

MAKALAH
PRAKTIKUM KOMPUTASI STATISTIKA II

Optimalisasi Deteksi Penipuan Kartu Kredit Menggunakan
Metode Klasifikasi *Random Forest* di Bank XYZ



Nama Mahasiswa (Nomor Induk Mahasiswa):

| | |
|--------------------------------|----------------------|
| ALLISYA MAHARANI ADINDA WIBOWO | (21/478078/PA/20729) |
| BRYAN FLORENTINO LEO | (21/473767/PA/20429) |
| NATASYA FATIMAH SALIM | (21/477164/PA/20634) |

LABORATORIUM KOMPUTASI MATEMATIKA DAN STATISTIKA
DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA

2023

DAFTAR ISI

| | |
|---------------------------------------|-----|
| DAFTAR ISI | i |
| ABSTRAK | 1 |
| LATAR BELAKANG | 2 |
| TUJUAN DAN MANFAAT | 4 |
| Tujuan..... | 4 |
| Manfaat | 4 |
| METODE ANALISIS | 5 |
| ANALISIS..... | 10 |
| Eksplorasi dan Visualisasi Data | 10 |
| Praproses Data..... | 13 |
| Hasil dan Pembahasan..... | 13 |
| Penerapan Model..... | 16 |
| KESIMPULAN | 17 |
| DAFTAR PUSTAKA | ii |
| LAMPIRAN | iii |

ABSTRAK

Kartu kredit merupakan salah satu opsi alat pembayaran yang penggunaannya terus bertumbuh pada masa kini. Populernya kartu kredit sebagai salah satu alat pembayaran digital tidak terlepas dari penyalahgunaan yang dikenal dengan sebutan *fraud* kartu kredit atau *carding*. Perlindungan berbasis teknologi merupakan salah satu bidang fokus untuk mencegah kasus ini. Sistem layanan dan perlindungan *fraud* yang proaktif perlu dikembangkan, salah satunya dengan pendekatan analitika data. Dalam penelitian ini, penulis membangun sebuah model pengklasifikasi *random forest* yang akurat dan tangkas untuk memprediksi kejadian *fraud* dari jutaan baris data transaksi dari Bank XYZ. Peneliti memperoleh bahwa model pengklasifikasi *random forest* yang didahului dengan praproses data sedemikian rupa menghasilkan akurasi dan presisi di atas 98 persen, sehingga model *random forest* sangat baik dan sesuai.

LATAR BELAKANG

Kartu kredit merupakan salah satu opsi alat pembayaran masa kini. Ide penemuan kartu kredit bermula dari sistem pembayaran kredit di Amerika Serikat pada awal tahun 1900-an. Beberapa dekade kemudian, berkat kerja sama bank-bank konvensional dengan VISA dan Mastercard International, kartu kredit mulai hadir di Indonesia.¹

BCG mencatat bahwa penetrasi kartu kredit di Indonesia tergolong rendah di region Asia Tenggara dengan tingkat hanya enam persen. Jumlah kartu kredit yang beredar di Indonesia juga menurun sebesar tiga persen pascakrisis pandemi COVID-19 pada tahun 2020.² Akan tetapi, tren penggunaan kartu kredit terus bangkit. Menurut Statistik Sistem Pembayaran dan Infrastruktur Pasar Keuangan (SPIP), hingga bulan Oktober 2022, volume dan nilai transaksi kartu kredit di Indonesia berturut-turut telah bertumbuh sebanyak 23 dan 33 persen secara *year-on-year*.³

Populernya kartu kredit sebagai salah satu alat pembayaran digital tidak terlepas dari penyalahgunaan. Menurut Otoritas Jasa Keuangan (OJK), *fraud* kartu kredit, atau yang juga dikenal dengan *carding*, ialah kasus kejahatan dengan mencuri atau menyalahgunakan identitas pemilik kartu kredit yang telah memasuki tren setidaknya sejak awal abad ke-21. Asosiasi Kartu Kredit Indonesia (AKKI) melaporkan bahwa selama periode Juli 2003 hingga April 2006, terjadi 89 kasus *fraud* kartu kredit dengan kerugian mencapai 4,6 juta USD. Nominal ini setara dengan kerugian akibat *fraud* kartu kredit pada tahun 2007 tersendiri. (Prabowo, 2012) Survei *fraud* global oleh ACI Worldwide pada tahun 2016 menyatakan Indonesia sebagai negara pada urutan ke-14 dunia dengan tingkat *fraud* kartu kredit nasional tertinggi, yakni sebesar 26%.⁴ Terkini, *Report to the Nations 2022* yang dirilis oleh *Association of Certified Fraud Examiners* (ACFE) menempatkan Indonesia pada urutan keempat sebagai negara di region Asia-Pasifik dengan kasus *fraud* tahunan terbanyak. Dalam berkas yang serupa dua tahun sebelumnya, Indonesia bahkan menempati urutan pertama melalui sumbangan 36 dari 198 kasus *fraud* regional terlapor.⁵

Prabowo (2012) menyatakan bahwa perlindungan berbasis teknologi merupakan salah satu bidang fokus untuk mencegah *fraud* kartu kredit. Berkenaan dengan aspek tersebut, Bank Indonesia telah mewajibkan institusi bank maupun

¹ Nofianti, M. (2023, 6 Jan). *Menilik Sejarah Kartu Kredit di Indonesia*. cekaja.com. Diakses pada 15 Juni 2023, dari <https://www.cekaja.com/info/menilik-sejarah-kartu-kredit-di-indonesia>.

² de Sartiges, D. dkk. (2020, 20 Mei). *Southeast Asian Consumers Are Driving a Digital Payment Revolution*. BCG. Diakses pada 15 Juni 2023, dari <https://www.bcg.com/publications/2020/southeast-asian-consumers-digital-payment-revolutions>.

