MAKALAH SAINS DATA

Implementasi Algoritma *Random Forest* Untuk Peningkatan Akurasi *Credit Scoring*

Disusun untuk memenuhi tugas akhir mata kuliah Sains Data

Dosen Pengampu:
Devvi Sarwinda, M.Kom.

Disusun oleh:
Kelompok 2 Sains Data (B)

Haifa Marwa Saniyyah	2206048783
Halimah As-Sajidah	2206048820
Hanny Awlia	2206048751
Rahma Chuzaima	2206048732
Reizka Fathia	2206052755

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS INDONESIA

2024

ABSTRAK

Credit scoring adalah metode penting yang digunakan lembaga keuangan untuk menilai kelayakan kredit individu atau perusahaan. Akurasi dalam penilaian ini sangat krusial karena berhubungan langsung dengan risiko kredit dan pengambilan keputusan finansial. Oleh karena itu, penelitian ini dilakukan guna menentukan dan mengembangkan model yang paling akurat dalam memberikan sebuah keputusan kredit, dengan membandingkan model Random Forest, Support Vector Machine, dan Decision Tree. Data yang digunakan dalam penelitian ini berjumlah 32.581 data dengan 12 fitur. Penelitian ini melibatkan tahap pra-pemrosesan data, termasuk imputasi *missing value* dan penanganan *outlier,* serta pembagian dataset menjadi data pelatihan (*training*) dan pengujian (*testing*). Model ini kemudian dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *f1-score* untuk mengukur kinerja dan keandalannya. Hasil menunjukkan bahwa algoritma Random Forest mampu meningkatkan akurasi *credit scoring* secara signifikan dengan akurasi sebesar 93%. Mengacu pada tingkat akurasi, maka model algoritma Random Forest termasuk kategori klasifikasi sangat baik.

Kata Kunci: Data, Klasifikasi, Random Forest, *Credit scoring*, Risiko kredit.

1. PENDAHULUAN

1.1 Latar Belakang

Menurut Undang-Undang Republik Indonesia nomor 10 tahun 1998 tentang perbankan, bank adalah badan usaha yang menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya kepada masyarakat dalam bentuk kredit dan atau bentuk-bentuk lainnya dalam rangka meningkatkan taraf hidup rakyat banyak. Kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dengan pihak lain yang mewajibkan pihak yang meminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga. Kredit memiliki risiko, di mana risiko tersebut dipengaruhi oleh latar belakang debitur (pemohon pengajuan kredit), seperti pendapatan, tujuan pengajuan kredit, riwayat kredit macet sesuai catatan biro kredit, dan masih banyak lainnya. Hal ini dapat menyebabkan terjadinya kredit macet yaitu situasi dimana debitur tidak mampu membayar utang pinjaman.

Oleh karena itu, penting bagi pihak bank sebagai kreditur untuk melakukan analisis kredit terhadap debitur untuk dapat memutuskan penyetujuan pengajuan kredit debitur. Sistem *credit scoring* sangat diperlukan dalam memutuskan pemberian kredit untuk menghindari kredit macet yang dapat menyebabkan kerugian bagi pihak kreditur (Aji & Dhini, 2019).

Dalam penelitian ini, kami akan membandingkan tiga metode klasifikasi untuk memprediksi kelancaran kredit debitur, yaitu Decision Tree, Support Vector Machine (SVM), dan Random Forest. Hasil akhirnya akan menunjukkan performa untuk setiap model, dan model dengan persentase performa paling tinggi akan digunakan untuk memprediksi kelancaran kredit debitur.

1.2 Tujuan

Penelitian ini bertujuan untuk memprediksi kelancaran kredit debitur menggunakan metode klasifikasi dengan

performa terbaik. Tujuan akhirnya adalah memberikan rekomendasi kepada pihak kreditur sehingga pihak kreditur dapat menganalisis kredit sebelum memutuskan penyetujuan kredit.

2. PEMBAHASAN

2.1 Data

Data yang digunakan merupakan dataset publik yang diperoleh dari *kaggle*, berisi beberapa informasi penting dan karakteristik pemohon kredit seperti usia, pendapatan, status kepemilikan rumah, jumlah pinjaman, suku bunga pinjaman, dan lain-lain. Dataset ini (credit_risk_dataset.csv) memiliki 32581 data dengan 12 fitur. Masing-masing fitur akan dijelaskan pada tabel berikut.

