ALOGARITMA *MACHINE LEARNING*

Abstract

Most people still have difficulty accessing finance because of a lack or even no credit history. This study aims to develop a data model that predicts a customer's ability to pa from various aspects other than credit history. This study uses the CRSIP-DM (Cross Industry Standard Process Model for Data mining) method. The ata used in this study is the Home Credit Default Risk dataset collected by documentation techn ues. The data were then analyzed using data modeling analysis techniques, namely logistic regressor, decision tree classifier, random forest classifier, and lgbm classifier. This study found that the best model for predicting client payment ability is the lgbm classifier or the Random Forest Classifier.

***Keywords**: Credit, Default, Data, Machine Learning*

Abstrak

Mayoritas masyarakat masih kesulitan mendapatkan akses pembiayaan karena kurang atau bahkan tidak adanya riwayat kredit. Penelitian ini bertujuan untuk mengembangkan permodelan data yang mampu memprediksi kemampan membayar nasabah dari berbagai aspek selain riwayat kredit. Penelitian ini menggunakan metode CRSIP-DM (*Cross Industry Standard Process Model for Data mining*). Data yang digunakan dalam penelitian ini adalah dataset *Home Credit Default Risk* yang dikumpulkan dengan teknik dokumentasi. Data kemudian dianalisis dengan menggunakan teknik analisis permodelan data yaitu *logistic regressor, decision tree classifier, random forest classifier*, dan *lgbm classifier*. Penelitian ini menghasilkan temuan bahwa model yang paling baik untuk memprediksi kemampuan pembayaran klien adalah *lgbm classifier* atau *Random Forest Classifier*.

Kata Kunci: kredit, gagal bayar, data, *machine learning*.

PENDAHULUAN

Kredit berperan penting dalam membantu masyarakat memenuhi kebutuhan perekomoniannya. Namun, banyak orang mengalami kesulitan untuk mendapatkan akses kredit karena kurang atau bahkan tidak adanya riwayat kredit (Beck, 2021). Hal ini menimbulkan kebingungan, bahwa untuk mendapatkan kredit masyarakat dituntut memiliki riwayat kredit, padahal riwayat kredit tidak serta merta bisa diperoleh jika tidak ada yang memberikan kredit. Meskipun demikian, tidak ada yang bisa disalahkan dalam masalah ini. Pihak pemberi kredit hanya berupaya mengurangi resiko sedemikian rupa untuk mencegah premasalahan kredit macet atau gagal bayar. Banyaknya nasabah yang gagal bayar akan mengakibatkan keuntungan dan modal dari pemberi kredit berkurang (Nirwana et al., 2022).

Resiko kredit sendiri termasuk dalam 5 teratas resiko yang menjadi tantangan terbesar industri perbankan atau lembaga keuangan lainnya (Syafudin et al., 2021). Penilaian resiko kredit berperan penting untuk menentukan apakah nasabah mampu melunasi hutangnya atau justru berpeluang gagal bayar. Penilaian resiko kredit juga dapat meminimalkan kemungkinan terjadinya agar pemberi kredit mampu menjaga kelangsungan usahanya dan mencegah kerugian (Rizki et al., 2020). Indikator utama yang dapat digunakan untuk menilai resiko kredit adalah rasio kredit bermasalah atau NPL (*non-performing loan*). Pemberi kredit harus senantiasa memonitor NPL karena semakin tinggi nilainya maka semakin tinggi pula jumlah nasabah yang beresiko kredit tinggi.

Prediksi terhadap status kredit gagal bayar (*default*) nasabah dibutuhkan sebagai upaya *monitoring* NPL. Prediksi status gagal bayar juga dapat menjadi alat penilaian resiko kredit. Tidak hanya berpatokan pada riwayat kredit, prediksi perlu dilakukan dengan mempertimbangkan faktor-faktor lainnya yang dapat memicu gagal bayar. Oleh karena itu, penting bagi pemberi kredit untuk mengembangkan prediksi status gagal bayar yang mampu memprediksi kemampuan pembayaran nasabah sehingga mampu menjangkau masyarakat yang belum memiliki riwayat kredit. Prediksi dengan menilai resiko kredit yang dilakukan secara manual cenderung bersifat subjektif dan rawan akan fraud (Syafudin et al., 2021).

