Implementasi Algoritma *Random Forest* untuk Peningkatan Akurasi Credit Scoring

Kelompok 2 Sains Data

Anggota Kelompok:

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

Daftar Isi

01

Pendahuluan

Latar Belakang

Analisis kredit adalah hal yang penting untuk dilakukan sebelum penyetujuan pengajuan kredit debitur, karena sifat dari kredit yang mempunyai risiko, dimana risiko tersebut dipengaruhi oleh latar belakang debitur. Maka, sistem credit scoring sangat diperlukan dalam memutuskan pemberian kredit untuk menghindari kredit macet yang dapat menyebabkan kerugian bagi pihak kreditur.

Tujuan

Penelitian ini bertujuan untuk meningkatkan akurasi metode klasifikasi dengan tingkat akurasi yang paling tinggi dan memberikan rekomendasi kepada pihak kreditur. Sehingga, pihak kreditur dapat menganalisis kredit sebelum memutuskan penyetujuan kredit.

02

Dataset dan EDA

Tentang Dataset

Sumber

Diambil dari kaggle: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Jumlah data

Data terdiri dari 32581 sampel dengan 12 fitur

Informasi pada data

Dataset berisi beberapa informasi penting serta karakteristik pemohon kredit seperti usia, pendapatan, status kepemilikan rumah, jumlah pinjaman, suku bunga pinjaman, dan lain-lain.

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1	0.59	Y	3
1	21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0	0.10	N	2
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1	0.57	N	3
3	23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1	0.53	N	2
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1	0.55	Y	4

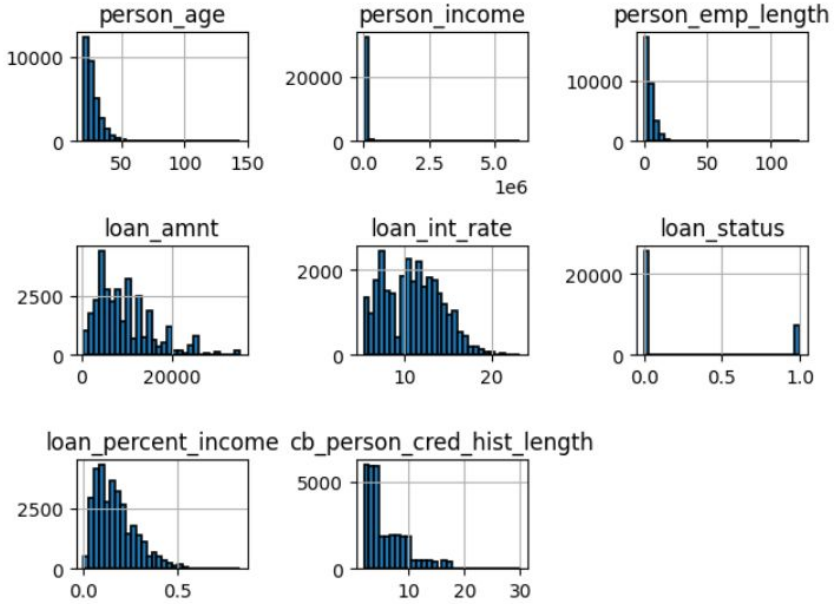
Penjelasan Fitur

person_age	Usia individu yang mengajukan kredit.
person_income	Penghasilan tahunan individu.
person_home_ownership	Jenis kepemilikan rumah individu (rent, mortgage, own, other).
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 (Grade A-G).

