Fraud Detection

Anggota Kelompok 6:
2702213530 - Sandy Agatha Indra Lim
2702209483 - Vania Oriana Tanoto
2702279086 - Ricky Atha Ajie Alvianto

I. Background Review

Cybercrime banking fraud adalah tindak kejahatan siber yang menargetkan sistem, layanan, atau nasabah perbankan dengan tujuan utama mencuri uang, data pribadi, atau informasi akses ke akun keuangan. Kejahatan ini sering dilakukan melalui teknologi digital seperti internet, malware, dan rekayasa sosial. Cybercrime banking fraud biasanya dilakukan dengan cara memanfaatkan kelemahan sistem perusahaan dan kelemahan manajemen yang ada. Bank Indonesia menyatakan bahwa setidaknya ada tiga modus kejahatan cyber yang sering menyerang sistem perbankan Indonesia, yaitu Skimming, Phising dan Malware. Skimming adalah tindakan mengambil data data para nasabah dengan menggunakan alat perekam. Biasanya hal tersebut terjadi di mesin EDC dan juga di mesin Anjungan Tunai Mandiri (ATM).

Phishing adalah modus kejahatan yang dilakukan dengan cara mengecoh para nasabah melalui pesan, email, atau situs palsu yang menyerupai institusi resmi seperti bank. Pelaku phising biasanya akan meminta korban untuk mengisi informasi sensitif seperti username, PIN, password, nomor kartu, atau kode OTP dengan alasan memverifikasi akun atau penanganan masalah keamanan, dimana padahal seharusnya data tersebut tidak dibagikan ke siapapun karena merupakan informasi yang berbahaya jika diketahui oleh orang lain. Tujuannya adalah mencuri data dan mengakses rekening korban secara ilegal. Teknik ini sangat berbahaya karena mengandalkan kelengahan dan kepercayaan korban, bukan kelemahan sistem teknis, sehingga bahkan bank pun tidak bisa melakukan tindakan apapun. Malware merupakan perangkat lunak berbahaya yang dikirim atau diinstal ke perangkat korban secara diam-diam. Malware dapat bekerja sebagai keylogger, yaitu merekam setiap ketikan pengguna termasuk informasi login, atau sebagai trojan, yaitu program yang tampak normal namun menjalankan aktivitas berbahaya di belakang layar.

Setelah terinfeksi, perangkat korban dapat dikendalikan atau dimata-matai oleh sang pelaku, sehingga ia dapat membuka akses terhadap informasi penting dan bahkan memungkinkan transaksi keuangan tanpa sepengetahuan korban. Ketiga modus ini—skimming, phishing, dan malware—sering kali dikombinasikan oleh pelaku untuk meningkatkan peluang keberhasilan. Oleh karena itu, perlindungan terhadap kejahatan perbankan digital tidak hanya bergantung pada keamanan sistem bank, tetapi juga pada kewaspadaan dan literasi digital para nasabah.

II. Literature Review

Dalam paper "Fraud Detection in Credit Card Transactions Using HDBSCAN, UMAP, and SMOTE", tantangan mendeteksi penipuan transaksi kartu kredit yang tersembunyi dalam dataset besar dan tidak berlabel secara langsung. Untuk menyelesaikan tantangan ini, peneliti menggunakan pendekatan unsupervised yang mengkombinasikan algoritma HDBSCAN sebagai metode clustering utama, UMAP untuk reduksi dimensi dan visualisasi pola, serta SMOTE sebagai teknik oversampling untuk mengatasi ketidakseimbangan kelas. Pendekatan ini berhasil mengidentifikasi anomali transaksi tanpa menggunakan label eksplisit. Meskipun tidak melibatkan model supervised secara langsung, hasil clustering dimanfaatkan untuk membentuk pseudo-label yang selanjutnya dapat digunakan dalam pelatihan model supervised. Paper ini menunjukkan bahwa pipeline unsupervised berbasis clustering dapat menjadi alternatif yang baik dalam mendeteksi penipuan, terutama saat data label sulit diperoleh.