Penelitian terdahulu yang berkaitan dengan prediksi status kredit gagal bayar telah mengembangkan model-model prediksi yang memanfaatkan alogaritma *machine learning*. Normah et al. (2022), menggunakan analisis CHAID (*Chi-square Automatic Interaction Detection Analysis*) untuk sekedar mengelompokkan nasabah potensial dan menentukan status kredit guna menekan jumlah kredit macet. Syafudin et al. (2021), memprediksi status pinjaman bank menggunakan *Deep Learning Neural* (DNN) dan menghasilkan akurasi 82.27%. Metode *K-Nearest Neighbor* (KNN) juga pernah digunakan untuk memprediksi pembayaran kredit

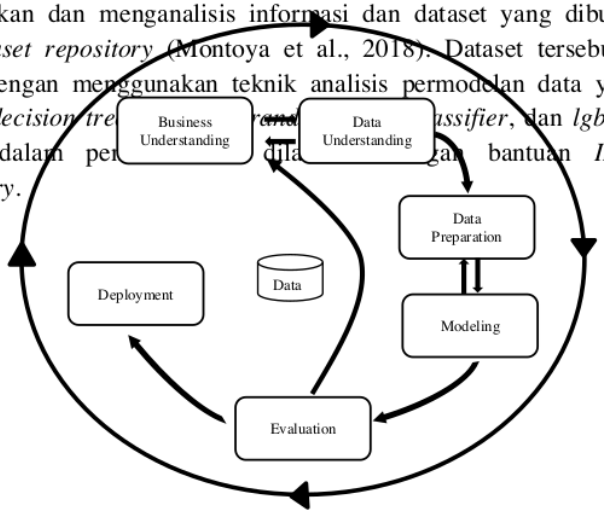
macet di PT FIF Goup Cabang Arjawinangun dan menghasilkan nilai akurasi sebesar 71% (Pratama et al., 2021). Beberapa penelitian sebelumnya tersebut masih memiliki keterbatasan dimana tidak adanya dataset *test* sehingga akurasi diluar *train* tidak diketahui. Selain itu, dataset yang digunakan terbatas pada dataset dengan ukuran kecil. Padahal dataset besar (*big dataset*) memungkinkan perusahaan untuk memantau secara lebih menyeluruh dari segala aspek kehidupan nasabah. Dataset besar juga dapat meningkatkan akurasi model secara signifikan (Beck, 2021).

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk mengembangkan prediksi status kredit gagal bayar dengan alogaritma *machine learning* dan dataset yang besar guna mendapatkan akurasi tinggi dalam memprediksi kemampuan pembayaran nasabah. Bagian selanjutnya akan memaparkan metodologi penelitian. Kemudian, disusul dengan sajian hasil penelitian dan diakhiri kesimpulan penelitian.

METODE PENELITIAN

Penelitian ini termasuk jenis penelitian kuantitatif dengan menggunakan metode CRSIP-DM (*Cross Industry Standard Process Model for Data mining*). CRSIP-DM merupakan metode yang telah digunakan oleh banyak ahli dalam permodelan pengembangan data untuk memecahkan suatu masalah (Givari et al., 2022). Adapun tahapan penelitian disesuaikan dengan *framework* CRSIP-DM yang dapat dilihat pada Gambar 1.

Data yang digunakan dalam penelitian ini adalah dataset *Home Credit Default Risk*. Pengumpulan data dilakukan dengan menggunakan teknik dokumentasi yaitu mengumpulkan dan menganalisis informasi dan dataset yang dibutuhkan dari *kaggle dataset repository* (Montoya et al., 2018). Dataset tersebut kemudian dianalisis dengan menggunakan teknik analisis permodelan data yaitu *logistic regressor, decision tre* *rand* *assifier*, dan *lgbm classifier*. Pengujian dalam pe* dila* *gan bantuan *IDE Google Colaboratory*.