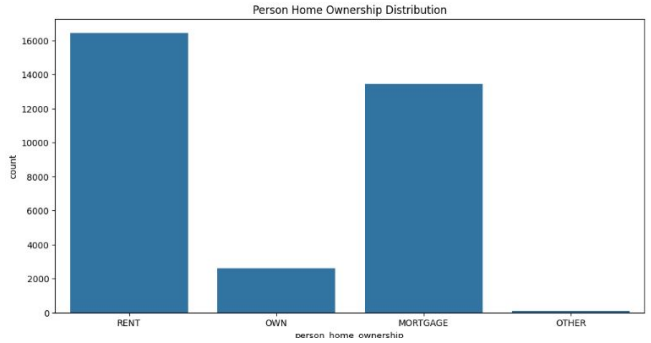
Penjelasan Fitur

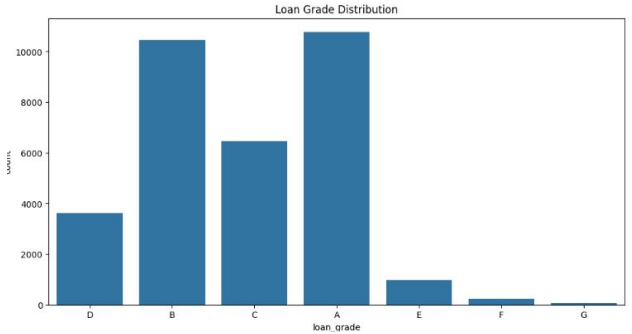
loan_amnt	Jumlah kredit/pinjaman yang diminta oleh individu.
loan_int_rate	Suku bunga yang terkait dengan kredit.
loan_status	Status kredit, dimana 0 menandakan tidak gagal bayar dan 1 menandakan gagal bayar.
loan_percent_income	Persentase pendapatan yang diwakili oleh jumlah pinjaman.
cb_person_default_on_file	Riwayat gagal bayar individu sesuai catatan biro kredit (Y/N)
cb_person_cred_hist_length	Panjang riwayat kredit untuk individu tersebut.

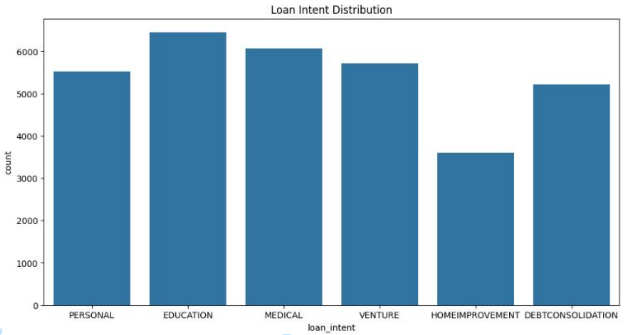
EDA : Distribusi Fitur Numerik

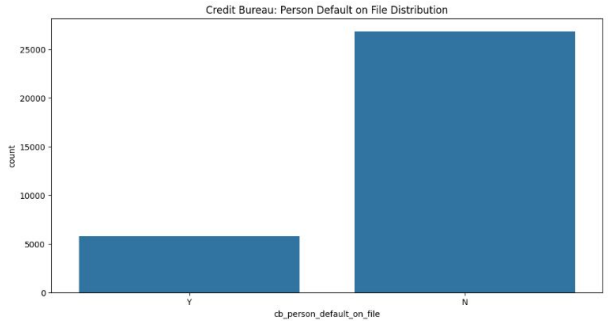


EDA : Distribusi Fitur Kategorik

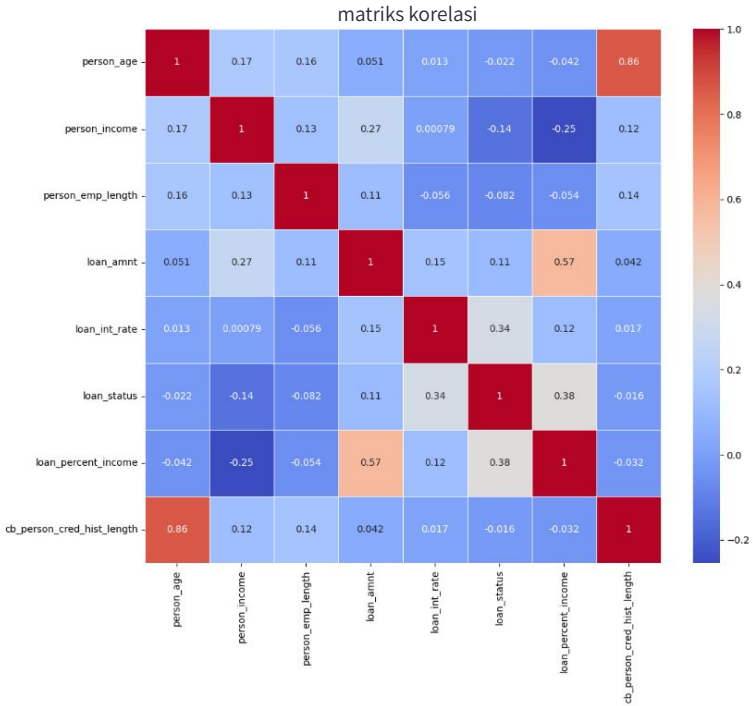








EDA : Heatmap



Berdasarkan heatmap, dapat dilihat bahwa nilai korelasi paling tinggi dihasilkan oleh hubungan antara fitur person_age dan cb_person_cred_hist_length dengan nilai 0.86