³ Damara, D. (2022, 29 Des). *Mantap! Volume dan Nilai Transaksi Kartu Kredit Tumbuh 2 Digit*. Bisnis.com. Diakses pada 15 Juni 2023, dari <https://finansial.bisnis.com/read/20221229/90/1613144/mantap-volume-dan-nilai-transaksi-kartu-kredit-tumbuh-2-digit>

⁴ Subeditor Business Plus. (2016, 13 Jul). *What Are Worst Countries For Credit Card Fraud?* Diakses pada 15 Juni 2023, dari <https://businessplus.ie/news/what-are-worst-countries-for-credit-card-fraud/>.

⁵ Triatmodjo, Y. (2021, 21 Sep). *Pencegahan Fraud di Indonesia*. Diakses pada 15 Juni 2023, dari <https://insight.kontan.co.id/news/pencegahan-fraud-di-indonesia>.

nonbank untuk mempunyai sistem layanan dan perlindungan *fraud* yang ketat. Survei Risiko *Fraud* oleh Kroll dan ACFE Indonesia mendapati bahwa 62 persen kejadian *fraud* di Indonesia dideteksi melalui sistem pelaporan *whistleblowing* (WBS).⁶ Akan tetapi, penerapan sistem ini merupakan suatu tantangan di tengah budaya timur masyarakat Indonesia yang menuntut terjaganya harmoni. Oleh karena itu, sistem layanan dan perlindungan *fraud* yang proaktif perlu dikembangkan, salah satunya dengan pendekatan analitika data.

Ide penerapan analitika data guna perlindungan *fraud* yang dicetuskan oleh Kroll bermula dari penentuan perusahaan akan parameter-parameter risiko *fraud* yang bersifat prioritas. Selanjutnya, analitika data, khususnya *machine learning*, akan menyaring transaksi-transaksi yang memerlukan investigasi lanjutan seputar *fraud*. Salah satu algoritma yang telah banyak digunakan untuk tujuan ini ialah klasifikasi *random forest*. Kurniawan dan Yulianingsih (2021) menyebut *random forest* disebut sebagai solusi terbaik dan akurat untuk memprediksi kejadian *fraud*. Atas dasar ini, penulis tertarik untuk membangun sebuah model pengklasifikasi *random forest* yang akurat dan tangkas untuk memprediksi kejadian *fraud* dari jutaan baris data.

⁶ Tama, D. R. (2022, 10 Agt). *Unreported Fraud: A Risky Blindspot for Indonesia*. Diakses pada 16 Juni 2023, dari <https://www.kroll.com/en/insights/publications/unreported-fraud-a-risky-blindspot-for-indonesia>.

TUJUAN DAN MANFAAT

Tujuan

Melakukan analisis data untuk membantu Bank XYZ dalam memprediksi dan mendeteksi *fraud* kartu kredit pada pelanggan menggunakan *machine learning*.

Manfaat

1. Bank XYZ dapat meningkatkan kemampuan mereka untuk mendeteksi dan memprediksi penipuan kartu kredit dengan lebih akurat sehingga dapat melindungi pelanggan dari penipuan.
2. Bank XYZ dapat mengurangi kerugian finansial yang mungkin disebabkan oleh penipuan dikarenakan bisa mendeteksi suatu penipuan lebih awal.
3. Bank XYZ dapat meningkatkan kepuasan dan kepercayaan pelanggan dikarenakan memiliki sistem deteksi penipuan yang lebih akurat.

METODE ANALISIS

Dimiliki suatu *dataset* yang terdiri dari 33 variabel. Dari *dataset* tersebut, ingin dilakukan analisis data untuk membantu Bank XYZ dalam memprediksi dan mendeteksi penipuan kartu kredit pada pelanggan menggunakan *machine learning*. Penjelasan mengenai variabel-variabel pada *dataset* tersebut adalah sebagai berikut.

| Variabel | Deskripsi |
|------------------------------|--|
| address_months_count | Jumlah bulan di alamat terdaftar pemohon sebelumnya, yaitu tempat tinggal pemohon sebelumnya, jika ada. (-1 adalah missing value). |
| age | Usia pelamar per dekade |
| app_24h | total permohonan dalam 24 jam terakhir |
| app_4w | total permohonan dalam 4 minggu terakhir |
| app_6h | total permohonan dalam 6 jam terakhir |
| bank_months | Berapa umur akun sebelumnya dalam bulan |
| credit_limit | batas kredit yang diusulkan dari pemohon |
| credit_score | risiko nilai kredit |
| current_address_months_count | Bulan di alamat pemohon yang terdaftar saat ini (-1 adalah missing value). |
| days_request | Jumlah hari sejak permohonan selesai |
| device_fraud | Jumlah aplikasi penipuan dengan <i>device</i> |
| distinct_birth_emails | Jumlah email untuk pelamar dengan tanggal lahir yang sama dalam 4 minggu terakhir |
| distinct_device_emails | Jumlah email berbeda di situs web perbankan |
| email_similarity | kesamaan antara email dan nama pelamar |
| email_status | status email (1 berbayar, 0 gratis) |
| employment | Status Pekerjaan |
| foreign | status asal permintaan permohonan (0 domestik) |
| fraud | Label penipuan (1 jika penipuan, 0 jika tidak) |
| housing | status perumahan bagi pemohon |
| id | id unique |
| income | Pendapatan tahunan pemohon (skala) |
| initial_amount | Jumlah transfer awal untuk permohonan |
| keep_alive | Opsi pengguna pada sesi logout. |
| mobile_status | ponsel yang disediakan (1 disediakan) |
| month_of_application | Bulan dimana permohonan dibuat |
| os | Sistem operasi perangkat |
| other_cards | status kepemilikan kartu lain (0 tidak punya) |
| payment | Jenis paket pembayaran kredit |
| phone_status | disediakan telepon rumah (1 disediakan) |
| session_length | Durasi sesi pengguna di situs web perbankan dalam hitungan menit |

| | |
|-----------------------|---|
| source_of_application | sumber permohonan |
| total_app_8w | total permohonan dalam 8 minggu terakhir |
| zip_count | Jumlah permohonan dengan kode pos yang sama dalam 4 minggu terakhir |

Tahapan yang dilakukan hingga memperoleh hasil prediksi adalah sebagai berikut.

