Analisis Perbandingan Penggunaan Model Machine Learning Pada Kasus Deteksi Kemampuan Calon Klien Dalam Membayar Kembali Pinjaman

[1]Muhammad Afifudin, [2]Agung Mustika Rizki
[1,2]Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: muhafifudin27@gmail.com[1]

Abstrak. *Semakin berkembangnya ekonomi suatu wilayah, makin tinggi pula kebutuhan belanja masayrakat diwilayah tersebut. Tak jarang hal ini memicu fenomena maraknya masyarakat yang mengajukan pinjaman kredit. Oleh karena itu, pihak peminjam kredit membutuhkan suatu metrik yang dapat memprediksi apakah calon kliennya mampu membayar Kembali pinjaman sebelum menyetujui ajuan kredit. Penelitian ini menganalisis perbandingan dari berbagai model machine learning sebagai alat prediksi calon klien peminjam kredit. Menggunakan dataset Home Credit Default Risk dari Kaggle, dan menerapkan metode CRISP-DM dalam pengembangannya. Beberapa model yang dipilih yaitu Regresi logistic, random forest, Gaussian naïvee bayes, decision tree, dan Multi-layer perceptron. Dari kelima model tersebut, random forest menunjukkan hasil skor evaluasi yang paling baik dengan metrik evaluasi ROC AUC. Yakni dengan nilai ROC AUC train score sebesar 1 dan ROC AUC test score sebesar 0,7233. .*

Kata Kunci: prediksi, kredit, machine learning, ROC AUC

Semakin berkembangnya perekonomian suatu negara, makin meningkat pula kebutuhan masyarakatnya. Sayangnya, tidak semua individu memiliki dalam masyarakat memiliki privilege yang sama dalam hal materi. Hal ini menyebabkan beberapa kelompok populasi dalam masyarakat sering melakukan aktivitas peminjaman uang atau kredit. Aktivitas peminjaman uang ini dilakukan dengan tujuan antara lain, membayar tagihan, memenuhi kebutuhan sehari-hari, membeli alat transportasi, hingga membeli barang-barang yang sifatnya konsumtif.

Dengan meningkatnya aktivitas peminjaman uang, banyak pihak yang harus bersusah payah dalam hal pengajuan pinjaman dikarenakan kurangnya atau bahkan tidak adanya Riwayat credit yang dimiliki. Masalah ini kemudian sering dimanfaatkan oleh oknum-oknum peminjam tak bertanggung jawab dengan memanfaatkan keadaan para klien yang terdesak. Dengan iming-iming persyaratan yang mudah dan cepat, calon klien dapat dengan mudah tergiur untuk mengajukan pinjaman. Masalah yang terjadi kemudian adalah oknum peminjam akan menerapkan bunga dan batas waktu pembayaran yang "mencekik" sehingga membuat klien tidak bisa memenuhi perjanjian awal dan menuju pada status gagal bayar.

Salah satu metode peminjaman yang paling marak digunakan adalah dengan kartu kredit. Sayangnya, peningkatan penggunaan kartu kredit juga diimbangi dengan peningkatan angka penipuan dan gagal bayar. Perkembangan internet pun menjadikan kepemilikan kartu kredit lebih mudah. Adanya penipuan maupun kien yang gagal bayar akan menjadi masalah yang serius bagi perekonomian masyarakat. Banyaknya kredit macet akan menyebabkan lumpuhnya suatu ekosistem ekonomi pada suatu wilayah. Oleh karena itu, pihak peminjam membutuhkan adanya suatu metode efektif guna mengurangi kerugian [1].