Tabel 1. Deskripsi Fitur

Fitur	Deskripsi Fitur
person_age	Usia individu yang mengajukan kredit.
person_income	Penghasilan tahunan individu.
person_home_ownership	Jenis kepemilikan rumah individu. - *rent*: Individu saat ini sedang menyewa properti. - *mortgage*: Individu tersebut memiliki hipotek atas properti yang mereka miliki. - *own*: Individu memiliki rumah mereka secara langsung. - *other*: Kategori lain dari kepemilikan rumah yang mungkin spesifik untuk dataset.
person_emp_length	Masa kerja individu dalam tahun.
loan_intent	Maksud di balik pengajuan kredit.
loan_grade	Nilai yang diberikan kepada pinjaman berdasarkan kelayakan kredit peminjam. - A: Peminjam memiliki kelayakan kredit yang tinggi, yang mengindikasikan risiko yang rendah. - B: Peminjam memiliki risiko yang relatif rendah, namun tidak memiliki kelayakan kredit seperti Grade A. - C: Peminjam memiliki kelayakan kredit yang moderat. - D: Peminjam dianggap memiliki risiko yang lebih tinggi dibandingkan dengan nilai sebelumnya. - E: Kelayakan kredit peminjam

	lebih rendah, mengindikasikan risiko yang lebih tinggi. - F: Peminjam memiliki risiko kredit yang signifikan. - G: Kelayakan kredit peminjam paling rendah, menandakan risiko paling tinggi.
loan_amnt	Jumlah kredit/pinjaman yang diminta oleh individu.
loan_int_rate	Suku bunga yang terkait dengan kredit.
loan_status	Status kredit, dimana 0 menandakan kredit lancar dan 1 menandakan kredit macet.
loan_percent_income	Persentase pendapatan yang diwakili oleh jumlah pinjaman.
cb_person_default_on_file	Riwayat kredit macet individu sesuai catatan biro kredit. - Y: Individu tersebut memiliki riwayat kredit macet pada file kredit mereka. - N: Individu tidak memiliki riwayat kredit macet.
cb_person_cred_hist_length	Panjang riwayat kredit untuk individu tersebut.

2.2 Metode

Dalam penelitian ini, kami menggunakan metode yang paling komprehensif untuk meneliti dan meningkatkan kinerja model prediksi kelayakan kredit. Metode tersebut melibatkan beberapa tahap penting, dimulai dengan *preprocessing data*. Selanjutnya, kami membagi dataset menjadi dua yaitu data untuk pelatihan (*training*) dan pengujian (*testing*) untuk memvalidasi model. Proses berikutnya adalah *model selection* dan *hyperparameter tuning*, di sini kami menggunakan metode GridSearch CV untuk mengoptimalkan model. Kami menyusun parameter-parameter model ke dalam kamus `model_params`, yang berisi daftar model yang akan dievaluasi bersama dengan *hyperparameter* yang akan dioptimalkan untuk setiap model. Setelah proses *tuning*, kami melakukan *fitting model* menggunakan algoritma Random Forest yang telah dioptimalkan.

2.2.1 Preprocessing

Pada tahap ini, kami mengatasi masalah nilai yang hilang (*missing value*), melakukan *encoding* pada data kategorik, serta menangani *outlier* untuk memastikan bahwa data yang digunakan dalam analisis adalah yang paling representatif.

- **Menangani Missing Values**

Adanya *missing values* pada data memengaruhi keakuratan dalam analisis. Untuk mengatasi masalah ini, kami menggunakan teknik imputasi *mean*. Teknik ini melibatkan perhitungan nilai

rata-rata dari setiap kolom yang memiliki data yang hilang, lalu menggantikan nilai-nilai yang hilang tersebut dengan nilai rata-rata yang telah dihitung.

- **Menangani Outliers**

Pada tahapan penanganan *outlier,* kami menggunakan *Inter Quartile Range* (IQR). Proses dimulai dengan menghitung kuartil pertama (Q1) dan kuartil ketiga (Q3) dari kolom data. IQR dihitung sebagai selisih antara Q3 dan Q1. Selanjutnya, ditentukan batas bawah (Q1 - (1.5 × IQR)) dan batas atas (Q3 + (1.5 × IQR)). Nilai yang berada di luar batas dianggap sebagai outlier dan dibatasi pada nilai batas bawah dan atas untuk mengurangi pengaruhnya. Selain itu, untuk mengurangi nilai ekstrem yang tidak realistis, seperti usia lebih dari 80 tahun atau lama kerja lebih dari 60 tahun, baris yang memenuhi kondisi ini dihapus dari dataset.

- **Encoding Fitur Kategorik**

Dalam tahap *encoding* fitur kategorik, kami menggunakan One Hot Encoder untuk mengubah variabel kategorikal menjadi bentuk numerik. Dengan One Hot Encoding, setiap nilai kategori diubah menjadi vektor biner yang hanya memiliki nilai 0 atau 1.