Dijelaskan juga pada paper"Unsupervised Label Generation for Severely Imbalanced Fraud Data", tantangan utama dalam mendeteksi fraud adalah distribusi label yang sangat timpang, di mana kasus penipuan hanya sebagian kecil dari seluruh transaksi. Untuk mengatasinya, peneliti menggunakan pendekatan unsupervised label generation yang menggunakan teknik clustering seperti HDBSCAN dan anomaly detection untuk membentuk label pseudo-labels pada data yang sebelumnya tidak berlabel. Pseudo-label ini kemudian digunakan untuk melatih model supervised seperti Random Forest dan Gradient Boosting. Hasil eksperimen menunjukkan bahwa model supervised yang dilatih menggunakan pseudo-label tersebut mampu mendekati performa model yang dilatih dengan label asli. Pendekatan ini menjadi solusi menjanjikan untuk aplikasi nyata, khususnya ketika label asli sulit atau mahal untuk didapatkan.

Tertulis pada paper "Explainable Deep Behavioral Sequence Clustering for Transaction Fraud Detection", para peneliti fokus pada pemodelan perilaku transaksi keuangan secara berurutan dengan pendekatan deep learning dan clustering. Mereka mengembangkan framework berbasis Bi-LSTM untuk mengekstraksi representasi dari urutan transaksi, yang kemudian diproses oleh algoritma HDBSCAN versi GPU HDBSCAN untuk membentuk cluster perilaku. Cluster tersebut digunakan untuk membedakan pola transaksi normal dan abnormal, sekaligus menghasilkan pseudo-label yang kemudian digunakan dalam pelatihan model klasifikasi berbasis supervised. Sistem ini tidak hanya efektif dalam mendeteksi penipuan baru yang belum pernah terlihat sebelumnya, tetapi juga dirancang agar hasilnya interpretable, memudahkan analis keuangan memahami alasan di balik deteksi. Studi ini menunjukkan bahwa kombinasi deep learning, unsupervised clustering, dan pseudo-labeling dapat menghasilkan sistem deteksi fraud yang efisien, akurat, dan mudah dijelaskan.

III. Purpose and Benefits

1. Memprediksi apakah suatu transaksi merupakan tindakan penipuan atau tidak berdasarkan data historis.
2. Mengembangkan model machine learning yang mampu mendeteksi jenis-jenis transaksi mencurigakan.
3. Menyediakan sistem deteksi penipuan yang dapat memberikan peringatan secara real-time kepada pihak bank.
4. Mengurangi jumlah transaksi penipuan dan meminimalkan kerugian finansial.
5. Memberikan wawasan untuk perbaikan berkelanjutan terhadap sistem keamanan transaksi bank.

IV. Datasets

Dataset yang digunakan dalam proyek ini berasal dari LOL Bank Pvt. Ltd., yang berisi informasi transaksi perbankan beserta atribut-atribut pendukung lainnya. Dataset ini mencakup berbagai elemen penting seperti identitas nasabah, rincian transaksi, informasi merchant, hingga perangkat dan lokasi saat transaksi dilakukan. Dataset dapat diakses melalui link ini.

Dataset ini mencakup atribut-atribut utama yang berkaitan dengan transaksi, informasi nasabah, merchant, dan perangkat yang digunakan. Beberapa kolom utama dalam dataset ini antara lain:

- Customer_ID: ID unik untuk setiap nasabah.
- Customer_Name: Nama lengkap dari nasabah.
- Gender: Jenis kelamin nasabah (Male atau Female).
- Age: Usia nasabah dalam tahun.
- State: Negara bagian tempat tinggal nasabah.
- City: Kota tempat tinggal nasabah.
- Bank_Branch: Nama cabang bank tempat akun nasabah terdaftar.
- Account_Type: Jenis akun yang dimiliki nasabah.
- Transaction_ID: ID unik untuk setiap transaksi yang dilakukan.
- Transaction_Date: Tanggal terjadinya transaksi.
- Transaction_Time: Waktu terjadinya transaksi.
- Transaction_Amount: Jumlah uang yang ditransaksikan.
- Merchant_ID: ID unik dari merchant (penjual atau penyedia layanan) tempat transaksi dilakukan.
- Transaction_Type: Jenis transaksi seperti Transfer, Debit, atau Bill Payment.
- Merchant_Category: Kategori bisnis merchant, contohnya Groceries, Restaurant, Entertainment.
- Account_Balance: Saldo terakhir nasabah setelah transaksi.
- Transaction_Device: Perangkat atau metode yang digunakan saat transaksi.
- Transaction_Location: Lokasi geografis transaksi.
- Device_Type: Tipe perangkat yang digunakan saat transaksi.
- Is_Fraud: Label transaksi, bernilai 1 jika terindikasi fraud, dan 0 jika normal.