Gambar 1. ***Framework CRSIP-DM***

HASIL DAN PEMBAHASAN

Hasil Penelitian

A. Business Understanding

Home Credit merupakan perusahaan yang berdiri sejak tahun 1997. *Home Credit* menyediakan pembiayaan untuk konsumen internasional yang beroperasi di delapan negara (Home Credit, 2022). *Home Credit* memiliki tujuan bisnis untuk mengembangkan layanan keuangan bagi masyarakat yang kurang terlayani dalam mengakses pembiayaan perbankan. Banyak masyarakat kesulitan mendapatkan pinjaman atau kredit karena kurang atau bahkan tidak adanya riwayat kredit (Givari et al., 2022)**.** Akibatnya, banyak pengajuan kredit yang ditolak. Untuk mengatasi situasi tersebut, *Home Credit* memanfaatkan alternatif data selain riwayat kredit untuk memprediksi kemampuan pembayaran nasabah, termasuk data telekomunikasi dan informasi lainnya. Pada tahun 2018 *Home Credit* dan merilis dataset *Home Credit Default Risk* yang berisi informasi mengenai nasabahnya di *kaggle dataset repository*. *Home Credit* membuka kesempatan penuh kepada semua orang untuk memanfaatkan data tersebut guna mengembangkan model prediksi yang dapat membantu perusahaan. Tujuan dari *data mining* dataset *Home Credit Default Risk* tersebut adalah untuk membantu perusahaan untuk mengetahui kemampuan pembayaran nasabah dan membantu perusahaan membuat keputusan tentang penyaluran kredit secara tepat, efektif, dan. Hal tersebut akan memastikan bahwa perusahaan tidak salah memberikan kredit kepada nasabah yang memiliki kemungkinan gagal bayar.

B. Data Undertsanding

Dataset yang dirilis oleh *Home Credit* merupakan dataset besar yang terdiri dari 9 dataset dengan format file csv. Masing-masing file berisi informasi tertentu terkait nasabah dari *Home Credit*. Dalam penelitian ini, dataset yang digunakan adalah dataset *train* yang digabungkan dengan dataset *bureau*. Dataset *bureau* digunakan karena dataset tersebut memuat informasi riwayat kredit nasabah yang dapat membantu ketika proses *modelling*. Adapun rincian dataset adalah sebagai berikut,

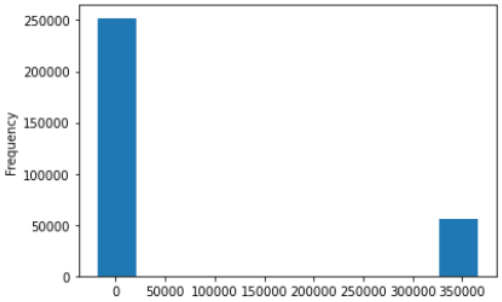
Tabel 1. Rincian Dataset

Dataset	Jumlah data	Jumlah Fitur (Kolom)
Train	307.511	122
Bureau	1.048.575	17
Join train and bureau	307.513	137

Dari 137 fitur data hasil *join train and bureau* tersebut, 77 fitur merupakan data bertipe float, 41 fitur bertipe integer, dan 19 fitur bertipe *object* yang juga disebut data *categorical*.

C. Data Preparation

Pada tahap *data preparation* ini, langkah pertama yang dilakukan adalah mengimport *library* yang dibutukan ke *IDE Google Colaboratory*. Setelah itu, dataset yang digunakn dimuat dengan menggunakan *library pandas*. Dataset yang telah dimuat yaitu *train* dan *bureau* kemudian digabungkan dengan menggunakan metode *left join*, dimana data digabungkan dengan berdasarkan pada dataset *train*. Hasil penggabungan dataset menghasilkan 307.513 data dengan 137 fitur. Setelah data digabungkan, langkah selanjutnya adalah *cleansing data* yaitu dengan mendrop *irrelevant columns*, *handling outliers, handling missing value,* dan *handling duplicate*. Langkah ini dilakukan agar alogaritma *machine learning* dapat memproses dan menghasilkan nilai terbaiknya (Zuama et al., 2022). Kolom atau fitur yang didrop adalah 'SK_ID_BUREAU' dan 'SK_ID_CURR' karena hanya berisi nomor identitas transaksi yang tidak relevan untuk digunakan dalam model prediksi. *Outlier* dalam dataset ini terdapat pada fitur 'DAYS_EMPLOYED' dan ditangani dengan mengubah (*replace*) menjadi nan.