- **Splitting Dataset**

Untuk dataset ini, fitur target utama yang ingin diprediksi adalah ‘loan status’. Fitur target tersebut dapat dipisahkan dari fitur-fitur lainnya, misal variabel y untuk fitur target dan variabel X untuk fitur-fitur lainnya.

Selain itu, diperlukan pula data *training* dan data *testing*. Oleh karena itu, perlu dilakukan *splitting* dataset, yaitu memecah dataset menjadi data *training* dan data *testing*. Pada proses ini, data dipecah menjadi dua bagian, yaitu data *training* sebanyak 80% dan data *testing* sebanyak 20%.

2.3 Implementasi Program dan Pembahasan

Sebelum membuat model, kami melakukan *model selection* untuk melihat model mana yang memiliki performa terbaik dari beberapa metode klasifikasi, yaitu SVM, Decision Tree, dan Random Forest.

	model	best_score
0	random_forest	0.931573
1	svm	0.761215
2	decision_tree	0.901907

Gambar 1. Hasil Model Selection

Terlihat bahwa model Random Forest adalah model terbaik dengan *best score* tertinggi, yaitu 93%. Ini menunjukkan bahwa model tersebut memiliki performa terbaik pada data *training* dan validasi selama *cross validation*.

Selanjutnya, dilakukan kembali *splitting* data setelah menemukan model dengan performa terbaik. Rasio pembagian data ditentukan berdasarkan nilai akurasi terbaik yang dihasilkan. Beberapa rasio train-test-split yang dicoba adalah 0.7, 0.75, 0.8, 0.85, dan 0.9. Masing-masing nilai akurasi hasil uji pada setiap rasio dataset termuat pada tabel berikut.

Tabel 2. Hasil Akurasi dari Tiap Rasio

Rasio Data Training	Rasio Data Testing	Nilai Akurasi
70%	30%	93,04%

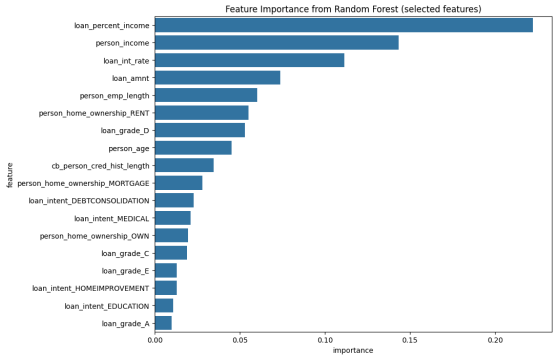
75%	25%	93,09%
80%	20%	93,18%
85%	15%	93,25%
90%	10%	93,33%

Lalu, kami mengidentifikasikan parameter terbaik dari rasio yang memiliki nilai akurasi tertinggi, yaitu 90:10. Didapat parameter terbaik dengan rincian sebagai berikut, di mana parameter *max_depth*, *min_samples_split*, dan *min_samples_leaf* yang diambil adalah *default* dari RandomForestClassifier.

Tabel 3. Hyperparameter Terbaik pada Model Random Forest

Hyperparameter	**Nilai**
n_estimators	500
max_depth	None
min_samples_split	2
min_samples_leaf	1

Kemudian, kami melakukan upaya untuk meningkatkan performa model dengan mengecek kepentingan fitur (*Feature Importance*). Dengan menggunakan *feature_importance_* yang ada pada scikit-learn, diperoleh urutan fitur dari fitur dengan kepentingan tertinggi sampai terendah. Kemudian, dibuat *threshold* sebesar 0.01, di mana fitur yang tingkat kepentingannya lebih tinggi dari *threshold* akan dipilih dalam *training* selanjutnya. Data *train* dan *test* kemudian akan diperbarui dengan hanya mencakup fitur-fitur yang terpilih. Urutan kepentingan fitur yang diperoleh adalah sebagai berikut.