- Transaction_Currency: Mata uang transaksi, contohnya INR.
- Customer_Contact: Nomor kontak atau telepon nasabah.
- Transaction_Description: Deskripsi transaksi, menjelaskan tujuan atau jenis transaksi.
- Customer_Email: Alamat email nasabah

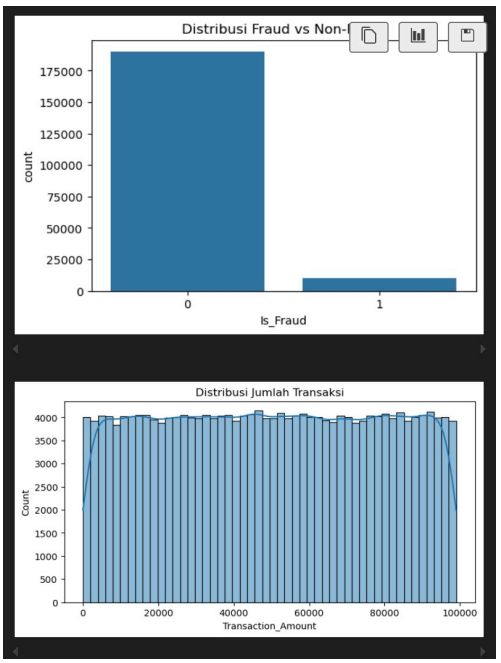
V. Methods

1. EDA

Sebelum memulai pemrosesan data, kami melakukan Exploratory Data Analysis (EDA) untuk memahami struktur dan karakteristik data transaksi perbankan dari LOL Bank Pvt. Ltd. Analisis awal mencakup pemetaan fitur kategorikal, visualisasi korelasi antar fitur, serta pengamatan distribusi variabel penting yang relevan terhadap deteksi fraud. Melalui langkah EDA ini, diperoleh gambaran menyeluruh mengenai struktur data dan fitur-fitur penting yang akan digunakan dalam tahap modeling, sekaligus menjadi landasan penting sebelum memasuki tahapan selanjutnya.

Dari visualisasi pertama, terlihat bahwa distribusi label Is_Fraud sangat tidak seimbang, mayoritas transaksi merupakan transaksi normal (label 0), sedangkan transaksi fraud (label 1) jumlahnya sangat sedikit. Ketidakseimbangan ini menandakan bahwa data bersifat imbalanced, sehingga dibutuhkan teknik khusus seperti resampling saat melatih model agar tidak bias terhadap kelas mayoritas.

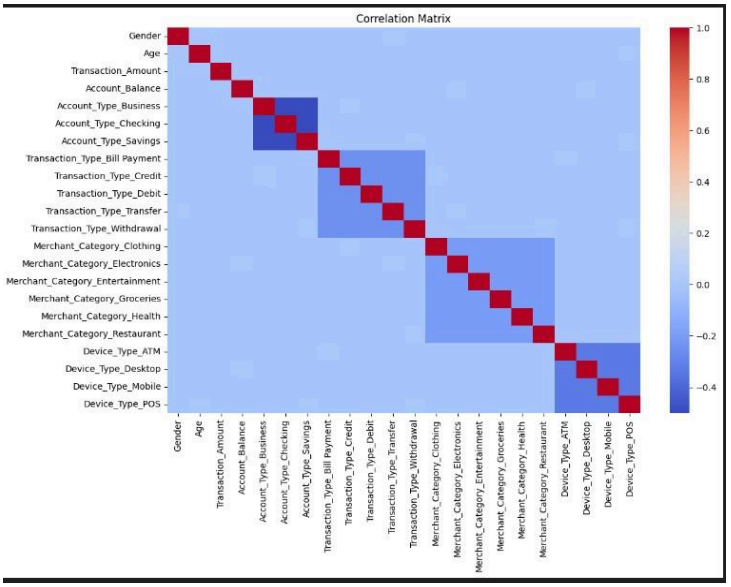
Sementara itu, distribusi jumlah transaksi relatif merata dari nominal kecil hingga besar, dengan bentuk distribusi yang cenderung simetris. Tidak terdapat outlier ekstrim yang mendominasi, sehingga model dapat memproses nilai numerik ini tanpa perlu banyak penyesuaian.