Gambar 2. Histplot days employed

Dari diagram plot tersebut, diketahui bahwa fitur bernilai negative lebih besar dibandingkan fitur bernilai positif. Dan ketika dilakukan pengecekan, jumlah *oiutlier* yang terdapat pada fitur 'DAYS_EMPLOYED' adalah sejumlah 55.374. Adapun untuk *missing value* ditangani dengan metode fillna menggunakan nilai mean untuk setiap fitur numeric, sedangkan untuk fitur categorical dilakukan fillna dengan nilai baru yaitu 'Data_Not_Available' atau data tidak tersedia. Pada dataset

hasil *join train and bureau* tidak ditemukan data yang *duplicate*, sehingga tidak diperlukan *handling duplicate*.

Setelah data dibersihkan, perlu dilakukan transforming data. Hal ini karena model hanya akan bekerja dengan baik ketika keseluruhan data berbentuk numerical. Oleh karena itu, data selain numerical atau data categorical perlu ditransformasikan terlebih dahulu. Dalam hal ini digunakan *label encoder* untuk mentransformasikan data secara otomatis menjadi nilai *dummy*, sebagai berikut

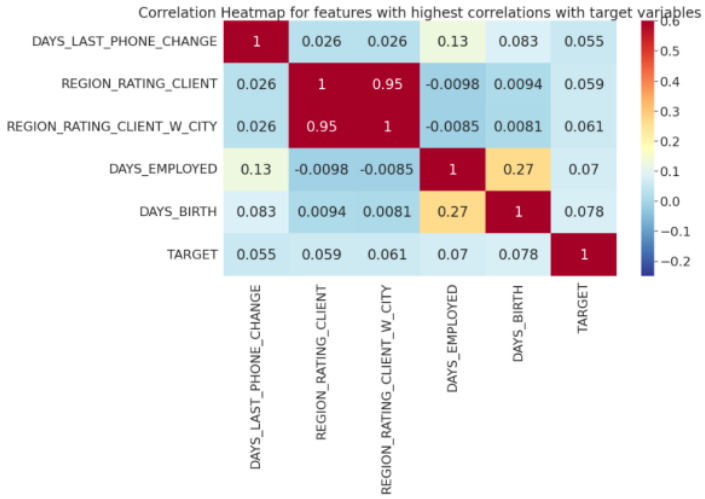
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	351000.0	Unaccompanied
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	1129500.0	Family
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	135000.0	Unaccompanied
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	297000.0	Unaccompanied
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	513000.0	Unaccompanied

⇩

	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_
0	1	0	1	0	1	0	202500.0	406597.5	24700.5	351000.0	6	
1	0	0	0	0	0	0	270000.0	1293502.5	35698.5	1129500.0	1	
2	0	1	1	1	1	0	67500.0	135000.0	6750.0	135000.0	6	
3	0	0	0	0	1	0	135000.0	312682.5	29686.5	297000.0	6	
4	0	0	1	0	1	0	121500.0	513000.0	21865.5	513000.0	6	

Gambar 3. *Data Encoding*

Langkah berikutnya adalah *Exploratory Data Analysis* (EDA) yang menghasilkan visualisasi *seaborn heatmap* yang menggambarkan 5 variabel dengan korelasi terkuat terhadap target, sebagai berikut:



Gambar 4. Heatmap Fitur dengan Korelasi Paling Positif

D. Modeling

Variabel dependen dalam permodelan data yang dikembangkan adalah 'TARGET', sedangkan variabel independennya adalah semua fitur kecuali 'TARGET' dan fitur-fitur yang telah didrop sebelumnya. Terdapat 4 permodelan alogaritma *machine learning* yang digunakan untuk memproses variabel tersebut yaitu *logistic regressor, decision tree classifier, random forest classifier*, dan *lgbm classifier*. Dari 4 model tersebut akan dicari model terbaik berdasarkan outputnya.