Home credit berupaya untuk mengatasi masalah kebutuhan peminjaman tersebut dengan memperluas inklusi keuangan bagi populasi masyarakat yang belum memiliki pengalaman dalam dunia perkreditan dengan menyediakan pengalaman kredit yang positif dan aman [2]. Dalam sejarahnya, Home credit telah menggunakan berbagai cara guna meningkatkan pengalaman kredit para klien. Salah satunya adalah dengan menggunakan data yang mereka punya untuk memprediksi kemampuan pembayaran calon klien mereka. Prediksi ini penting untuk dilakukan guna memastikan bahwa klien yang mampu membayar dijamin tidak akan ditolah saat mengajukan pinjaman dan mengantisipasi adanya potensi mangkir/telat bayar/gagal bayar pada klien

Guna mewujudkan prediksi calon klien, dapat digunakan berbagai metode statistic dan machine learning. Machine learning merupakan disiplin ilmu dari rumpun Computer Science

yang mendalami bagaimana merancang mesin/computer hingga memiliki kecerdasan menyerupai manusia. Salah satu metode dalam machine learning adalah concept learning, metode ini membutuhkan data training dan mampu mengatasi data negative maupun positif karena termasuk ke dalam Supervised Learning [3]. Supervised learning merupakan algoritma yang bergantung pada data input berlabel guna mempelajari data dan menghasilkan lable yang sesuai ketika diberi data baru tanpa label [4].

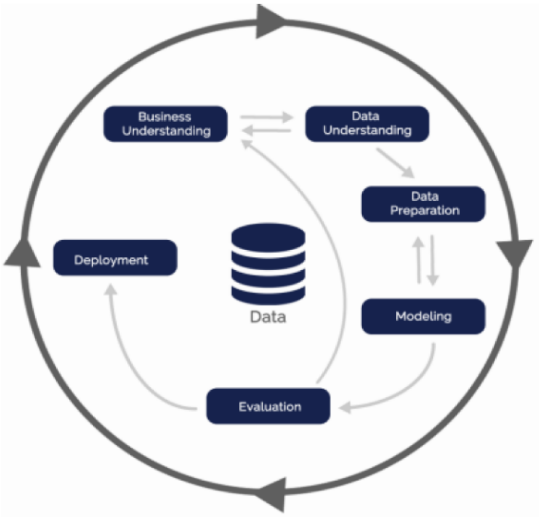
Dalam kasus prediksi, jika tujuan dari prediksi adalah untuk menentukan kategori data dari satu baris table, dengan kata lain melabeli/menamai data, dapat digunakan metode klasifikasi [5]. Sebaliknya, jika tujuan akhir dari suatu project machine learning adalah memprediksi variable dari data dalam suatu baris kolom yang bernilai kontinyu, maka digunakan metode regresi.

Religia, Nugroho, dan Hadikristanto dalam penelitiannya berjudul Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. Penerapan algortima Random forest sebagai klasifikasi prediksi penerimaan pengajuan pinjaman menyebutkan bawhwa, pengujian telah dilakukan menghasilakn performa dari klasifikasi bank marketing dataset. Algoritma random forest yang dipilih menghasilkan akurasi sebesar 88,30% dan AUC sebesar 0,500[6].

Sahroni, Seftiani, dan Fitriana dalam penelitian membandingkan penggunaan tiga algoritma Naïve Bayes, k-Nearest Neighbor, dan Neural Network untuk permasalahan imbalanced data pada dataset credit card fraud. Hasil pengujian menunjukan bahwa algituma Neural Network menghasilkan akurasi 93,59% dan AUC sebesar 0,977, Naïve bayes dengan skor akurasi 91,26% dan AUC sebesar 0,956, dan k-NN menghasilakn akurasi 64,84% dan AUC 0,709 [7]. Dalam penelitian ini, Algotima yang digunakan untuk mencapai prediksi mampu tidaknya klien dalam membayar pinjaman adalah algoritma klasifikasi. Pada penelitian kali ini, akan digunakan beberapa algoritma supervised learning antara lain, regresi logistic, random forest, Gaussian Naïve Bayes, dan Multi-Layer Perceptron.