1. Eksplorasi dan Visualisasi Data

Sebelum melakukan *modelling* dengan *random forest classifier*, eksplorasi dan visualisasi data memiliki beberapa fungsi penting, antara lain:

1) Memahami karakteristik data

Eksplorasi dan visualisasi data membantu pemahaman akan karakteristik data, seperti distribusi, korelasi, dan pencilaan. Hal ini dapat membantu pemilihan fitur yang tepat untuk dimasukkan ke dalam model dan memastikan bahwa data yang digunakan sudah bersih dan siap untuk diproses.

2) Menentukan fitur yang relevan

Dengan visualisasi data, dapat ditentukan fitur-fitur yang paling relevan untuk dimasukkan ke dalam model. Hal ini dapat membantu meningkatkan akurasi model dan mengurangi *overfitting*.

3) Meningkatkan interpretasi model

Visualisasi data dapat membantu memahami bagaimana model bekerja dan mengapa model mengeluarkan hasil tertentu. Hal ini dapat membantu meningkatkan interpretasi model dan memastikan bahwa model yang dihasilkan dapat dipahami oleh khalayak umum yang tidak memiliki latar belakang teknis.

2. Praproses Data

Praproses data adalah proses mengubah data mentah menjadi bentuk yang lebih mudah dipahami dan diproses oleh mesin. Praproses data sangat penting karena kualitas data mempengaruhi keberhasilan dari suatu analisis. Dengan melakukan praproses data, data lebih mudah dibaca dan berformat tertentu, sehingga dapat mengeluarkan hasil yang lebih akurat dan optimal. Tahapan praproses data yang akan dilakukan adalah sebagai berikut.

1) Menghilangkan data duplikat

Data duplikat harus dihilangkan karena dapat memengaruhi kualitas dan akurasi hasil analisis data. Data duplikat dapat menghasilkan informasi yang tidak akurat dan analisis menjadi tidak efisien.

2) Menghilangkan atribut yang tidak diperlukan

Untuk menghilangkan atribut yang tidak diperlukan, dilakukan seleksi fitur. Seleksi fitur adalah proses memilih subhimpunan fitur yang paling relevan dan signifikan untuk dimasukkan ke dalam model. Atribut yang tidak diperlukan atau terlalu banyak mengandung *missing value* harus dihilangkan karena juga dapat membuat analisis menjadi tidak akurat dan tidak efisien.

3) **Memperbaiki *missing values***

Memperbaiki *missing values* sangat penting untuk dilakukan karena data yang tidak lengkap dapat mempengaruhi kualitas dan akurasi analisis data. Lalu, terdapat beberapa algoritma yang tidak mengizinkan kumpulan data dengan *missing value*. Cara yang dapat dilakukan untuk memperbaiki *missing values* adalah dengan mengisi nilai-nilai hilang menggunakan mean dan median untuk data numerik serta modus untuk data kategorik.

4) **Penskalaan data**

Penskalaan data atau penskalaan fitur adalah proses menyeragamkan skala fitur-fitur dalam data. Hal ini penting dilakukan karena beberapa algoritma sangat sensitif terhadap skala data, sehingga penskalaan perlu dilakukan agar hasilnya lebih akurat.

5) ***One Hot Encoding***

One-hot encoding adalah proses mengubah variabel kategorik menjadi nilai numerik biner. Proses ini dilakukan untuk menghindari bias yang mungkin terjadi pada variabel kategorik dan memungkinkan algoritma untuk memproses data kategorik.

3. **Modelling dengan *Random Forest Classifier***

Random Forest Classifier adalah algoritma *machine learning* yang digunakan untuk klasifikasi data dalam jumlah besar. Algoritma ini merupakan kombinasi dari beberapa pohon keputusan (*decision tree*) yang dikombinasikan menjadi satu model. *Random Forest Classifier* dapat diterapkan pada kasus klasifikasi maupun regresi dan menerapkan teknik ansambel untuk menggabungkan banyak penggolong (*classifiers*) guna memberikan solusi terhadap masalah yang kompleks. Berikut adalah langkah-langkah untuk melakukan *modelling* dengan *Random Forest Classifier*:

1) **Mempersiapkan data**

Melakukan praproses data, seperti menghilangkan data duplikat, menghilangkan atribut yang tidak diperlukan, dan memperbaiki *missing value*.

2) **Menskalakan data**

Melakukan penskalaan data agar algoritma dapat menghasilkan hasil yang akurat dan konsisten.

3) **Membuat model**

Membuat model *Random Forest Classifier* menggunakan *library* seperti *Scikit-learn* dalam bahasa Python.

4) **Melatih model**

Melatih model dengan data *train*.

5) **Mengevaluasi model**

Mengevaluasi model dengan menggunakan data *test* dan metrik-metrik evaluasi, seperti akurasi, presisi, *recall*, dan *F1-score*.