2. Pre-Processing

Untuk memastikan dataset transaksi LOL Bank dapat digunakan secara efektif dalam pemodelan deteksi penipuan, dilakukan tahapan pre-processing data secara menyeluruh. Beberapa kolom yang dianggap tidak relevan untuk analisis maupun pemodelan dihapus, seperti: Customer_ID, Customer_Name, Transaction_ID, Merchant_ID, Customer_Contact, Customer_Email, Transaction_Date, Transaction_Time, Transaction_Location, dan Transaction_Description. Penghapusan ini dilakukan untuk menyederhanakan struktur dataset dan menghindari noise yang tidak diperlukan dalam proses pelatihan model. Selanjutnya, fitur kategorikal ditangani dengan pendekatan encoding yang disesuaikan dengan karakteristik masing-masing fitur. Untuk fitur Gender, digunakan label encoding, di mana nilai 'Male' dikonversi menjadi 1 dan 'Female' menjadi 0, karena hanya memiliki dua kategori yang bersifat eksklusif. Sementara itu, fitur kategorikal lainnya seperti Account_Type, Transaction_Type, Merchant_Category, dan Device_Type memiliki lebih dari dua kategori yang tidak memiliki hubungan ordinal. Oleh karena itu, fitur-fitur tersebut dikodekan menggunakan One-Hot Encoding agar setiap kategori direpresentasikan sebagai kolom biner yang dapat dibaca dan diproses secara optimal oleh algoritma machine learning.

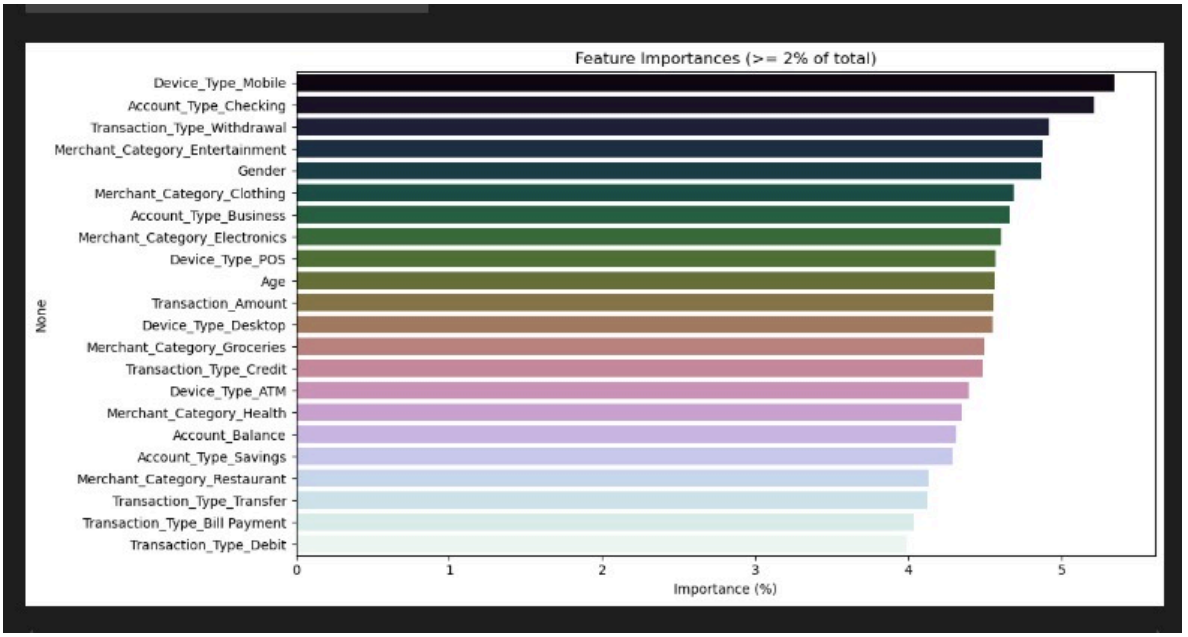
Setelah proses encoding dan penggabungan fitur numerik dan kategorikal, dilakukan analisis korelasi antar fitur menggunakan matriks korelasi (correlation matrix). Hasil visualisasi menunjukkan bahwa sebagian besar fitur memiliki korelasi yang sangat rendah satu sama lain (nilai mendekati nol), yang ditandai dengan warna biru terang pada plot. Tidak ditemukan korelasi tinggi (≥ 0.8) antar fitur, sehingga tidak ada multikolinearitas yang mengganggu. Beberapa fitur dari kategori yang sama, seperti antar tipe akun atau antar kategori device, memang menunjukkan korelasi negatif karena hasil dari proses one-hot encoding yang saling eksklusif. Namun, ini bersifat normal dan tidak mengindikasikan redundansi yang merugikan. Dengan hasil ini, seluruh fitur dalam dataset dianggap cukup independen dan layak untuk digunakan dalam proses pelatihan model tanpa perlu dilakukan penghapusan tambahan. Dataset akhir sudah berada dalam format numerik, seimbang secara struktur, dan siap digunakan dalam pipeline machine learning.