I. Metodologi

Penelitian ini dilaksanakan menggunakan acuan metode CRISP-DM (Cross Industry Standard Process Model for Data Mining) dengan enam tahapan yaitu (1) Business Understanding; (2) Data Understanding; (3) Data Preparation; (4) Modeling; (5) Evaluation; (6)[7], [8]. Berikut gambaran alur proses tahapan-tahapan CRISP-DM:



Gambar 1. alur Proses CRISP DM

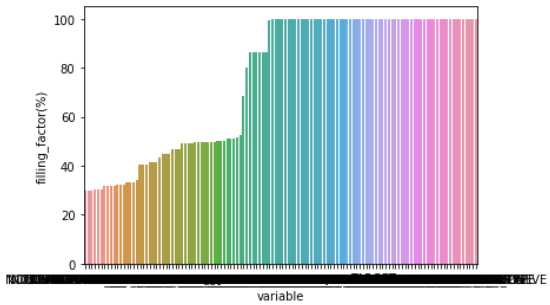
Sumber data diperoleh dari dataset Kaggle dengan judul “Home Credit Default Risk”. Dalam penelitian ini, hanya digunakan 2 dari 9 table, yaitu table application_train.csv dan application_test.csv. Kedua table tersebut merupakan table utama yang dibagi menjadi 2 bagian, table train (dilengkapi kolom TARGET) dan test (tanpa kolom TARGET). Setiap baris data dalam dataset mewakili satu kali transaksi peminjaman.

Business Understanding

Permasalahan yang dialami pada kasus ini adalah sulitnya menentukan metrik yang tepat dalam menentukan diterima atau tidaknya acuan pinjaman dari calon klien. Salah satu cara menentukan matrik tersebut adalah dengan memprediksi apakah calon klien dapat membayar pinjaman dengan tepat waktu. Prediksi tersebut dapat dilakukan dengan melakukan metode-metode statistic dan machine learning terhadap data-data dan dokumen yang disertakan calon klien saat mengajukan pinjaman.

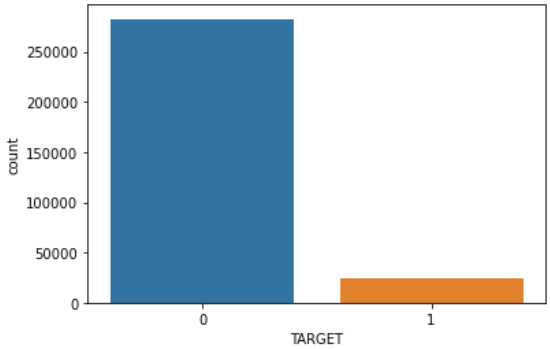
Data Understanding

Dataset yang dipakai memiliki 122 kolom yang terkait dengan semua data klien, ditambah satu kolom ‘TARGET’ yang berisi variable yang menunjukkan apakah klien membayar pinjaman tepat waktu atau tidak (0: tepat waktu; 1: telat)



Gambar 2. Filling factor dari masing-masing kolom

Dilihat dari filling factornya, terdapat banyak missing value dalam dataset sehingga pada tahap selanjutnya diperlukan proses-proses seperti feature engineering, feature importance, dan imputasi.



Gambar 3. Bar chart perbandingan data

Gambar diatas merupakan hasil visualisasi dari kolom TARGET. Dapat dilihat dari visualisasi tersebut bahwa dataset yang digunakan tidak seimbang atau imbalance. Oleh sebab itu, pada tahap evaluasi nanti tidak cocok bila menggunakan metrik evaluasi accuracy score. Metrik evaluasi yang lebih cocok untuk data imbalance dengan kasus klasifikasi pada penelitian ini adalah metrik ROC AUC.

Data Preparation

Data preparation adalah tahapana yang paling memakan waktu serta Langkah paling penting dalam seluruh siklus CRISP-DM [9]. Dalam tahap ini, akan terjadi serangkaian proses guna mendapatkan dataset yang bersih dan layak digunakan dalam algoritma. Tahap ini meliputi beberapa proses diantaranya:

- Pemilihan data, Proses ini dilakukan dengan menyeleksi beberapa atribut atau kolom yang akan dimasukkan dalam algoritma. Dari 122 kolom yang ada dalam dataset, akan dipilih 20 kolom. Kolom-kolom dipilih berdasarkan nilai korelasi dan presentase missing value yang dimiliki.
- Penanganan Missing Value, pada tahap sebelumnya sudah diketahui bahwa banyak kolom yang memiliki missing value. Missing value merupakan keadaan dimana sejumlah nilai dalam kolom di dataset kosong [10]. Penanganan missing value bisa dilakukan dengan metode imputasi. Untuk data numerical akan digunakan metode median, untuk data categorical akan digunakan metode mode(modus).
- Transformasi data, ada dua metode yang sering dipakai pada tahapan ini, yaitu one-hot encoding dan label encoding. Berbeda dengan one-hot encoding yang membuat terciptanya kolom baru untuk setipa nilai unik pada kolom, label encoding hanya melakukan pengubahan nilai kolom menjadi integer. Pada penelitian kali ini, transformasi data dilakukan dengan label encoding.

Modeling

Pada tahap ini, data yang telah disiapkan, dibersihkan, dan ditransformasi akan dimasukkan pada algoritma klasifikasi yang telah dipilih yaitu , regresi logistic, random forest, Gaussian Naïve Bayes, Decision Tree, dan Multi-Layer Perceptron. Proses ini dilakukan dengan pengkodean Bahasa pemograman python dengan bantuan library scikit learn.

Evaluation

Pada tahap ini akan dilakukan pengukuran terhadapa hasil prediksi yang dihsilkan berbagai algoritma yang digunakan dengan metrik evaluasi. Metrik evaluasi yang digunakan adalah Metode ROC AUC.

Deployment

Tahap ini adalah tahap akhir dari penilitian ini. Pada tahap ini akan disajikan hasil penelitian melalui tulisan ini dan memaparkan saran-saran dan rekomendasi yang sekiranya diperlukan untuk membantu penelitian yang akan dilakukan kedepannya

II. Hasil dan Pembahasan

Telah dilakukan modeling dengan algoritma klasifikasi yang telah dipilih yaitu, Regresi logistic, random forest, Gaussian Naïve Bayes, dan Multi-Layer Perceptron. Setelah modeling akan dilakukan evaluasi terhadap hasil modeling dengan metrik ROC (Receiver

Operating Characteristics) AUC (Area Under the Curve). Tabel 1 menunjukkan data hasil evaluasi dari semua model machine learning yang dipilih.

Tabel 1. Hasil evaluasi ROC AUC score

Algortima	ROC AUC train score	ROC AUC test score
Logistic Regression	0,6221	0,6244
Random Forest	1	0,7233
Gaussian Naive Bayes	0,6126	0,6129
Decision Tree	1	0,5389
Multi-Layer Perceptron	0,5163	0,5219

Pada table 1, dapat dilhat hasil evaluasi dari semua algoritma yang dipilih. Berdasarkan hasil evaluasi ROC AUC, model terbaik yang bisa dipakai dalam kasus prediksi kemampuan calon klien untuk membayar Kembali pinjaman adalah algoritma Random forest dengan ROC AUC score train sebesar 1, dan ROC AUC score test sebesar 0,7233

Sementara model algortima yang paling tidak cocok untuk digunakan dalam prediksi kemampuan calon klien untuk membayar Kembali pinjaman adalah Multi-Layer Preceptron. Bisa dilihat dari hasil ROC AUC score train 0,5163 dan ROC AUC score test sebesar 0,5219.

Sementara itu, 3 model lain yang dipakai, yaitu regresi logistic, gaussian naïve bayes, dan decision tree, menghasilakn skor pada evaluasi ROC AUC masing-masing sebesar 0,6221; 0,6126; dan 1 pada ROC AUC score train, dan sebesar 0,6244; 0,6129; dan 0,5389 pada ROC AUC score test masing-masing.