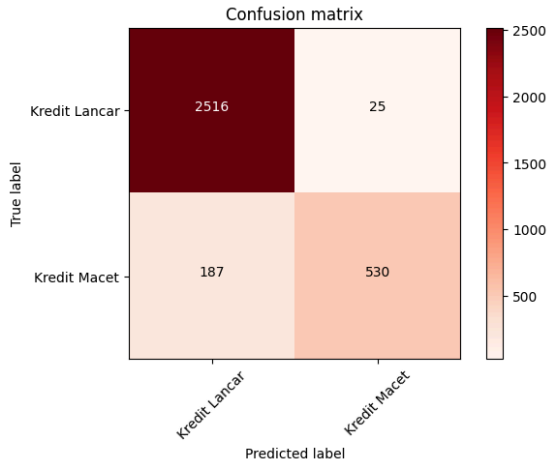


Gambar 2. Urutan Kepentingan Fitur

Setelah itu, kami juga menghitung interaksi fitur. Hal ini bertujuan untuk mengeksplorasi potensi tambahan dalam meningkatkan performa model. Dalam hal ini, interaksi fitur dilakukan dengan mengalikan dua fitur dengan tingkat kepentingan tertinggi, yaitu 'loan_percent_income' dan 'person_income' dan dibentuk fitur baru yang disebut 'interaction_loan_income'. Fitur ini dimasukkan ke dalam dataset yang telah dipilih untuk melatih kembali model Random Forest. Setelah pelatihan ulang, model tersebut digunakan untuk melakukan prediksi terhadap data *testing*.

2.4 Evaluasi Model

Setelah model kelayakan kredit menggunakan Random Forest telah berhasil dikembangkan, kami melakukan evaluasi untuk menilai sejauh mana model tersebut dapat memprediksi kelayakan peminjam dengan akurat. Evaluasi dilakukan dengan merujuk pada *confusion matrix* yang menggambarkan hasil prediksi model.



Gambar 3. Confusion Matrix

Dari *confusion matrix*, diperoleh hasil sebagai berikut:

True Positive (TP): Sebanyak 2516 kasus berhasil diprediksi sebagai “Kredit Lancar”.

False Positive (FP): Terdapat 25 kasus yang salah diprediksi sebagai “Kredit Lancar”.

False Negative (FN): Terdapat 187 kasus yang salah diprediksi sebagai “Kredit Macet”.True Negative (TN): Sebanyak 530 kasus berhasil diprediksi sebagai “Kredit Macet”.

Selain Confusion Matrix, diperoleh juga *Classification Report* yang memuat nilai akurasi, presisi, recall (sensitivitas), F1-score, dan lainnya.

$$precision = \frac{(TP)}{(TP+FP)}$$

$$recall = \frac{(TP)}{(TP+FN)}$$

$$f1\text{-}score = \frac{2 \times (recall \times precision)}{(recall + precision)}$$

$$accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

	precision	recall	f1-score	support
0	0.93	0.99	0.96	2541
1	0.95	0.74	0.83	717
accuracy			0.93	3258
macro avg	0.94	0.86	0.90	3258
weighted avg	0.94	0.93	0.93	3258

Gambar 4. Classification Report

Berdasarkan *Classification Report* di atas, model ini sangat efektif dalam mengklasifikasikan "kredit lancar" dengan *precision* dan *recall* yang sangat tinggi. Namun, untuk "kredit macet", meskipun *precision* cukup tinggi, *recall*-nya lebih rendah, menunjukkan bahwa model ini cenderung melewatkan beberapa kasus "kredit macet". Tetapi secara keseluruhan, kinerja model ini sangat baik dengan akurasi 93%.

3. KESIMPULAN

Berdasarkan hasil penelitian, diperoleh model klasifikasi Random Forest dengan tingkat performa tertinggi dibandingkan dengan model Decision Tree dan Support Vector Machine (SVM), dengan persentase performa sebesar 93%.

Evaluasi model dilakukan dengan Confusion Matrix, yang menunjukkan hasil prediksi model dengan jumlah True Positive (TP) sebesar 2516, False Positive (FP) sebesar 25, False Negative (FN) sebesar 187, dan True Negative (TN) sebesar 530. Hasil ini menunjukkan bahwa model Random Forest yang dikembangkan sangat efektif dalam mengklasifikasikan peminjam ke dalam kategori "kredit lancar" dan "kredit macet".

Dengan demikian, model Random Forest terbukti menjadi pilihan terbaik untuk memprediksi kelancaran kredit debitur dalam penelitian ini. Model yang telah dikembangkan akan membantu pihak kreditur dalam mengurangi risiko kredit macet serta memungkinkan pengambilan keputusan pemberian kredit yang lebih tepat sehingga mengurangi potensi kerugian finansial.

DAFTAR PUSTAKA

Aji, N. A., & Dhini, A. (2019). *Credit scoring through data mining approach: A case study of mortgage loan in Indonesia*. 2019 16th International Conference on Service Systems and Service Management, ICSSSM 2019, 1–5. https://doi.org/10.1109/ICSSSM.2019.8887731

Firmansah, N. and Indahyanti, U. (2023) *Prediction of Credit Eligibility Using the Random Forest Method*. doi:10.21070/ups.3515.

Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media : Canada.

LAMPIRAN

Link coding pada Google Colab:
Kelompok 2_UAS Sains Data.ipynb