

**UNIVERSITAS ESA UNGGUL**

**PENERAPAN MODEL *MACHINE LEARNING* UNTUK *FRAUD DETECTION*: STUDI KASUS PADA DATA TRANSAKSI KEUANGAN *MOBILE* SINTETIS**

**TUGAS AKHIR**

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana**

**NAMA : IMAM MULYANA**

**NIM : 20210801234**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS ESA UNGGUL**

**TAHUN 2025**

# HALAMAN PERNYATAAN KEASLIAN

Tugas Akhir ini adalah hasil karya saya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama : Imam Mulyana

NIM : 20210801234

Tanda Tangan :

|  |
| --- |
| (Materai 6000) |

Tanggal : 23 Januari 2025

# HALAMAN PENGESAHAN TUGAS AKHIR

Tugas Akhir ini diajukan oleh

Nama : Imam Mulyana

NIM : 20210801234

Program Studi : Teknik Informatika

Fakultas Ilmu Komputer - Universitas Esa Unggul

Judul Tugas Akhir : Penerapan Model *Machine Learning* untuk *Fraud Detection*: Studi Kasus pada Data Transaksi Keuangan *Mobile* Sintetis

**Telah berhasil dipertahankan dihadapan Tim Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Teknik Informatika Fakultas Ilmu Komputer, Universitas Esa Unggul.**

**TIM PENGUJI**

Pembimbing : M. Bahrul Ulum, S.Kom, M.Kom ( )

Penguji I : Dr. Vitri Tundjungsari, ST., M.Sc., M.M ( )

Penguji II : Agung Mulyo Widodo, S.T, M.Sc, Ph.D ( )

Ditetapkan di : Jakarta

Wakil Dekan : Diah Aryani ST, M.Kom

Tanggal : 19 Januari 2025

# HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

**UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademika Universitas Esa Unggul, saya yang bertanda tangan di bawah ini:

Nama : Imam Mulyana

NIM : 20210801234

Program Studi : Teknik Informatika

Fakultas : Ilmu Komputer

Jenis Karya Ilmiah : Tugas Akhir

demi pengembangan ilmu pengetahuan, teknologi, dan seni, menyetujui untuk memberikan kepada Universitas Esa Unggul Hak Bebas Royalti Non eksklusif atas karya ilmiah saya yang berjudul:

*APPLICATION OF MACHINE LEARNING MODELS FOR FRAUD DETECTION IN SYNTHETIC MOBILE FINANCIAL TRANSACTIONS*

beserta perangkat yang ada (apabila diperlukan). Dengan Hak Bebas Royalti Non eksklusif ini, Universitas Esa Unggul berhak menyimpan, mengalihmediakan, mengelola dalam bentuk pangkalan data, merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Jakarta

Pada tanggal : 23 Januari 2025

Yang menyatakan

(Imam Mulyana)

# KATA PENGANTAR

Puji syukur ke hadirat Allah SWT yang telah memberikan kesehatan, karunia serta hidayah-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul “Penerapan Model *Machine Learning* untuk *Fraud Detection*: Studi Kasus pada Data Transaksi Keuangan *Mobile* Sintetis”. Sebagai syarat untuk memperoleh gelar sarjana komputer (S.Kom) pada program studi Teknik Informatika Universitas Esa Unggul.

Penulis menyadari bahwa Tugas Akhir ini tidak dapat terselesaikan tanpa bantuan dari pihak lain, oleh karena itu izinkan saya selaku penulis untuk mengucapkan terima kasih yang sebesar besarnya kepada semua pihak yang telah membantu dalam menyelesaikan Tugas Akhir ini, terutama kepada:

1. Allah SWT atas segala Rahmat dan karunia-Nya.
2. Bapak Dr. Gerry Firmansyah, S.T, M.Kom, selaku Dekan Fakultas Ilmu Komputer Universitas Esa Unggul.
3. Ibu Diah Aryani ST, M.Kom selaku plt Kepala Program Studi Teknik Informatika dan Wakil Dekan Fakultas Ilmu Komputer Universitas Esa Unggul.
4. Bapak M. Bahrul Ulum, S.Kom, M.Kom selaku Dosen Pembimbing yang telah memberikan bimbingan, arahan, dorongan, dan semangat kepada penulis dalam penyusunan Tugas Akhir ini.
5. Ibu Dr. Vitri Tundjungsari, ST, M.Sc dan Bapak Agung Mulyo Widodo, S.T, M.Sc, Ph.D selaku dosen penguji.
6. Bapak Jefry Sunupurwa Asri, S.Kom, M.Kom selaku Dosen Pembimbing Akademik penulis.
7. Bapak Romy Ardianto, M.Sc selaku Mentor dan Data Scientist Kementrian Kelautan dan Perikanan Republik Indonesia yang selalu memberikan dukungan dan bimbingannya kepada penulis.
8. Kedua Orang Tua dan keluarga penulis yang selalu memberikan doa, dukungan moril, materil dan selalu memberikan motivasi yang luar biasa.
9. Seluruh Dosen dan Staff Fakultas Ilmu Komputer yang tidak bisa disebutkan satu persatu namanya.
10. Sahabat serta teman-teman terutama Ananda Dwi Rizkyta, Shefia Anggraeni, Bonifasius Dandy Krisnanda, dan Neng Yely Mulki K yang selalu memberikan doa, dukungan, saran, dan bantuannya dalam menyelesaikan Tugas Akhir ini.

Penulis menyadari bahwa proyek Tugas Akhir ini masih jauh dari kesempurnaan, maka saran dan kritik yang membangun dari semua pihak sangat diharapkan demi penyempurnaan selanjutnya.