4. Evaluasi model

Evaluasi model sangat penting dilakukan dalam *modeling* dengan *Random Forest Classifier* karena dapat memberikan informasi tentang seberapa baik model yang dibuat dapat memprediksi data yang belum pernah dilihat sebelumnya. Evaluasi model juga dapat membantu pemilihan parameter yang tepat untuk model dan memastikan bahwa model yang dihasilkan dapat digunakan untuk memprediksi data yang belum pernah dilihat sebelumnya dengan akurasi yang tinggi. Metrik-metrik evaluasi, seperti akurasi, presisi, *recall*, dan *F1-score* digunakan untuk mengevaluasi kinerja model.

a) *Confusion matrix*

Confusion matrix terdiri dari empat bagian, yaitu *true positive* (TP), *false positive* (FP), *true negative* (TN), dan *false negative* (FN). *True positive* adalah jumlah data yang diklasifikasikan dengan benar sebagai positif, *false positive* adalah jumlah data yang diklasifikasikan dengan salah sebagai positif, *true negative* adalah jumlah data yang diklasifikasikan dengan benar sebagai negatif, dan *false negative* adalah jumlah data yang diklasifikasikan dengan salah sebagai negatif.

b) Akurasi, *Precision*, dan *Recall*

Akurasi, *precision*, dan *recall* adalah metrik-metrik evaluasi yang digunakan untuk mengukur kinerja model *machine learning*, khususnya pada tugas klasifikasi.

- **Akurasi**

Akurasi adalah rasio antara jumlah prediksi benar dengan jumlah total data. Metrik ini mengukur seberapa baik model dapat memprediksi kelas target secara keseluruhan. Namun, akurasi tidak cocok digunakan pada dataset yang tidak seimbang (jumlah data pada kelas target yang tidak seimbang). Rumus akurasi yaitu sebagai berikut.

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%$$

- ***Precision***

Precision adalah rasio antara jumlah prediksi benar positif dengan jumlah total prediksi positif. Metrik ini mengukur seberapa baik model dapat memprediksi kelas positif secara akurat. *Precision* cocok digunakan pada dataset yang memiliki jumlah data pada kelas target yang tidak seimbang. Rumus *precision* yaitu sebagai berikut.

$$\text{Precision} = \frac{TP}{TP+FP}$$

- ***Recall***

Recall adalah rasio antara jumlah prediksi benar positif dengan jumlah total data positif. Metrik ini mengukur seberapa baik model dapat mengidentifikasi semua data positif. *Recall* cocok digunakan pada dataset yang memiliki jumlah data pada kelas target yang tidak seimbang. Rumus *recall* yaitu sebagai berikut.

$$\text{Recall} = \frac{TP}{TP+FN}$$

c) *F1-Score*

F1-score adalah metrik evaluasi yang menggabungkan *precision* dan *recall* menjadi satu nilai tunggal. *F1-score* digunakan untuk mengukur kinerja model klasifikasi dengan mempertimbangkan kedua metrik tersebut secara seimbang. *F1-score* dihitung dengan menggunakan *harmonic mean* dari *precision* dan *recall*. *Harmonic mean* digunakan karena lebih cocok untuk menghitung rata-rata dari dua nilai yang berbeda, seperti *precision* dan *recall*. *F1-score* memiliki rentang nilai antara 0 dan 1, di mana nilai 1 menunjukkan kinerja model yang sempurna. Rumus *F1-score* yaitu sebagai berikut.

$$F1-Score = \frac{2*Precision*Recall}{Precision+Recall}$$

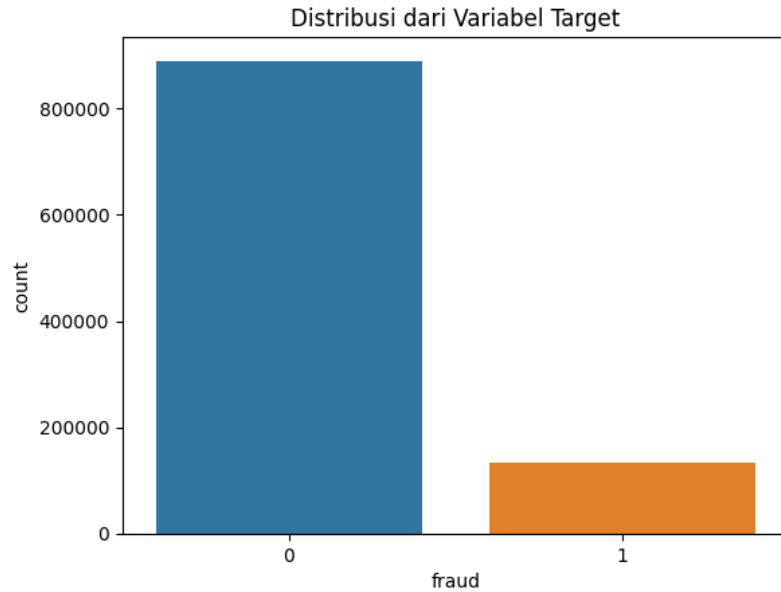
d) *Receiver Operator Characteristic (ROC)*

Receiver Operating Characteristic (ROC) adalah kurva yang digunakan untuk mengevaluasi kinerja model klasifikasi pada berbagai *threshold* atau batas keputusan. ROC menggambarkan hubungan antara *true positive rate (TPR)* dan *false positive rate (FPR)* pada berbagai nilai *threshold*. TPR adalah rasio antara jumlah prediksi benar positif dengan jumlah total data positif, sedangkan FPR adalah rasio antara jumlah prediksi salah positif dengan jumlah total data negatif. ROC digunakan untuk memilih *threshold* yang optimal untuk model klasifikasi, yaitu *threshold* yang memberikan keseimbangan antara TPR dan FPR yang optimal.