3. Feature Selection

Dalam tahap feature selection, dilakukan seleksi fitur berdasarkan nilai feature importance dari model XGBoost untuk meningkatkan efisiensi dan akurasi model deteksi fraud. Selain itu seleksi fitur juga berguna untuk mengurangi noise dan menghindari overfitting. Pemilihan threshold ≥ 2% pada feature importance digunakan sebagai strategi untuk menyeimbangkan antara mempertahankan fitur yang benar-benar relevan tanpa membanjiri model dengan fitur yang kontribusinya kecil. Hal ini penting karena pada dataset fraud detection, banyak fitur memiliki pengaruh yang sangat rendah terhadap prediksi dan justru berpotensi menjadi noise. Dengan menggunakan ambang batas 2%, hanya fitur-fitur yang menyumbang minimal 2% terhadap total bobot model yang dipertahankan.

Pendekatan ini juga berfungsi sebagai bentuk regularisasi manual untuk mencegah overfitting, khususnya pada data yang sangat tidak seimbang seperti data penipuan. Jika seluruh fitur digunakan, termasuk yang nilai importance nya di bawah 0.5%, maka model berisiko mempelajari pola yang tidak relevan. Selain itu, metode threshold 2% membantu menjaga interpretabilitas dan efisiensi model, karena biasanya hanya menghasilkan sekitar 8 hingga 15 fitur utama yang lebih mudah divisualisasikan dan diimplementasikan. Dengan threshold 2%, kita mengambil fitur yang memiliki kontribusi signifikan terhadap prediksi dan membuang "noise". (Misalnya, jika total 100 fitur, maka threshold 2% akan memilih hanya fitur-fitur yang menyumbang minimal 2% dari total bobot (berarti paling tidak satu fitur menyumbang 2% ke arah model decision).



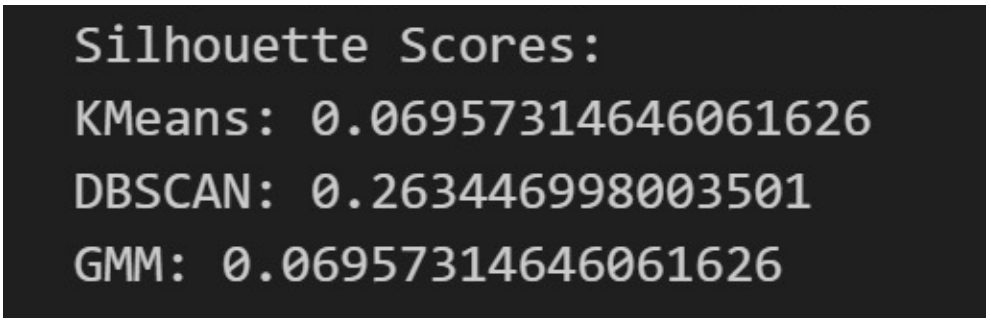
Dibandingkan dengan metode pemilihan fitur berdasarkan top-N (misalnya top 10 fitur teratas), penggunaan threshold ini lebih fleksibel karena disesuaikan dengan distribusi kontribusi fitur. Jika hanya ada enam fitur yang penting, threshold 2% akan memilih enam tersebut, tanpa memaksa menambahkan fitur lain yang sebenarnya kurang relevan.

```
Fitur yang dipilih (importance >= 2%):
Device_Type_Mobile                 5.34
Account_Type_Checking              5.21
Transaction_Type_Withdrawal        4.92
Merchant_Category_Entertainment    4.88
Gender                             4.87
Merchant_Category_Clothing         4.69
Account_Type_Business              4.66
Merchant_Category_Electronics      4.61
Device_Type_POS                    4.57
Age                                4.56
Transaction_Amount                 4.56
Device_Type_Desktop                4.55
Merchant_Category_Groceries        4.49
Transaction_Type_Credit            4.48
Device_Type_ATM                    4.39
Merchant_Category_Health           4.35
Account_Balance                    4.31
Account_Type_Savings               4.29
Merchant_Category_Restaurant       4.13
Transaction_Type_Transfer          4.12
Transaction_Type_Bill Payment      4.03
Transaction_Type_Debit             3.99
dtype: float32
```