III. Kesimpulan

Berdasarkan hasil penelitian dari prediksi kemampuan calon klien untuk membayar Kembali pinjaman dengan metode algortima Regressi logistic, Random forest, Gussian Naïve Bayes, Decision tree dan Multi-layer perceptron, dapat diambil kesimpulan bahwa:

- Dari Home credit dataset yang tersedia dari Kaggle yang terdiri dari 122 kolom. Merupakan data yang masih kotor dan imbalance, sehingga diperlukan pembersihan dan persiapan terlebih dahulu untuk menggunakan data tersebut. Karena keadaanya yang imbalance pula diperlukan treatment-teatment tertentu untuk dapat menggunakan data tersebut, contohnya adalah pemilihan metode evaluasi metrik ROC AUC.
- Dari kelima algortima yang dipakai dalam penelitian ini, algortima Random forest memperoleh skor tertinggi dari evaluasi ROC AUC score train dan test sebesar 1 dan 0,7233. Tinggi nya skor yang diperleh oleh random forest juga dipengaruhi oleh kelebihan-kelebihan yang ditawarkan algoritma ini, antara lain: (1) Tidak terpengaruh dengan outlier; (2) Dapat bekerja optimal pada data non-linier; (3) Risiko Overfitting rendah; (4) Dapat bekerja dengan baik pada kumpulan data yang besar;

Adapun saran yang bisa dilakukan untuk penelitian selanjutnya, adalah untuk mencoba menggunakan metode algoritma lain yang sifatnya lebih advance. Dikarenakan kelima algoritma yang digunakan pada penelitian ini merupakan algoritma klasifikasi dasar. Disarankan menggunakan Nearest neighboor atau XGBoost lalu bandingkan Kembali hasil evaluasinya dengan algoritma random forest.

I. Daftar Pustaka

[1] K. M. K. Anna Montoya inversion, “Home Credit Default Risk.” Kaggle, 2018. [Online]. Available: https://kaggle.com/competitions/home-credit-default-risk

[2] A. Montoya, inversion, K. Odintsov, and M. Kotek, “Home Credit Default Risk,” 2018. https://kaggle.com/competitions/home-credit-default-risk (accessed Dec. 21, 2022).

[3] J. A. Pratama et al., “The Analisis Sentimen Sosial Media Twitter Dengan Algoritma Machine Learning Menggunakan Software R,” Jurnal Fourier, vol. 6, no. 2, pp. 85–89, Oct. 2017, doi: 10.14421/FOURIER.2017.62.85-89.

[4] K. Kristiawan, D. D. Somali, A. Widjaja, and others, “Deteksi Buah Menggunakan Supervised Learning dan Ekstraksi Fitur untuk Pemeriksa Harga,” Jurnal Teknik Informatika dan Sistem Informasi, vol. 6, no. 3, 2020.

[5] “Supervised learning: predicting an output variable from high-dimensional observations — scikit-learn 1.2.0 documentation.” https://scikit-

learn.org/stable/tutorial/statistical_inference/supervised_learning.html (accessed Dec. 22, 2022).

[6] Y. Religia, A. Nugroho, and W. Hadikristanto, "Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 1, pp. 187–192, Feb. 2021, doi: 10.29207/RESTI.V5I1.2813.

[7] M. Y. Sahroni, N. A. Setifani, and D. N. Fitriana, "Analisis perbandingan algoritma Naïve Bayes, k-Nearest Neighbor dan Neural Network untuk permasalahan class-imbalanced data pada kasus credit card fraud dataset," Teknologi: Jurnal Ilmiah Sistem Informasi, vol. 11, no. 2, pp. 69–73, Jun. 2021, doi: 10.26594/TEKNOLOGI.V11I2.2393.

[8] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," Jurnal Teknologi Informatika dan Komputer, vol. 6, no. 2, pp. 12–20, 2020.

[9] A. Bengnga and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," Jambura Journal of Electrical and Electronics Engineering, vol. 4, no. 2, pp. 169–174, Jul. 2022, doi: 10.37905/JJEEE.V4I2.14403.

[10] A. Alfarisi, A. R. Alfarisi, H. Tjandrasa, and I. Arieshanti, "Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor," Jurnal Teknik ITS, vol. 2, no. 1, pp. A73–A76, Mar. 2013, doi: 10.12962/j23373539.v2i1.2735.