Akhirnya hanya kepada Allah SWT kita kembalikan semua urusan dan semoga penulisan Tugas Akhir ini dapat bermanfaat bagi semua pihak, khususnya bagi penulis dan para pembaca pada umumnya.

Jakarta, 23 Januari 2025

Imam Mulyana

(20210801234)

# ABSTRAK

Judul : Penerapan Model *Machine Learning* untuk *Fraud Detection*: Studi Kasus pada Data Transaksi Keuangan *Mobile* Sintetis

Nama : Imam Mulyana

Program Studi : Teknik Informatika

Industri keuangan menghadapi tantangan dalam mendeteksi *fraud*. Laporan Indeks Basel Anti-Money Laundering (AML), pada tahun 2023 menunjukkan tren risiko pencucian uang yang memburuk selama lima tahun terakhir pada 107 negara. Dan menurut Financial Action Task Force (FATF) pada tahun 2023, hal ini diperparah oleh lembaga keuangan yang memiliki masalah dalam rendahnya pelaporan transaksi keuangan mencurigakan *(Suspicious Transaction Report)*. Keterbatasan akses terhadap data transaksi keuangan yang bersifat rahasia menjadi hambatan dalam pengembangan model *fraud detection* berbasis *machine learning*. Untuk mengatasi tantangan ini, penelitian menggunakan *dataset* sintetis PaySim yang meniru pola transaksi keuangan asli. Pendekatan CRISP-DM digunakan, mencakup tahap *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Algoritma yang digunakan adalah *Decision Tree*, *Random Forest*, dan *XGBoost*. Evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, *specifity,* *cross-validation* dan *ROC-AUC*. Hasil menunjukkan bahwa algoritma *XGBoost*  memiliki performa terbaik dengan akurasi 99,29%, diikuti oleh *Random Forest* (99,06%) dan *Decision Tree* (98,86%). Analisis data menunjukkan bahwa jenis transaksi *cash-out* dan *transfer* memiliki risiko *fraud* tertinggi. Model ini terbukti efektif dalam mendeteksi transaksi keuangan mencurigakan dengan tingkat akurasi tinggi. Penelitian ini memberikan kontribusi signifikan dalam mitigasi risiko keuangan, mendukung kebijakan *anti-fraud*, dan mendorong inovasi deteksi *fraud* menggunakan data sintetis.

Kata Kunci: *Anti-money laundering, Dataset sintetis, Fraud detection, Machine learning, PaySim, Random Forest.*

***ABSTRACT***

*Title : Application of Machine Learning Models for Fraud Detection: Case Study on Synthetic Mobile Financial Transaction Data*

*Name : Imam Mulyana*

*Study program : Informatics Engineering*

*The financial industry faces challenges in detecting fraud. The 2023 Basel Anti-Money Laundering (AML) Index report shows a worsening money laundering risk trend over the last five years in 107 countries. And according to the Financial Action Task Force (FATF) in 2023, this is exacerbated by financial institutions which have problems with low reporting of suspicious financial transactions (Suspicious Transaction Report). Limited access to confidential financial transaction data is an obstacle in developing machine learning-based fraud detection models. To overcome this challenge, the research uses PaySim synthetic datasets that mimic real financial transaction patterns. The CRISP-DM approach is used, including the Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment stages. The algorithms used are Decision Tree, Random Forest, and XGBoost. Model evaluation is carried out using accuracy, precision, recall, F1-score, specificity, cross-validation and ROC-AUC metrics. The results show that the XGBoost algorithm has the best performance with an accuracy of 99.29%, followed by Random Forest (99.06%) and Decision Tree (98.86%). Data analysis shows that cash-out and transfer transactions have the highest risk of fraud. This model has proven effective in detecting suspicious financial transactions with a high level of accuracy. This research makes a significant contribution to mitigating financial risks, supporting anti-fraud policies, and encouraging innovation in fraud detection using synthetic data.*

*Keywords :* *Anti-money laundering, Dataset sintetis, Fraud detection, Machine learning, PaySim, Random Forest.*

# DAFTAR ISI

Halaman

HALAMAN PERNYATAAN KEASLIAN ii

HALAMAN PENGESAHAN TUGAS AKHIR iii

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH iv

KATA PENGANTAR v

ABSTRAK vii

DAFTAR ISI ix

DAFTAR TABEL xi

DAFTAR GAMBAR xii

BAB 1 PENDAHULUAN 1

1.1 Latar Belakang 1

1.2 Identifikasi Masalah 4

1.3 Tujuan Tugas Akhir 4

1.4 Manfaat Tugas Akhir 4

1.5 Lingkup Tugas Akhir 5

1.6 Kerangka Berpikir 6

1.7 Sistematika Penulisan Tugas Akhir 6

BAB 2 TINJAUAN PUSTAKA 8

2.1 Penelitian Sebelumnya 8

*2.2* *Money Laundering* 13

*2.3* *Anti Money Laundering* 15

*2.4* *Machine Learning* 20

*2.5* *Fraud Detection* 20

2.6 Data Sintetis 20

*2.7* CRIPS-DM *(Cross-Industry Standard Process for Data Mining)* 21

*2.7.1* *Business Understanding* 22

*2.7.2* *Data Understanding* 22

*2.7.3* *Data Preparation atau Data Preprocessing* 22

*2.7.4* *Data Modeling* 27

*2.7.5* *Data Evaluation* 30

*2.7.6* *Deployment* 32

2.8 Tools 32

2.8.1 Python 32

2.8.2 Numpy 33

2.8.3 Pandas 33

2.8.4 Matplotlib 33

2.8.5 Scikit Learn 34

BAB 3 METODE 35

3.1 Rencana Penelitian 35

3.2 Objek Penelitian 41

3.3 Teknik Pengumpulan Data 42

3.4 Pendekatan Penelitian 45

3.5 Instrumen Penelitian 45

BAB 4 HASIL DAN PEMBAHASAN 47

*4.1* *Load Data* 47

4.2 *Preprocessing Data* dan Analisis Data 47

4.3 *Modeling* 56

4.4 *Evaluation* 61

*4.5* *Deployment* 68

BAB 5 KESIMPULAN DAN SARAN 71

5.1 Kesimpulan 71

5.2 Saran 71

DAFTAR REFERENSI 73

# DAFTAR TABEL

Halaman

[Tabel 2.1 Penelitian Sebelumnya 17](#_Toc190800567)