03

Preprocessing Data

Menangani Missing Values

Pemeriksaan Missing Values

person_age	0
person_income	0
person_home_ownership	0
person_emp_length	895
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	3116
loan_status	0
loan_percent_income	0
cb_person_default_on_file	0
cb_person_cred_hist_length	0

Dilakukan pemeriksaan jumlah missing values yang terdapat pada tiap-tiap fitur.

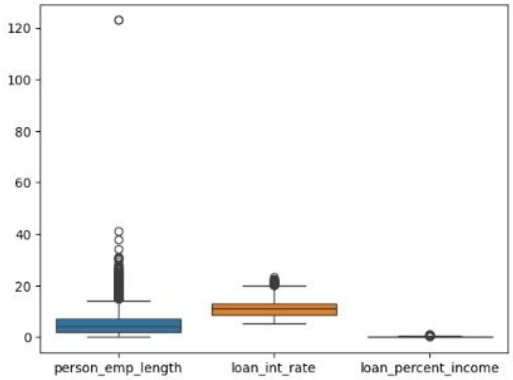
Menangani missing values dengan strategi mean

person_age	0
person_income	0
person_home_ownership	0
person_emp_length	0
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	0
loan_status	0
loan_percent_income	0
cb_person_default_on_file	0
cb_person_cred_hist_length	0

Missing values diisi dengan menggunakan strategi mean kemudian dicek kembali.

Menangani Outliers

Pemeriksaan Outliers



Dilakukan pemeriksaan outliers yang terdapat pada fitur-fitur dengan tipe data ‘float64’ dengan visualisasi boxplot.

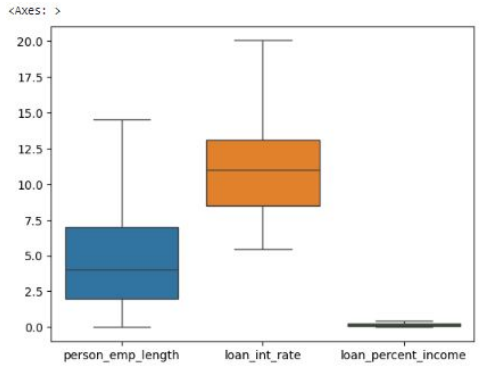
Observasi Outliers

Observasi:

* person_age: Sebagian besar individu berusia 20 hingga 60 tahun. Sehingga, agar lebih umum, individu dengan usia > 80 tahun akan dihapus.
* person_emp_length: Sebagian besar individu memiliki pengalaman kerja kurang dari 40 tahun. Sehingga, individu dengan pengalaman kerja > 60 tahun akan dihapus.

Menangani Outliers

Menangani outliers dengan metode Inter Quartile Range (IQR)



Setelah dilakukan metode IQR dan penghapusan sampel berdasarkan observasi, keberadaan outliers diperiksa kembali menggunakan boxplot.