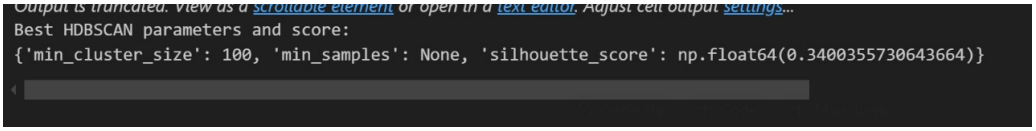
VI. Experiments

1. HDBSCAN

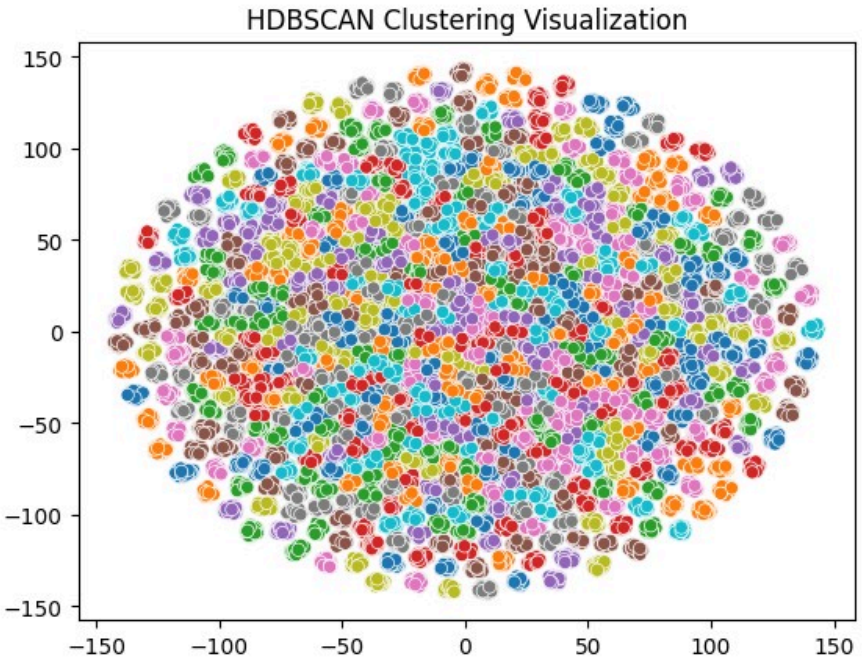
Pada tahap awal eksplorasi, kami mencoba beberapa algoritma unsupervised learning untuk melakukan clustering terhadap data transaksi yang tidak berlabel. Tiga metode utama yang diuji adalah KMeans, DBSCAN, dan Gaussian Mixture Model (GMM). Namun, setelah dilakukan evaluasi, pendekatan tersebut tidak memberikan hasil yang cukup baik dalam memisahkan transaksi normal dan mencurigakan, baik secara visual maupun berdasarkan metrik seperti silhouette score.



Sebagai alternatif, kami kemudian mencoba HDBSCAN, yang merupakan pengembangan dari DBSCAN dan dikenal lebih unggul dalam menangani data dengan kepadatan yang tidak merata. Dalam eksperimen awal, HDBSCAN menunjukkan hasil terbaik di antara semua metode.
Setelah itu, kami mencoba melakukan hyperparameter tuning terhadap HDBSCAN, terutama pada parameter min_cluster_size dan min_samples yang sangat berpengaruh dalam pembentukan cluster dan deteksi noise. Salah satu konfigurasi yang paling stabil adalah min_cluster_size = 100. Namun, proses tuning ini sangat memakan waktu, satu kombinasi parameter bisa memakan waktu hingga 4 jam, sehingga tidak dimasukkan ke dalam kode utama demi efisiensi proses dan keterbacaan proyek.