E. Evaluation

Untuk mengevaluasi keempat model pengembangan data, digunakan penilaian *Skor accuracy, Skor recall, Skor AUC Performance, Skor F1, Skor precision, dan Confusion matrics* sebagai berikut,

Tabel 2. Skor Evaluasi Model

Model	Skor				
	Accuracy	***Recall***	***AUC Performance***	***F1***	***Precision***
Logistic Regressor	0,92	0,01	0,50	0,02	0,46
Decision Tree Classifier	0,85	0,16	0,54	0,02	0,14
Random Forest Classifier	0,92	0,00	0,50	0,02	0,73
Lgbm Classifier	0,92	0,00	0,50	0,02	0,73

```
Confusion logistic regressor:  [[84667    87]
 [ 7427    75]]

Confusion decision tree regressor:  [[77209  7545]
 [ 6303  1199]]

Confusion random forest classifier:  [[84750     4]
 [ 7491    11]]

Confusion lgbm classifier:  [[84750     4]
 [ 7491    11]]
```

Gambar 5. Skor Confusion metrics

Dari hasil evaluasi di atas, diketahui bahwa dari keempat model yang dikembangkan, model yang paling baik digunakan adalah *lgbm classifier* atau *Random Forest Classifier*. Berikut merupakan output prediksi menggunakan dua model terbaik tersebut, yaitu sebagai berikut,

	Actual Outcome	prob_0	prob_1	predicted_TARGET
64944	NaN	0.994987	0.005013	0.0
64694	NaN	0.994423	0.005577	0.0
55294	NaN	0.994308	0.005692	0.0
78840	NaN	0.994229	0.005771	0.0
63165	NaN	0.994212	0.005788	0.0

Gambar 6. Output prediksi dengan model *lgbm classifier*

	Actual Outcome	prob_0	prob_1	predicted_TARGET
0	NaN	1.0	0.0	0.0
59954	NaN	1.0	0.0	0.0
59962	NaN	1.0	0.0	0.0
59961	NaN	1.0	0.0	0.0
59960	NaN	1.0	0.0	0.0

Gambar 7. Output prediksi dengan model *Random Forest Classifier*

F. Deployment

Deployment merupakan proses pelaporan hasil dari *modelling* dan *evaluation* pada pengembangan data yang telah dilakukan. Pelaporan ini dapat digunakan sebagai bahan pertimbangan dalam membuat keputusan pemberian kredit yang tepat dengan memprediksi dan mengidentifikasi segala aspek yang terkait dan mencegah kemungkinan gagal bayar. (Givari et al., 2022). *Deployment* dapat dilakukan dengan menggunakan *software* seperti *power BI* atau google data studio. Adapun *deployment* dari penelitian ini adalah sebagai berikut,

Gambar 8. ***Dahsboard Home Credit Default Risk***

Dashboard tersebut merepresentasikan ringkasan data, distribusi data, dan statistic deskriptif dari dataset *Home Credit Default Risk*.

Pembahasan

Berdasarkan hasil pengembangan model prediksi yang dilakukan, diketahui bahwa model terbaik untuk memprediksi dataset *Home Credit Default Risk* yang berukuran besar adalah model *lgbm classifier* atau *Random Forest Classifier*. *Lgbm classifier* menghasilkan skor masing-masing sama yaitu 0,92 atau 91,88%. Hal tersebut berarti dari keseluruhan data, sebesar 91,88% telah diklasifikasikan dengan benar dan diprediksi tidak akan menimbulkan gagal bayar atau kredit macet. Hasil tersebut dapat dikatakan baik karena skor yang mendekati 100%.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan untuk memprediksi kemampuan pembayaran kredit nasabah, dapat disimpulkan bahwa keempat model yang dikembangkan yaitu *logistic regressor, decision tree classifier, random forest classifier*, dan *lgbm classifier* menghasilkan nilai *accuracy* lebih dari 50%. Namun, dengan mempertimbangkan skor evaluasi lainnya, maka model yang paling baik digunakan adalah *lgbm classifier* atau *Random Forest Classifier*. Keterbatasan dalam penelitian ini adalah dataset yang digunakan dan digabungkan terbatas pada

train dan *buerau*. Oleh karena itu, saran bagi penelitian mendatang hendaknya mengeksplore dataset lainnya untuk mendapatkan fitur-fitur yang dapat mendukung model prediksi yang dikembangkan.

REFERENSI

296-Bistek-Turnitin

ORIGINALITY REPORT

5%	5%	1%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	github.com Internet Source	2%
2	repository.usahidsolo.ac.id Internet Source	1%
3	Asif Ahmed Neloy, Sazid Alam, Rafia Alif Bindu, Nusrat Jahan Moni. "Machine Learning based Health Prediction System using IBM Cloud as PaaS", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019 Publication	1%
4	www.scribd.com Internet Source	1%
5	3lib.net Internet Source	1%
6	journal.student.uny.ac.id Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		