Encoding Fitur Kategorik

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32572 entries, 0 to 32571
Data columns (total 27 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   person_age                      32572 non-null  int64
 1   person_income                   32572 non-null  int64
 2   person_emp_length               32572 non-null  float64
 3   loan_amnt                       32572 non-null  int64
 4   loan_int_rate                   32572 non-null  float64
 5   loan_status                     32572 non-null  int64
 6   loan_percent_income             32572 non-null  float64
 7   cb_person_cred_hist_length      32572 non-null  int64
 8   person_home_ownership_MORTGAGE  32572 non-null  float64
 9   person_home_ownership_OTHER     32572 non-null  float64
 10  person_home_ownership_OWN       32572 non-null  float64
 11  person_home_ownership_RENT      32572 non-null  float64
 12  loan_intent_DEBTCONSOLIDATION   32572 non-null  float64
 13  loan_intent_EDUCATION           32572 non-null  float64
 14  loan_intent_HOMEIMPROVEMENT     32572 non-null  float64
 15  loan_intent_MEDICAL             32572 non-null  float64
 16  loan_intent_PERSONAL            32572 non-null  float64
 17  loan_intent_VENTURE             32572 non-null  float64
 18  loan_grade_A                    32572 non-null  float64
 19  loan_grade_B                    32572 non-null  float64
 20  loan_grade_C                    32572 non-null  float64
 21  loan_grade_D                    32572 non-null  float64
 22  loan_grade_E                    32572 non-null  float64
 23  loan_grade_F                    32572 non-null  float64
 24  loan_grade_G                    32572 non-null  float64
 25  cb_person_default_on_file_N     32572 non-null  float64
 26  cb_person_default_on_file_Y     32572 non-null  float64
dtypes: float64(22), int64(5)
memory usage: 6.7 MB
```

Fitur-fitur kategorik yang terdapat di dalam dataset adalah person_home_ownership, loan_intent, loan_grade, dan cb_person_default_on_file. Dengan metode one-hot encoding, fitur-fitur tersebut berhasil di-encoding, menambahkan jumlah fitur sehingga berjumlah 27 fitur. Setelah itu, dipastikan semua data sudah bersifat numerik dan tidak terdapat missing values.

04

Modelling

Train-Test-Split

```
[23] # menghapus kolom loan_status dan menyimpan data dalam X
     X = df_preprocessed.drop(['loan_status'], axis=1)

     # memilih kolom loan_status dan menyimpannya dalam Y
     y = df_preprocessed['loan_status']
```

```
[24] # membagi data menjadi dua bagian, yaitu 80% data training dan 20% data testing
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[64] print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

     (29314, 26) (3258, 26) (29314,) (3258,)
```

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. Sebelum membentuk model, diperlukan 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%.

Model Selection dan Hyperparameter Tuning

Sebelum memutuskan untuk menggunakan metode Random Forest, kami melakukan Model Selection untuk melihat metode mana yang memiliki best_score tertinggi dari beberapa metode, yaitu SVM, Decision Tree, dan Random Forest.

	model	best_score
0	random_forest	0.931841
1	svm	0.376582
2	decision_tree	0.901992

Terlihat bahwa metode Random Forest adalah model terbaik dengan best score tertinggi, yaitu 93,18%. Ini menunjukkan bahwa model tersebut memiliki performa terbaik pada data training dan validasi selama cross validation.

Perbandingan Rasio

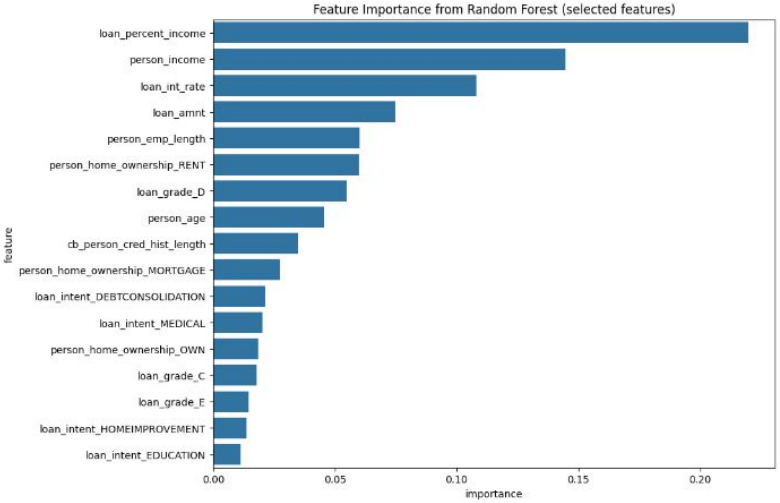
Selanjutnya, dilakukan kembali splitting data setelah menemukan metode yang paling efektif. Rasio ditentukan berdasarkan nilai akurasi yang dihasilkan. Masing-masing nilai akurasi hasil uji tiap rasio dataset termuat pada tabel berikut.

Rasio Data Testing	Rasio Data Training	Nilai Akurasi
70%	30%	93,04%
75%	25%	93,09%
80%	20%	93,18%
85%	15%	93,25%
90%	10%	93,33%

Didapatkan rasio terbaik adalah 90:10. Lalu, diidentifikasikan parameter terbaik dari rasio tersebut.

Hyperparameter	Nilai
n_estimators	100
max_depth	None
min_samples_split	2
min_samples_leaf	1

Feature Importance



Untuk meningkatkan performa model dilakukan pengecekan kepentingan fitur (Feature Importance). Dengan menggunakan feature_importance_ yang ada pada python, 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 dilampirkan pada gambar diatas.

Feature Interaction

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.