Meskipun performa awal HDBSCAN cukup baik, hasil akhirnya menunjukkan bahwa algoritma ini membentuk banyak cluster secara otomatis, karena memang tidak dirancang untuk menghasilkan dua kelas yang jelas. Hal ini menjadi tantangan, karena dalam proyek ini kami membutuhkan pemisahan biner (fraud vs non-fraud) untuk pelatihan model selanjutnya. Harapan bahwa transaksi fraud akan terkonsentrasi dalam satu atau dua cluster kecil tidak terpenuhi; justru data fraud tersebar ke berbagai cluster, bahkan ke cluster besar, sehingga strategi berbasis jumlah cluster tidak dapat digunakan secara langsung untuk membentuk label yang andal.



2. **Solusi Pertama: Membagi Cluster Kecil ke Kelompok Besar**

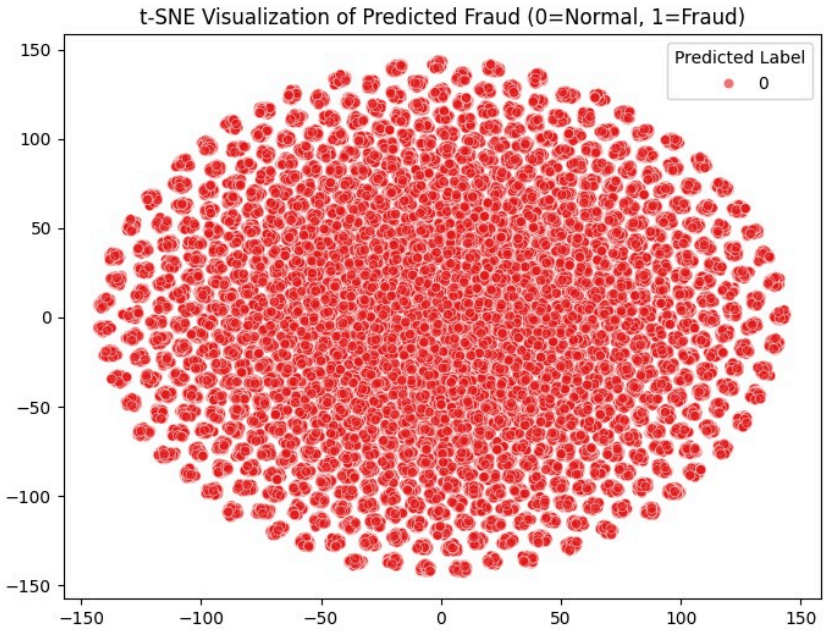
Percobaan awal kami dalam mendeteksi transaksi fraud secara unsupervised dilakukan dengan asumsi bahwa fraud merupakan anomali yang langka, yang seharusnya akan membentuk cluster kecil atau bahkan dianggap sebagai noise (-1) oleh algoritma HDBSCAN. Pendekatan ini berlandaskan pada pemahaman bahwa jika fraud adalah kejadian yang jarang terjadi, maka transaksi tersebut akan terkonsentrasi di dalam cluster yang berukuran lebih kecil dibandingkan transaksi normal. Oleh karena itu, kami mengasumsikan bahwa cluster kecil dalam hasil clustering HDBSCAN dapat digunakan sebagai penanda transaksi fraud, sementara cluster besar mencerminkan transaksi normal.

Untuk menguji asumsi ini, kami melakukan eksperimen dengan tuning threshold ukuran cluster, yaitu dengan mencoba berbagai nilai ambang batas dari 10 hingga 290, meningkat setiap 10 poin. Setiap nilai threshold digunakan untuk mengidentifikasi cluster yang dianggap sebagai fraud (misalnya: semua cluster dengan jumlah anggota < threshold dianggap fraud). Prediksi label dari strategi ini kemudian dibandingkan dengan label asli Is_Fraud, dan dievaluasi menggunakan F1-score untuk kelas fraud.

Namun, dari seluruh threshold yang diuji, tidak satu pun yang menghasilkan F1-score di atas 0.000. Ini menunjukkan bahwa strategi ini sepenuhnya gagal menangkap transaksi fraud, bahkan ketika cluster kecil telah ditandai sebagai fraud secara eksplisit. Bahkan setelah dilakukan pencarian threshold terbaik berdasarkan F1-score, nilai terbaik yang diperoleh tetap 0.000, dan classification report tidak menghasilkan metrik yang layak. Dengan

kata lain, tidak ada konfigurasi threshold yang mampu mengidentifikasi fraud dengan benar.