[Tabel 3.1 Data variabel PaySim 51](#_Toc190800568)

[Tabel 4.1 Variabel baru dataset PaySim 61](#_Toc190800569)

# DAFTAR GAMBAR

Halaman

[Gambar 1.1 Sebaran Negara Risiko AML 1](#_Toc190787092)

[Gambar 1.2 Kerangka Berpikir Penelitian 6](#_Toc190787093)

[Gambar 2.1 *Money Laundering Cycle*  14](#_Toc190787094)

[Gambar 2.2 Tahapan penanganan AML 16](#_Toc190787095)

[Gambar 2.3 Tahapan Metode CRISP-DM 21](#_Toc190787096)

[Gambar 2.4 Proses *Preprocessing* 23](#_Toc190787097)

[Gambar 2.5 Proses *Data Modeling* 27](#_Toc190787098)

[Gambar 2.6 *Confusion Matrik* 30](#_Toc190787099)

[Gambar 2.7 Rumus Evaluasi 31](#_Toc190787100)

[Gambar 2.8 Kurva ROC AUC 31](#_Toc190787101)

[Gambar 3.1 Proses Pengembangan M*achine Learning* 35](#_Toc190787102)

[Gambar 3.2 Pembagian Dataset 37](#_Toc190787103)

[Gambar 3.3 Cara kerja *Decision Tree* 37](#_Toc190787104)

[Gambar 3.4 Cara Kerja *Random Forest*  40](#_Toc190787105)

[Gambar 3.5 Cara Kerja *XGBoost* 40](#_Toc190787106)

[Gambar 4.1 Loading dataset 47](#_Toc190787107)

[Gambar 4.2 Mengecek informasi data 48](#_Toc190787108)

[Gambar 4.3 Mengecek nilai null 48](#_Toc190787109)

[Gambar 4.4 Mengecek duplikat data 49](#_Toc190787110)

[Gambar 4.5 Rename kolom 49](#_Toc190787111)

[Gambar 4.7 Jumlah *transaksi fraud* dan *non fraud* 50](#_Toc190787112)

[Gambar 4.8 Distribusi transaksi *fraud* dan *non fraud* berdasarkan type 51](#_Toc190787113)

[Gambar 4.9 *Feature Engineering* 51](#_Toc190787114)

[Gambar 4.10 *Droping* kolom 52](#_Toc190787115)

[Gambar 4.11 Distribusi transaksi *fraud* dan *non fraud* berdasarkan type2 53](#_Toc190787116)

[Gambar 4.12 Frekuensi Transaksi dalam 24 Jam 54](#_Toc190787117)

[Gambar 4.13 Melakukan *resampling* dan *label encoding* 54](#_Toc190787118)

[Gambar 4.14 *Splitting data* 55](#_Toc190787119)

[Gambar 4.15 *Corelation Matrik* 55](#_Toc190787120)

[Gambar 4.16 *Modelling Decision Tree* 56](#_Toc190787121)

[Gambar 4.17 *Modeling Random Forest* 57](#_Toc190787122)

[Gambar 4.18 *Modeling XGBoost* 58](#_Toc190787123)

[Gambar 4.19 Waktu pelatihan model 59](#_Toc190787124)

[Gambar 4.20 Visualisasi pohon model *decision tree* 60](#_Toc190787125)

[Gambar 4.21 *Cross-Validation* 61](#_Toc190787126)

[Gambar 4.22 Evaluasi model *decision tree* 62](#_Toc190787127)

[Gambar 4.23 Signifikan faktor *Decision Tree* 63](#_Toc190787128)

[Gambar 4.24 Evaluasi model *Random Forest* 64](#_Toc190787129)

[Gambar 4.25 Signifikan faktor *Random Forest* 65](#_Toc190787130)

[Gambar 4.26 Evaluasi model *XGBoost* 66](#_Toc190787131)

[Gambar 4.27 Signifikan faktor *XGBoost* 67](#_Toc190787132)

[Gambar 4.28 *Page prediction* 68](#_Toc190787133)

[Gambar 4.29 *Page Currency Exchange* 69](#_Toc190787134)