ANALISIS

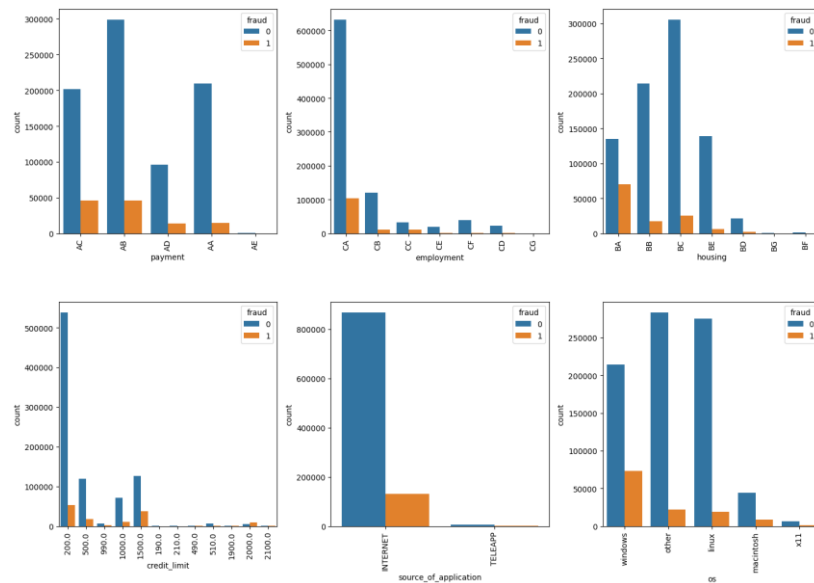
Eksplorasi dan Visualisasi Data

1. Distribusi variabel target

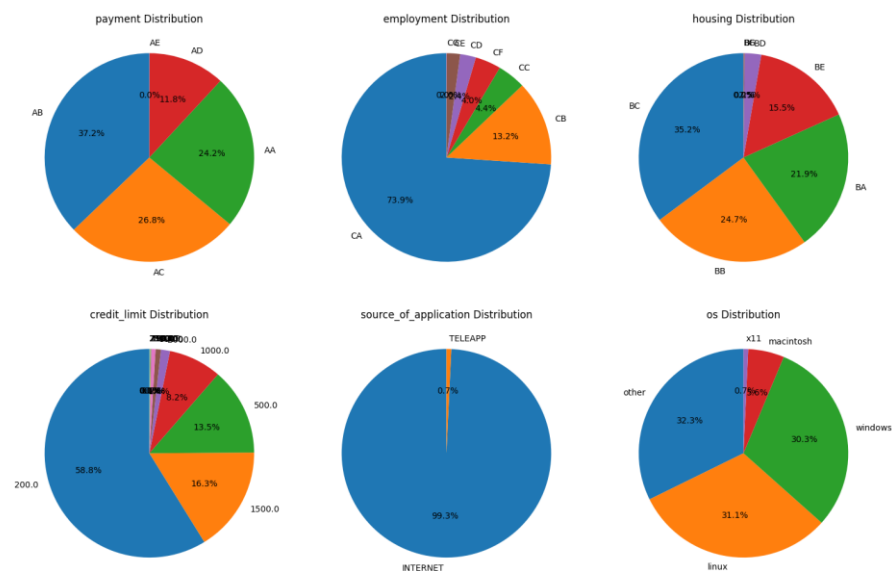


Dari data *train* yang telah diberikan, terlihat dari diagram batang di atas bahwa nilai variabel *fraud* yang dilabelkan dengan '1' memiliki jumlah yang jauh signifikan lebih sedikit daripada data *not fraud* yang dilabelkan dengan angka '0'. Oleh karena itu, disimpulkan bahwa data bersifat *imbalance* atau tidak rata.

2. Distribusi variabel prediktor kategorik



Dari visualisasi grafik di atas, dapat dijelaskan banyaknya jumlah masing-masing nilai pada setiap jenis data dan dihubungkan dengan output dari variabel target yang ingin dicari. Secara sekilas, terlihat bahwa terdapat beberapa nilai pada variabel prediktor tertentu yang memiliki pengaruh yang cukup besar terhadap status dari variabel prediktor. Dari variabel *payment*, dapat dilihat bahwa *payment* menggunakan metode AB mempunyai kesempatan besar untuk diidentifikasi sebagai penipuan. Pada variabel *employment*, dicurigai nilai CA. Pada variabel *housing*, dicurigai nilai BC. Dicurigai juga variabel *credit_limit* dengan nilai 200.0, *source of application* dengan nilai “internet”, dan *os* dengan pilihan “other”.



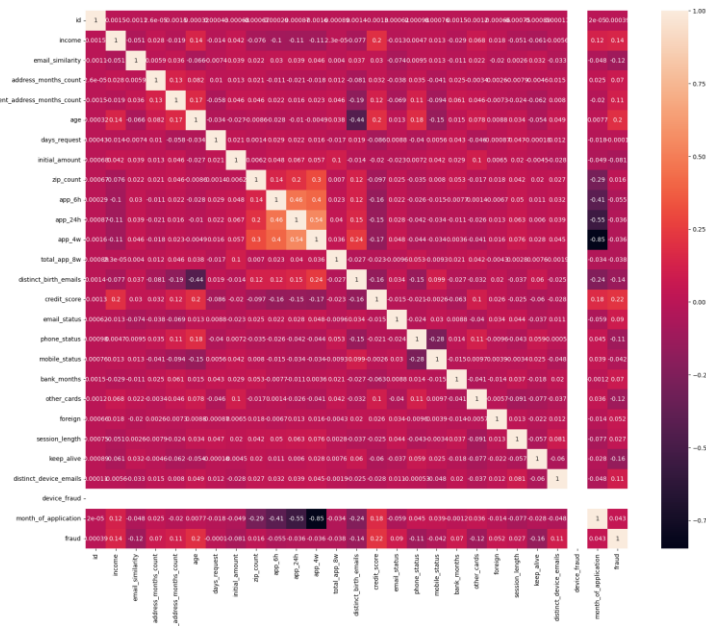
Melalui diagram lingkaran di atas, terlihat bahwa terdapat beberapa kategori data yang mendominasi pada setiap variabel prediktor. Pada variabel *payment*, data yang mendominasi adalah AB. Pada variabel *employment*, data yang mendominasi adalah CA. Pada variabel *housing*, data yang mendominasi adalah BC. Pada *credit limit*, nilai yang mendominasi adalah 200.0. Pada kolom *source_of_application*, mayoritas *user* berasal dari internet. *os* yang paling banyak digunakan adalah *other*.