Eksperimen ini tidak disertakan dalam kode akhir demi menjaga efisiensi dan keterbacaan proyek. Namun, proses ini menjadi bagian penting dari eksplorasi awal yang menunjukkan bahwa ukuran cluster bukan indikator yang dapat diandalkan untuk mendeteksi transaksi fraud. Distribusi fraud yang tersebar dan tidak membentuk cluster khusus menjadikan pendekatan ini tidak efektif. Kegagalan ini mendorong kami untuk beralih ke strategi pseudo-labeling, yang lebih adaptif dalam menangani distribusi fraud yang tidak beraturan dan mendukung pembentukan model prediktif berbasis klasifikasi biner.



3. Solusi Kedua: Pseudo-Labeling

Setelah dilakukan clustering menggunakan metode unsupervised, langkah selanjutnya adalah menetapkan label semu (pseudo-label) berdasarkan hasil clustering. Dalam pendekatan ini, cluster terbesar yang terbentuk dianggap merepresentasikan transaksi normal, sementara cluster lainnya diasumsikan sebagai indikasi potensi fraud. Asumsi ini didasarkan pada sifat umum distribusi data transaksi, di mana sebagian besar data merepresentasikan perilaku normal dan hanya sebagian kecil yang merupakan outlier. Untuk menjaga proporsi data yang seimbang dalam pelatihan model berikutnya, dilakukan proses balancing dengan cara resampling data pseudo-label.

a. Supervised as Comparison

Sebagai pembanding terhadap pendekatan unsupervised dan semi-supervised, kami membangun model supervised menggunakan label asli Is_Fraud. Setelah membagi data dan menyeimbangkan kelas

dengan SMOTE, dilakukan pelatihan model XGBoost disertai hyperparameter tuning menggunakan RandomizedSearchCV dengan metrik recall agar model lebih sensitif terhadap fraud. Namun, hasil evaluasi menunjukkan bahwa meskipun label asli digunakan, model tidak mampu menghasilkan F1-score yang memadai pada kelas fraud. Hal ini menunjukkan bahwa pendekatan supervised penuh belum efektif dalam konteks data fraud yang sangat tidak seimbang. Oleh karena itu, model supervised ini hanya digunakan sebagai baseline, sementara pendekatan utama tetap difokuskan pada metode unsupervised dan semi-supervised melalui pseudo-label augmentation.

b. Semi-Supervised Learning via Pseudo-Label Augmentation

Pendekatan utama dalam proyek ini adalah semi-supervised learning menggunakan teknik pseudo-label augmentation. Proses diawali dengan melakukan clustering menggunakan HDBSCAN. Cluster terbesar dianggap mewakili transaksi normal (label 0), sedangkan cluster lainnya diberi label fraud (1). Untuk menyeimbangkan distribusi kelas, dilakukan oversampling terhadap data pseudo-label sebelum model dilatih.

Model klasifikasi biner kemudian dilatih menggunakan data hasil pseudo-labeling. Sebagian kecil data berlabel asli Is_Fraud digunakan secara terbatas sebagai referensi, namun tidak dijadikan label utama dalam pelatihan model yang akan diterapkan ke dalam aplikasi. Model supervised penuh yang menggunakan label asli tetap dibangun, namun hanya berfungsi sebagai baseline untuk evaluasi. Pendekatan utama tetap menjaga sifat unsupervised, namun diperkuat secara strategis melalui augmentasi terbatas dan evaluasi terkontrol.

Secara teknis, pendekatan ini dikategorikan sebagai semi-supervised learning. Model utama dilatih dengan label semu dari hasil clustering, lalu diperluas dengan pendekatan supervised. Dengan ini, sistem menggabungkan kekuatan eksplorasi pola unsupervised dan akurasi prediksi dari supervised, sangat cocok untuk kasus fraud detection dengan keterbatasan label.

```
--- EVALUASI SETELAH PSEUDO LABEL ---
Optimal threshold: 0.64608365
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.97     45579
           1       1.00      0.93      0.96     45579

    accuracy                           0.96     91158
   macro avg       0.97      0.96      0.96     91158
weighted avg       0.97      0.96      0.96     91158

Confusion Matrix:
 [[45566    13]
 [ 3280 42299]]
ROC AUC Score: 0.9715544357878938
Recall (fraud class): 0.9280370345992672
Accuracy: 0.9638759077645407
```

VII. Application

https://fraudulent-bank-transaction.streamlit.app/