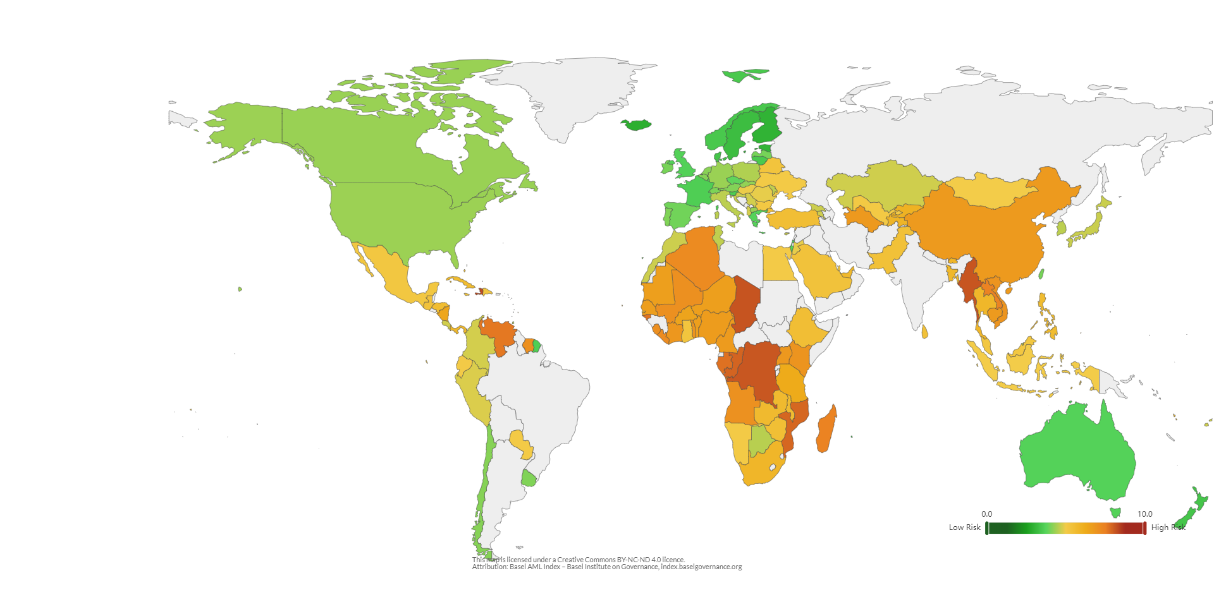
[Gambar 4.30 *Page History* 69](#_Toc190787135)

[Gambar 4.31 *Page Training* 70](#_Toc190787136)

# PENDAHULUAN

## Latar Belakang

Bidang keuangan menghadapi ancaman serius dari aktivitas penipuan terorganisir yang berdampak luas secara global. Laporan Indeks Basel *Anti-Money Laundering* (AML) (Basel, 2023), pada tahun 2023 menunjukkan tren risiko pencucian uang yang memburuk selama lima tahun terakhir pada 107 negara.



Gambar 1.1 Sebaran Negara Risiko AML (Basel, 2023)

Sedangkan menurut Financial Action Task Force (FATF) (*Executive Summary 1.*, 2023), pada laporan tahun 2023 lembaga keuangan memiliki masalah dengan rendahnya pelaporan transaksi mencurigakan (*Suspicious Transaction Report*) yang perlu ditingkatkan. Meskipun AML menjadi fokus utama, tetapi hanya sebagian kecil dari aktivitas tersebut yang berhasil dicegah. Karena data keuangan dalam bidang ini dilindungi dan tersembunyi, aksesnya sangat terbatas dan diawasi oleh otoritas regulasi (Naveed et al., 2023). *Fraud detection* adalah tugas yang krusial dalam industri keuangan untuk memastikan integritas dan kelangsungan organisasi keuangan (Rakesh Pandit, Sheetal Bawane, Jayesh Surana, 2024). *Money laundering* atau pencucian uang adalah proses mengaburkan sumber dana yang diperoleh dari kegiatan kriminal agar tampak legal dengan menggunakan cara-cara legal, seperti melalui struktur bisnis yang rumit atau transaksi internasional untuk menyamarkan uang ilegal (Isolauri & Ameer, 2023). Selain pelaku kriminal, berbagai pihak seperti firma hukum, lembaga, dan pejabat terkait juga dapat terlibat baik secara langsung maupun tidak langsung dalam manipulasi sistem hukum untuk mempermudah pencucian uang. Hal ini telah menjadi tantangan yang signifikan dan kompleks bagi perkembangan bisnis serta masyarakat global.

Metode konvensional seringkali tidak mampu menilai risiko dengan tepat dan mendeteksi aktivitas pencucian uang secara cepat. Dalam beberapa tahun terakhir, *machine learning* telah muncul sebagai alat yang efektif untuk meningkatkan proses ini (Rakesh Pandit, Sheetal Bawane, Jayesh Surana, 2024). Namun, hal ini terkendala terhadap akses data transaksi keuangan dalam melakukan pemrosesan data. Bagian dari masalah ini adalah sifat intrinsik yang bersifat pribadi dari data transaksi keuangan yang menyebabkan tidak adanya *dataset* publik yang tersedia secara terbuka. Data transaksi keuangan bersifat rahasia dan terbatas karena mengandung informasi sensitif seperti identitas pribadi dan data aliran transaksi keuangan, sehingga hanya lembaga jasa keuangan dan otoritas pemerintah terkait yang bisa mengaksesnya. Hal ini menjadi tantangan bagi para pengembang untuk mendapatkan akses data transaksi keuangan dalam melakukan *fraud detection* berbasis *machine learning*.

Terdapat satu model pendekatan untuk mengatasi permasalahan akses data transaksi keuangan yaitu melalui pembuatan *dataset* sintetis. Data sintetis merupakan suatu data yang dihasilkan melalui proses kecerdasan buatan untuk menciptakan data baru yang memiliki kemiripan dengan data aslinya. PaySim merupakan salah satu contoh data sintetis yang dihasilkan melalui simulator keuangan canggih yang dirancang untuk mensimulasikan transaksi keuangan *mobile* dengan akurasi tinggi berdasarkan data transaksi asli (Lopez-Rojas et al., 2016) (Lopez-Rojas & Barneaud, 2019). Simulator ini menghasilkan *dataset* sintetis yang sangat mirip dengan pola transaksi keuangan *mobile* yang sebenarnya sehingga dapat memberikan wawasan penting bagi peneliti dan praktisi dalam menganalisis perilaku transaksi. Dalam data transaksi yang tersedia hanya terdapat seperempat dari keseluruhan data asli, hal ini memungkinkan pengujian yang lebih terfokus.