3. Distribusi variabel prediktor numerik



Dari plot histogram di atas, terlihat bahwa terdapat beberapa variabel numerik yang telah terdistribusi secara normal, sedangkan beberapa lainnya yang tidak berdistribusi normal, terutama pada variabel numerik yang belum di-*scaling*.

4. Correlation heat map



Dari visualisasi korelasi di atas, dapat terlihat terdapat beberapa pasang variabel yang memiliki korelasi yang kuat dan beberapa pasang lainnya memiliki korelasi yang lemah.

Praproses Data

Dari *data set* yang telah diberikan, terdapat beberapa tahapan praproses data pada analisis pemodelan *Random Forest Classifier*, yaitu:

1. Menghilangkan data duplikat

Data *train* memiliki 1.083.761 baris dan data *test* memiliki 286.803 baris. Setelah dilakukan praproses dengan menghilangkan data duplikat, didapatkan hasil bahwa banyak baris data *train* sebanyak 1.023.690 dan data *test* sebanyak 197.795.

2. Menghilangkan atribut yang tidak diperlukan

Dalam pemodelan ini, terdapat dua atribut yang dihilangkan, yaitu variabel *id* dan variabel *address_month_count*. Variabel *id* dihapus karena tidak memiliki pengaruh terhadap pemodelan, yang mana variabel tersebut hanya menjadi identitas dari setiap subjek pengamatan, sedangkan variabel *address_month_count* dihapus karena terdapat terlalu banyak *missing value*, yaitu sebanyak 80% dari data asli.

3. Mengisi nilai yang hilang (*missing values*)

Digunakan beberapa metode untuk mengisi nilai yang hilang atau *missing values*. Beberapa di antaranya adalah menggunakan mean dan median untuk data berskala numerik dan modus untuk data yang kategorik.

4. Penskalaan data

Digunakan fungsi *standard scaler* untuk melakukan penskalaan data. Variabel yang diskalakan adalah umur.

5. One Hot Encoding

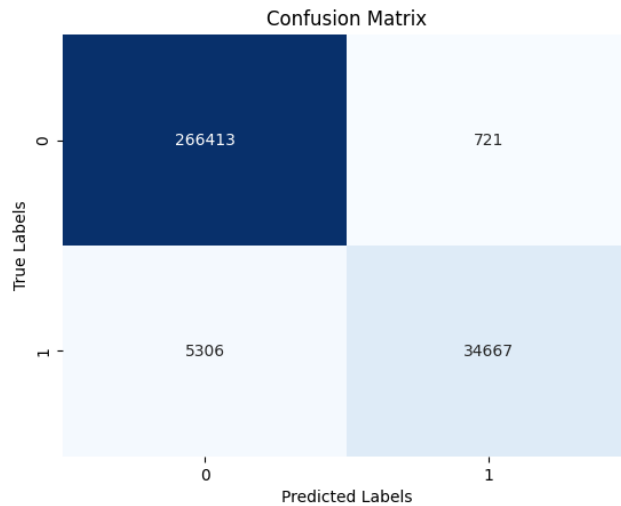
Proses *one hot encoding* diterapkan pada data kategorik untuk mempermudah proses pemodelan dengan mengubah data berskala nominal menjadi numerik. Atribut yang melalui proses *one hot encoding* adalah *payment*, *employment*, *housing*, *credit_limit*, *source_of_application*, dan *os*.

Hasil dan Pembahasan

Digunakan beberapa *metric* perhitungan untuk mengevaluasi model *Random Forest Classifier* yang telah digunakan. *Metric* yang akan digunakan antara lain *confusion matrix*, *precision*, *recall*, *F1-score*, dan *Receiver Operator Characteristic* (ROC).

1. Confusion Matrix

Confusion matrix dari pengklasifikasi pada penelitian ini ditunjukkan pada gambar di bawah. Matriks tersebut menunjukkan jumlah prediksi yang benar (TP = 266413 dan TN = 34667) dan jumlah prediksi salah (FP = 721 dan FN = 5306) yang dibangun oleh model.