```
# interaksi fitur
X_train_selected['interaction_loan_income'] = X_train_selected['loan_percent_income'] * X_train_selected['person_income']
X_test_selected['interaction_loan_income'] = X_test_selected['loan_percent_income'] * X_test_selected['person_income']
```

```
# pelatihan ulang model
rf_selected = RandomForestClassifier(n_estimators=100, random_state=42)
rf_selected.fit(X_train_selected, y_train)
```

```
▾ RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
# melakukan prediksi menggunakan model yang telah diperbarui
y_pred_selected = rf_selected.predict(X_test_selected)
```

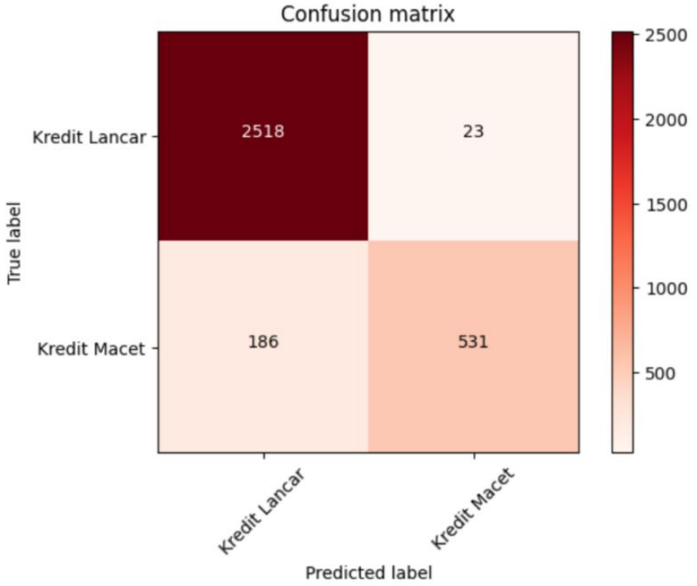
05

Evaluasi

Confusion Matrix

Dari Confusion Matrix, diperoleh hasil sebagai berikut:

- True Positive (TP): Sebanyak 2518 kasus berhasil diprediksi sebagai “Kredit Lancar”.
- False Positive (FP): Terdapat 23 kasus yang salah diprediksi sebagai “Kredit Lancar”.
- False Negative (FN): Terdapat 186 kasus yang salah diprediksi sebagai “Kredit Macet”.
- True Negative (TN): Sebanyak 531 kasus berhasil diprediksi sebagai “Kredit Macet”.



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 94%.

	precision	recall	f1-score	support
0	0.93	0.99	0.96	2541
1	0.96	0.74	0.84	717
accuracy			0.94	3258
macro avg	0.94	0.87	0.90	3258
weighted avg	0.94	0.94	0.93	3258

Kesimpulan

Penelitian ini bertujuan untuk memprediksi kelancaran kredit debitur menggunakan metode klasifikasi yang paling akurat, serta memberikan rekomendasi kepada pihak kreditur untuk menganalisis kredit sebelum memutuskan penyetujuan kredit. Dalam penelitian ini, tiga metode klasifikasi dibandingkan: Decision Tree, Support Vector Machine (SVM), dan Random Forest. Hasil menunjukkan bahwa Random Forest memiliki tingkat akurasi tertinggi, yaitu 93,33%, dengan rasio data training dan testing 90:10. Evaluasi model dilakukan dengan Confusion Matrix dan Classification Report. Hasilnya menunjukkan bahwa model Random Forest yang dikembangkan sangat efektif dalam mengklasifikasikan peminjam ke dalam kategori "kredit lancar" dan "kredit macet". Dengan demikian, metode Random Forest terbukti sebagai metode terbaik untuk memprediksi kelancaran kredit debitur dalam penelitian ini. Model yang dikembangkan dapat membantu pihak kreditur dalam mengurangi risiko kredit macet dan membuat keputusan pemberian kredit yang lebih tepat, sehingga mengurangi potensi kerugian finansial.

Thank you

(for listening). have a nice day!