Pengklasifikasian dalam *machine learning* menggunakan program komputer cerdas yang dapat digunakan untuk memeriksa transaksi *transfer* uang melalui ponsel. Tujuannya adalah untuk mengidentifikasi transaksi yang mungkin merupakan *fraud* atau transaksi keuangan mencurigakan yang berhubungan dengan pencucian uang. Program ini belajar dari data transaksi sebelumnya untuk mengenali pola-pola yang menunjukkan adanya potensi transaksi *fraud* atau transaksi keuangan mencurigakan, sehingga bisa memberitahu jika ada transaksi yang mencurigakan atau tidak biasa (Lokanan, 2023).

Penelitian terkait penerapan *machine learning* untuk *fraud detection* pada data transaksi keuangan *mobile* sintetis sudah pernah dilakukan oleh beberapa peneliti diantaranya oleh Lokanan, dimana hasilnya menunjukkan bahwa *Random Forest* adalah model paling akurat dengan tingkat akurasi 89% (Lokanan, 2023). Selanjutnya Naveed, dimana hasilnya model GCN mampu mencapai akurasi 77-79% dalam mendeteksi *fraud detection* (Naveed et al., 2023). Penelitian dari Hajek, Botcheey, dan Nobel menunjukan hasil yang paling optimal, dimana hasilnya mampu mencapai tingkat akurasi pada *fraud detection* masing-masing sebesar 99,98%, 99,88%, dan 99,90% dengan model *XGBoost* (Hajek et al., 2023) (Nobel et al., 2024) (Botchey et al., 2020).

Lembaga jasa keuangan dapat memanfaatkan pendekatan ini untuk merancang kebijakan yang lebih efektif dalam mengurangi risiko *fraud* atau aktivitas keuangan yang mencurigakan. Kemampuan untuk secara cepat memprediksi risiko pencucian uang menjadi sangat penting agar masalah tersebut dapat ditangani dengan lebih efektif dan efisien. Oleh karena itu, dalam penelitian ini, kami memperkenalkan metode inovatif yang menggunakan teknik *machine learning* untuk mendeteksi *fraud* atau transaksi keuangan mencurigakan. Metode ini diterapkan pada data transaksi keuangan *mobile* yang bersifat sintetis dengan tujuan meningkatkan ketepatan dan kecepatan dalam mengidentifikasi potensi pelanggaran keuangan. Dengan demikian, lembaga keuangan dapat mengambil tindakan pencegahan lebih awal dan memperkuat sistem pengawasan mereka terhadap risiko kejahatan keuangan.

## Identifikasi Masalah

Berdasarkan permasalahan sebelumnya, maka dapat diidentifikasi masalah yang dibahas sebagai berikut:

1. Bagaimana mengatasi keterbatasan akses terhadap data keuangan yang bersifat rahasia untuk mendukung pengembangan model *fraud detection* berbasis *machine learning*?.
2. Bagaimana metode berbasis *machine learning* dapat menggantikan metode tradisional dalam mendeteksi *fraud* atau transaksi keuangan yang mencurigakan secara cepat dan efektif?.
3. Bagaimana memilih model *machine learning* yang tepat untuk mendeteksi *fraud* atau transaksi keuangan yang mencurigakan dengan hasil yang akurat dan sesuai kebutuhan dalam pengembangan *fraud detection*?*.*

## Tujuan Tugas Akhir

Dalam penelitian ini terdapat beberapa tujuan yang ingin dicapai adalah sebagai berikut:

1. Melakukan pengumpulan dan identifikasi data transaksi keuangan *mobile* sintetis yang dihasilkan oleh simulator PaySim.
2. Mengembangkan model *fraud detection* atau transaksi keuangan mencurigakan berbasis *machine learning* yang dapat mengatasi keterbatasan akses data transaksi keuangan secara efektif.
3. Mengimplementasikan model *machine learning* dalam *fraud detection* yang mampu menilai risiko secara lebih efisien dan meningkatkan kemampuan lembaga keuangan dalam mengidentifikasi *fraud* atau transaksi keuangan mencurigakan.
4. Mengevaluasi efektivitas model *machine learning* dalam mendeteksi aktivitas *fraud* secara cepat dan akurat dengan membandingkan hasil model.

## Manfaat Tugas Akhir

Dengan adanya penelitian ini diharapkan dapat memberikan manfaat, yaitu:

1. Penelitian ini akan memberikan kontribusi signifikan dalam meningkatkan kemampuan lembaga keuangan untuk mendeteksi *fraud* atau transaksi keuangan mencurigakan secara lebih akurat dan cepat sehingga dapat mengurangi risiko kerugian finansial akibat aktivitas *fraud.*
2. Dengan menggunakan model *fraud detection* atau transaksi keuangan mencurigakan berbasis *machine learning* yang dihasilkan dari *dataset* sintetis PaySim, lembaga keuangan dapat merumuskan kebijakan keamanan yang lebih efektif dalam menangani berbagai bentuk *fraud* terutama yang terkait dengan transaksi keuangan *mobile*.
3. Penelitian ini akan menyediakan kerangka kerja yang dapat digunakan oleh pengembang lain untuk mengembangkan model *fraud detection* atau transaksi keuangan mencurigakan yang lebih canggih meskipun dihadapkan pada keterbatasan akses data keuangan yang bersifat rahasia sehingga dapat mendorong inovasi lebih lanjut dalam bidang keamanan keuangan global.