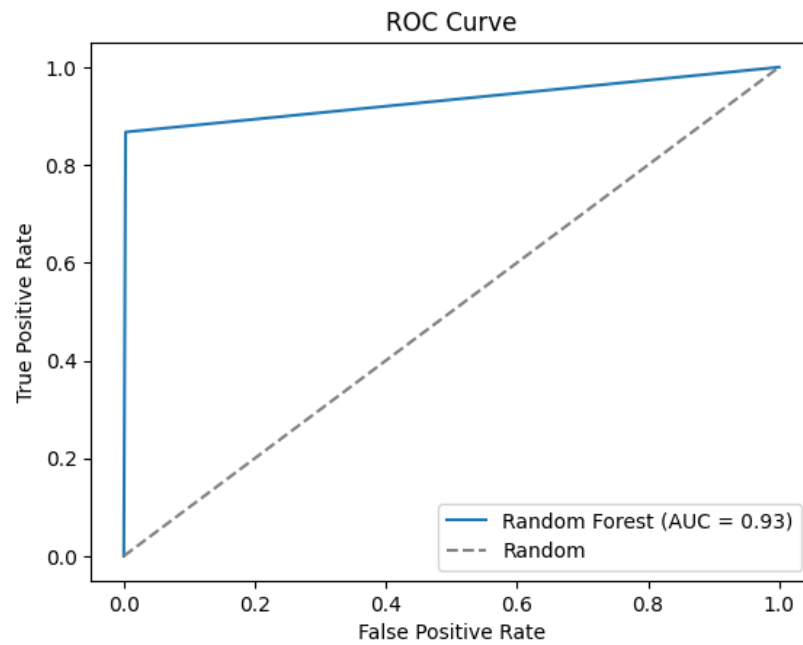
2. Akurasi, *Precision*, *Recall*, dan *F1- Score*

| Metrik | Skor |
|------------------|--------|
| Akurasi | 98,04% |
| <i>Precision</i> | 98,03% |
| <i>Recall</i> | 98,03% |
| <i>F1-score</i> | 98,03% |

Pemodelan menggunakan *Random Forest Classifier* memperoleh akurasi sebesar 98,04% untuk memprediksi penipuan, dengan skor presisi, *recall*, dan *F1* memiliki nilai yang sama, yaitu sebesar 98,03%. Berdasarkan hasil di atas, dapat dikatakan bahwa model dapat memprediksi variabel target yang diinginkan dengan baik.

3. *Receiver Operator Characteristic (ROC)*

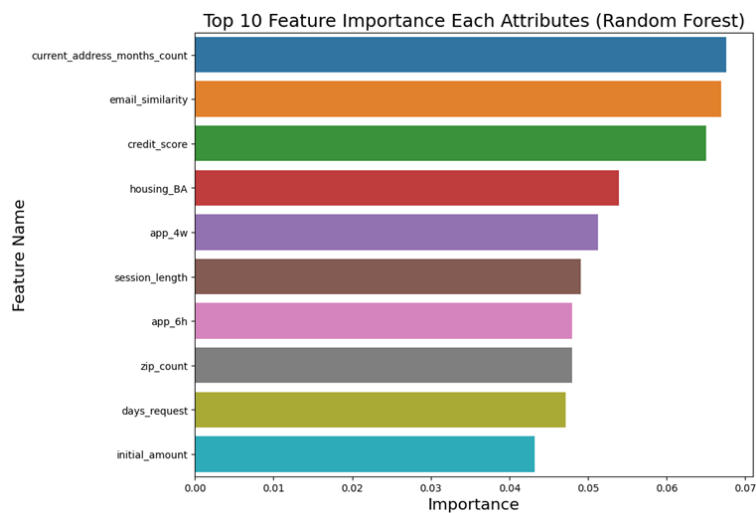
ROC curve menggambarkan hubungan antara *True Positive Rate (TPR)* dan *False Positive Rate (FPR)* pada berbagai *threshold* yang berbeda. Selain *ROC curve*, metrik evaluasi lain yang dihitung dari kurva ROC adalah AUC (*Area Under the ROC Curve*). AUC merupakan luas area di bawah kurva ROC. Nilai AUC berada dalam rentang 0 hingga 1. Semakin besar nilainya menunjukkan kinerja model yang semakin baik.



Berdasarkan gambar di atas, hasil AUC menggunakan *Random Forest Classifier* adalah 0,93. Artinya, model mampu mengklasifikasikan dengan benar 93% contoh positif sebagai hasil yang positif dan contoh negatif sebagai hasil yang negatif. Nilai ini juga memiliki arti bahwa tingkat positif palsu rendah, yaitu tingkat di mana model salah mengklasifikasikan contoh negatif sebagai hasil yang positif.

4. *Feature Importance*

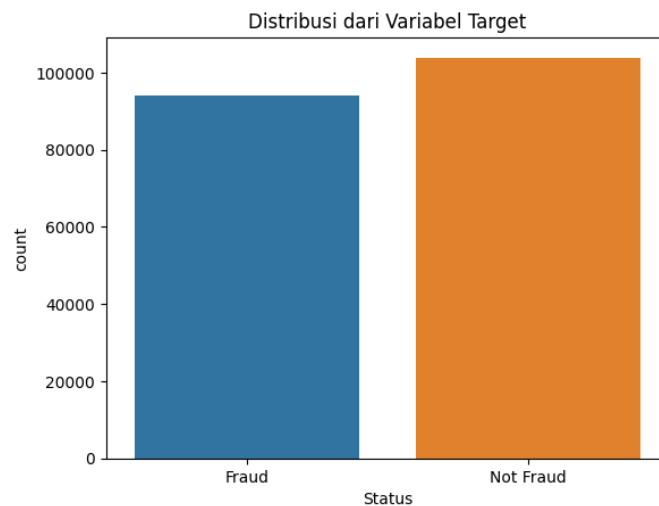
Melalui data *train*, didapatkan juga urutan sepuluh fitur atau variabel prediktor yang paling berpengaruh pada penentuan klasifikasi variabel target.



Terlihat bahwa sepuluh fitur terbaik yang memiliki nilai penting untuk pengklasifikasian adalah *current address months count*, *email similarity*, *credit score*, *housing BA*, *app 4w*, *session length*, *app 6h*, *zip count*, *days request*, dan *initial amount*.