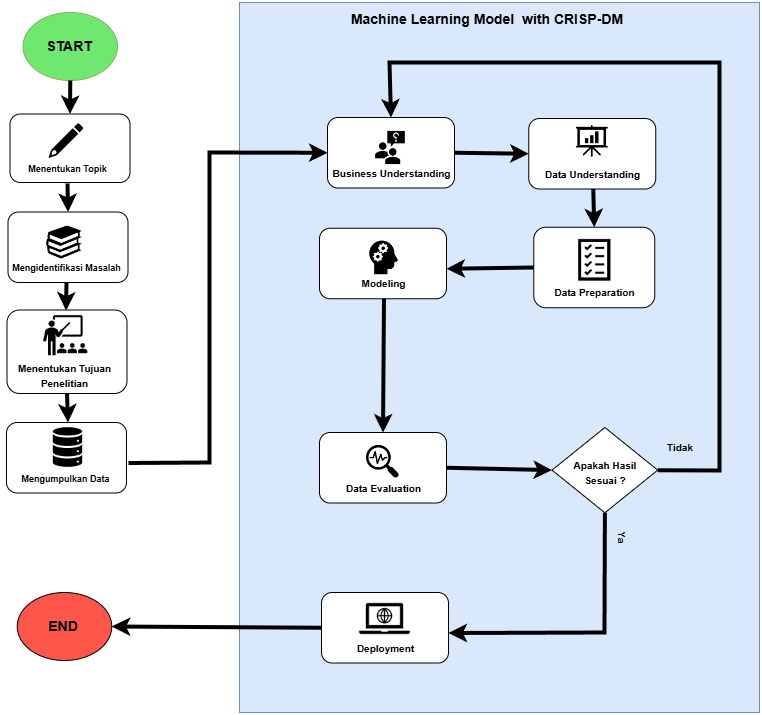
## Lingkup Tugas Akhir

Lingkup tugas akhir yang dirumuskan dalam penelitian yaitu:

1. Penelitian ini akan fokus pada penggunaan *dataset* sintetis yang dihasilkan oleh simulator PaySim untuk mengembangkan dan menguji model *fraud detection* atau transaksi keuangan mencurigakan. *Dataset* lain yang tidak tersedia secara publik atau yang memerlukan izin khusus tidak akan digunakan dalam penelitian ini.
2. Penelitian ini akan membatasi *fraud detection* atau transaksi keuangan mencurigakan pada transaksi keuangan yang dilakukan melalui keuangan *mobile*. Bentuk-bentuk *fraud* atau transaksi keuangan mencurigakan lainnya seperti *fraud* melalui *transfer* bank tradisional atau *platform* pembayaran digital lainnya tidak akan menjadi fokus utama dalam penelitian ini.

## Kerangka Berpikir

Berdasarkan identifikasi masalah yang telah dibuat, maka dibuatlah kerangka berpikir guna untuk memperjelas apa saja yang menjadi tujuan penelitian.



Gambar 1.2 Kerangka Berpikir Penelitian

## Sistematika Penulisan Tugas Akhir

Penyusunan skripsi ini dibagi dalam lima bab, uraian dan penjelasan secara singkat sebagai berikut:

**BAB 1 PENDAHULUAN**

Bab ini berisi uraian tentang latar belakang, identifikasi masalah, tujuan, lingkup, dan manfaat dibuatnya penulisan skripsi ini.

**BAB II TINJAUAN PUSTAKA**

Bagian ini menyajikan tinjauan pustaka dan landasan teori sebagai kerangka acuan untuk memahami permasalahan penelitian serta hasil-hasil penelitian terdahulu yang relevan.

**BAB III METODE PENELITIAN**

Pada bab ini membahas rencana penelitian, obyek penelitian, teknik pengumpulan data, metode pengembangan sistem.

**BAB IV HASIL DAN PEMBAHASAN**

Pada bab ini membahas tentang hasil dan pembahasan dari *fraud detection* yang telah dilakukan dengan menerapkan metode *machine learning.*

**BAB V KESIMPULAN DAN SARAN**

Pada bab ini membahas tentang kesimpulan dan saran dari penelitian yang telah dibuat.

# TINJAUAN PUSTAKA

## Penelitian Sebelumnya

Penelitian terkait *fraud detection* telah banyak dilakukan di bidang keuangan, oleh karena itu diperlukan studi literatur terhadap penelitian sebelumnya. Studi literatur adalah teknik pengumpulan data dengan mengadakan studi penelaahan terhadap buku-buku, literatur-literatur, catatan-catatan, dan laporan-laporan yang ada hubungannya dengan masalah yang dipecahkan (Yeni & Hartati, 2020). Penulis menggunakan beberapa penelitian terdahulu sebagai bahan analisis dan juga referensi dalam pembuatan tugas akhir.

Berikut adalah hasil kajian pustaka yang diperoleh melalui pengumpulan beragam penelitian terdahulu yang terkait dengan topik yang dibahas. Beberapa diantaranya mencakup:

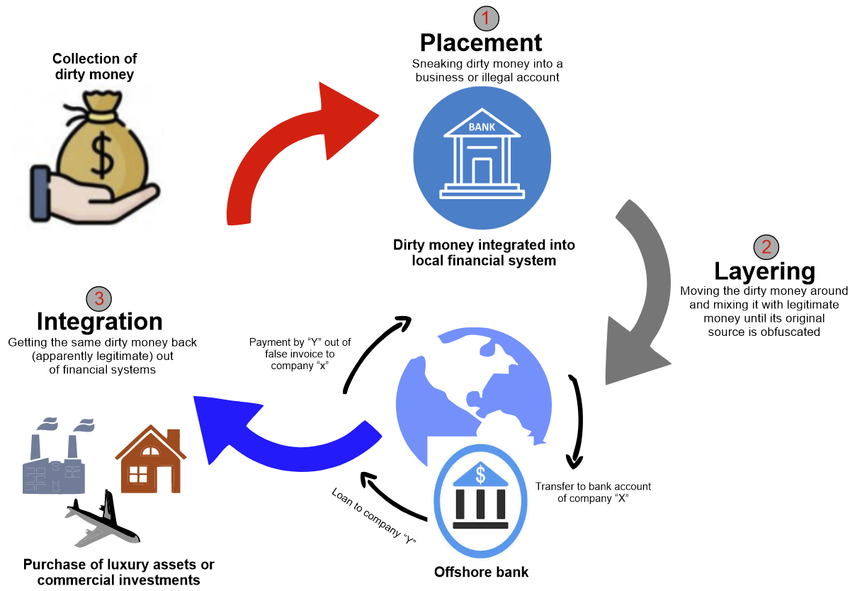
Tabel 2.1 Penelitian Sebelumnya

|  |  |  |  |
| --- | --- | --- | --- |
| NO | Nama Jurnal | Peneliti | Hasil Penelitian |
| 1 | *Predicting Mobile Money Transaction Fraud using Machine Learning Algorithms* | (Lokanan, 2023) | Penelitian ini menggunakan berbagai algoritma *machine learning*, seperti *Logistic Regression, Random Forest*, dan S*tochastic Gradient Descent* untuk mendeteksi *fraud* dalam transaksi uang elektronik dengan menggunakan data yang dihasilkan dari PaySim dan teknik SMOTE-ENN digunakan untuk menangani ketidak seimbangan kelas. Hasilnya menunjukkan bahwa *random forest* adalah model paling akurat dengan tingkat akurasi 89% sehingga menjadikannya model yang efektif dalam mendeteksi aktivitas *fraud* dan pencucian uang. |
| 2 | *Estimating Financial Fraud through Transaction-Level Features and Machine Learning* | (Alwadain et al., 2023) | Penelitian ini menggunakan 27 algoritma *machine learning* termasuk *XGBoost* dan *Random Forest* serta menghasilkan sampel sintetis menggunakan CTGAN untuk mengatasi ketidak seimbangan data transaksi *fraud.* Hasilnya menunjukkan bahwa *XGBoost* memberikan kinerja terbaik dengan *accuracy* 99,9% dan *F1-score* 0,999 dalam mendeteksi transaksi *fraud.* |
| 3 | *Intelligent Anti-Money Laundering Fraud Control Using Graph-Based Machine Learning Model for the Financial Domain* | (Naveed et al., 2023) | Penelitian ini menggunakan model *Graph Convolutional Networks* (GCN) dengan metode *semi-supervised learning* untuk mendeteksi aktivitas pencucian uang menggunakan data sintetis dari AMLSim yang mencakup 100 ribu akun dan 5,3 juta transaksi. Hasilnya menunjukkan bahwa model GCN mampu mencapai akurasi 77-79% dalam mendeteksi transaksi keuangan yang mencurigakan. |
| 4 | *Unmasking Banking Fraud: Unleashing the Power of Machine Learning and Explainable AI (XAI) on Imbalanced Data* | (Nobel et al., 2024) | Penelitian ini menggunakan empat algoritma *machine learning* termasuk *XGBoost,* *SVM, Decision Tree,* dan *Logistic Regression* dengan teknik SMOTE untuk menangani ketidak seimbangan data serta *SHAP* dan *LIME* untuk menjelaskan keputusan model *XGBoost*. Hasilnya menunjukkan bahwa *XGBoost* adalah model yang paling unggul dengan *accuracy* 99,88%, *precision* 0,96, dan *recall* 0,88 dalam mendeteksi *fraud.* |
| 5 | *Financial Fraud Detection through the Application of Machine Learning Techniques: A Literature Review* | (Hernandez Aros et al., 2024) | Penelitian ini menggunakan metode tinjauan literatur sistematis (SLR) dengan pendekatan PRISMA dan Kitchenham, mengkaji 104 artikel tentang deteksi penipuan keuangan menggunakan *machine leraning*. Hasilnya menunjukkan bahwa model *machine learning* seperti *Random Forest, Logistic Regression,* dan SVM adalah model yang memberikan hasil terbaik dalam mendeteksi *fraud*, diukur menggunakan metrik seperti *Precision, Recall,* dan *F1 Score.* |
| 6 | *Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework* | (Hajek et al., 2023) | Penelitian ini menemukan bahwa metode *XGBoost* dengan teknik *random under-sampling* mampu mencapai tingkat akurasi deteksi penipuan sebesar 99,98%. Disisi lain, metode *XGBOD* berhasil memperoleh nilai AUC (*Area Under Curve*) sebesar 99,58%. |
| 7 | *Mobile Money Fraud Prediction—A Cross-Case Analysis on the Efficiency of Support Vector Machines, Gradient Boosted Decision Trees, and Naïve Bayes Algorithms* | (Botchey et al., 2020) | Penelitian ini menggunakan algoritma *machine learning* SVM*,* GBDT*, dan Naïve Bayes*, dengan teknik penanganan ketidak seimbangan data seperti *undersampling* dan *oversampling* untuk mendeteksi *fraud* pada transaksi keuangan *mobile*. Hasilnya menunjukkan bahwa GBDTmemberikan performa terbaik dengan akurasi 99,90% dan *F1-Score* 99,95%, sedangkan *Naïve Bayes* juga menunjukkan hasil yang baik setelah *oversampling,* dengan akurasi 88,97% dan *F1-Score* 90%. |
| 8 | *A Synthetic Data Set to Benchmark Anti-Money Laundering Methods* | (Jensen et al., 2023) | Penelitian ini menggunakan dataset sintetis SynthAML yang dibangun dengan perpustakaan *Synthetic Data Vault* (SDV) dan *Gaussian copulas* yang meniru pola transaksi *anti-money laundering* (AML) berdasarkan data asli dari Spar Nord Denmark. Hasilnya menunjukkan bahwa model *machine learning* yang dilatih pada data sintetis seperti *Logistic Regression* dan *Gradient Boosted Trees* dapat men*transfer* performa ke data asli dengan konsistensi yang baik mencapai ROC AUC hingga 64,48% |
| 9 | *Advantages of the PaySim Simulator for Improving Financial Fraud Controls* | (Lopez-Rojas & Barneaud, 2019) | Penelitian ini menggunakan simulator PaySim yang menghasilkan *dataset* sintetis dari data keuangan anonim yang menyerupai dinamika dan statistik transaksi asli, kemudian digunakan untuk mempelajari skenario *fraud* dan mengevaluasinya. Hasil simulasi menunjukkan bahwa kontrol *fraud* yang lebih ketat mengurangi *fraud* yang tidak terdeteksi, tetapi meningkatkan jumlah kondisi dimana sistem memprediksi hasil positif, tetapi kenyataannya hasil tersebut negatif. |
| 10 | *PaySim: A Financial Mobile Money Simulator for Fraud Detection* | (Lopez-Rojas et al., 2016) | Penelitian ini menggunakan simulator PaySim berbasis *Multi-Agent Based Simulation* (MABS) untuk menghasilkan data sintetis yang menyerupai transaksi keuangan *mobile* asli, dengan jenis transaksi seperti *Cash-in, Cash-out, Debit, Payment,* dan *Transfer,* berdasarkan data asli dari log transaksi keuangan perusahaan telekomunikasi multinasional di Afrika. Hasil simulasi menunjukkan bahwa dataset sintetis yang dihasilkan sangat mirip dengan data asli, memungkinkan penggunaan untuk penelitian *fraud detection.* |