Penerapan Model

Setelah dilakukan evaluasi model menggunakan data *train*, dilakukan penerapan model yang telah diuji pada data *test* yang diberikan. Melalui praproses data yang sama, didapatkan hasil prediksi penipuan dengan menggunakan data prediktor yang ada pada data *test* adalah sebagai berikut.



Terlihat bahwa melalui praproses dan pengklasifikasian menggunakan *Random Forest Classifier*, didapatkan hasil bahwa terdapat 93.938 akun yang diprediksi melakukan penipuan dan 103.857 akun yang diprediksi tidak melakukan penipuan.













KESIMPULAN

Dari pelatihan model pengklasifikasi *random forest*, didapatkan hasil evaluasi metrik berupa nilai skor akurasi sebesar 98,04%, presisi sebesar 98,03%, *recall* sebesar 98,03%, *F1-score* sebesar 98,03%, dan AUC (*Area Under the ROC Curve*) sebesar 93%. Sepuluh fitur terbaik yang mengindikasikan klasifikasi baik adalah *current address months count*, *email similarity*, *credit score*, *housing BA*, *app 4w*, *session length*, *app 6h*, *zip count*, *days request*, dan *initial amount*. Dari data uji, model memprediksi 93.938 akun mengalami penipuan dan 103.857 akun lainnya tidak mengalami penipuan.

DAFTAR PUSTAKA

- Aburbeian, A. M., & Ashqar, H. I. (2023). Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced Data. *arXiv*, 1-11.
- Ariyoga, D. (2022). *PERBANDINGAN METODE SELEKSI FITUR FILTER, WRAPPER, DAN EMBEDDED PADA KLASIFIKASI DATA NIRS MANGGA MENGGUNAKAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE (SVM)*. Sleman: Universitas Islam Indonesia.
- Aziz, W. A. (2021). *IMPLEMENTASI METODE RANDOM FOREST PADA KLASIFIKASI DATA ULASAN KONSUMEN PERUSAHAAN (Studi Kasus: Aplikasi KAI Access)*. Jakarta: Universitas Islam Negeri Syarif Hidayatullah.
- Fawcett, A. (2021, Februari 11). *Data Science in 5 Minutes: What is One Hot Encoding?* Diambil kembali dari educative: <https://www.educative.io/blog/one-hot-encoding>
- Informatika. (2012, Desember 12). *Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning*. Diambil kembali dari nursahid.com: <https://www.nursahid.com/mengenal-accuracy-precision-recall-dan-specificity-septa-yang-diprioritaskan-dalam-machine-learning>
- Kurniawan, A., & Yulianingsih. (2021). Pendugaan Fraud Detection pada Kartu Kredit dengan Machine Learning. *KILAT*, 320-325.
- Muttaqin, F. (2022, September 28). *4 Langkah Data Preprocessing Agar Data Lebih Mudah Dibaca*. Diambil kembali dari EKRUT media: <https://www.ekrut.com/media/data-preprocessing>
- Muttaqin, F. A., & Bachtiar, A. M. (2016). IMPLEMENTASI TEXT MINING PADA APLIKASI PENGAWASAN PENGGUNAAN INTERNET ANAK "DODO KIDS BROWSER". *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*, 1-8.
- Prabowo, H. Y. (2012). A better credit card fraud prevention strategy for Indonesia. *Emerald Insight*, 267-292.
- Trivusi. (2022, Juli 16). *Metriks Evaluasi Sistem Menggunakan Confusion Matrix*. Diambil kembali dari Trivusi: <https://www.trivusi.web.id/2022/04/evaluasi-sistem-dengan-confusion-matrix.html?m=1>
- van Plaosan, S. (n.d.). *Random Forest*. Retrieved from LearningBox: https://learningbox.coffeecup.com/05_2_randomforest.html
- Yunus, M. (2020, Januari 12). *#3 Machine Learning Evaluation*. Retrieved from Medium: <https://yunusmuhammad007.medium.com/3-machine-learning-evaluation-239426e3319e>

LAMPIRAN

| Submission and Description | Private Score ⓘ | Public Score ⓘ | Selected |
|--|-----------------|----------------|--------------------------|
|  submission (2).csv Complete · Yayayyy · 14d ago | 0.99148 | 0.99148 | <input type="checkbox"/> |
|  submission (1).csv Complete · Yayayyy · 14d ago | 0 | 0 | <input type="checkbox"/> |
|  bismillah5.csv Complete · Natasya Fatimah Salim · 14d ago | 0.9918 | 0.9918 | <input type="checkbox"/> |
|  submission.csv Complete · Yayayyy · 14d ago | 0.90424 | 0.90424 | <input type="checkbox"/> |
|  4-UAS-A - Version 6 Error · Bryan Florentino Leo · 14d ago · Bryan - Submitted 2 | | | |
|  4-UAS-A - Version 5 Complete · Bryan Florentino Leo · 14d ago | 0.69345 | 0.69345 | <input type="checkbox"/> |
|  bismillah4.csv Complete · Natasya Fatimah Salim · 18d ago | 0.99201 | 0.99201 | <input type="checkbox"/> |
|  bismillah3.csv Complete · Natasya Fatimah Salim · 18d ago | 0.90226 | 0.90226 | <input type="checkbox"/> |
|  bismillah2.csv Complete · Natasya Fatimah Salim · 18d ago | 0.93428 | 0.93428 | <input type="checkbox"/> |
|  bismillah1.csv Complete · Natasya Fatimah Salim · 18d ago | 0.94758 | 0.94758 | <input type="checkbox"/> |
|  bismillah.csv Error · Natasya Fatimah Salim · 18d ago | | | |
|  Submission_format - Submission_format.csv.csv Complete · Natasya Fatimah Salim · 19d ago | 0.91564 | 0.91564 | <input type="checkbox"/> |