## *Money Laundering*

*Money Laundering* atau Pencucian uang adalah upaya menyembunyikan atau menyamarkan uang atau dana yang diperoleh dari suatu aksi kejahatan atau hasil tindak pidana sehingga seolah-olah tampak menjadi harta kekayaan yang sah (Ginting et al., 2023).

Dalam prakteknya, kegiatan pencucian uang mencakup tiga langkah yang menjadi dasar operasional pencucian uang (N. Gilmour, 2023), yaitu :



Gambar 2.1 *Money Laundering Cycle* (Karim et al., 2024)

1. *Placement*

*Placement* atau penempatan uang adalah proses masuknya uang tunai ke dalam sistem finansial yang sah (Rose, 2020). Pada tahapan ini, pergerakan uang sangat rawan untuk dideteksi maka untuk menghindari terdeteksinya transaksi keuangan, cara yang biasa dilakukan adalah dengan memecah uang menjadi satuan yang lebih kecil agar tidak mudah dicurigai (Gunawan, 2024). Di samping itu terdapat cara lain, yaitu dengan menempatkan uang tersebut ke dalam instrumen penyimpanan uang yang berbeda-beda seperti cek dan deposito, menyelundupkan uang atau harta hasil tindak pidana ke negara lain, melakukan penempatan secara elektronik, dan menggunakan beberapa pihak lain dalam melakukan transaksi.

1. *Layering*

*Layering* adalah proses menghilangkan jejak uang yang berasal dari aktivitas ilegal dengan membuat berbagai lapisan yang mengaburkan asal-usul uang dengan cara memisahkan uang dari sumber aslinya (P. M. Gilmour, 2023). Lapisan-lapisan ini bisa mencakup berbagai transaksi atau metode lain yang digunakan untuk menyembunyikan asal-usul uang. Sering kali metode ini menggunakan bank atau lokasi di negara-negara lain yang memiliki aturan kerahasiaan bank yang ketat sehingga semakin sulit bagi pihak berwenang untuk melacak jejak aliran dana tersebut (Laowo, Sebastian, 2022).

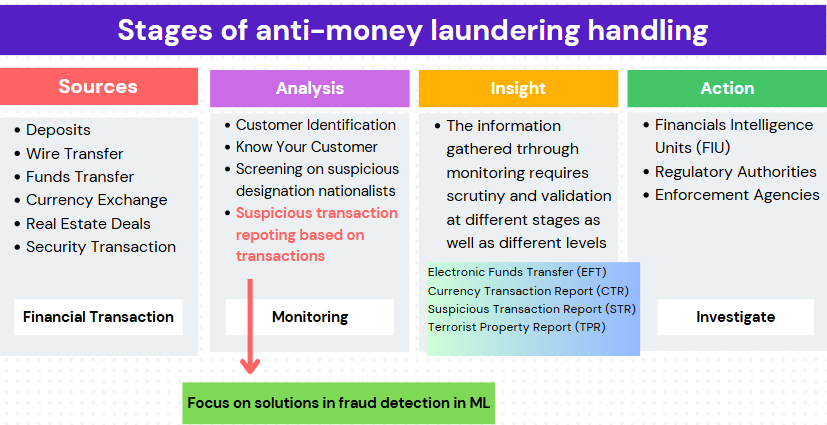
1. *Integration*

*Integration* adalah upaya untuk menyatukan atau memanfaatkan kekayaan yang telah tampak legal setelah melalui tahap *placement* dan *layering*. Pada tahap ini, pelaku mengalirkan dana yang telah dimanipulasi transaksi agar tampak legal ke dalam kegiatan yang sah, seperti investasi dalam produk keuangan, pembelian aset, atau pembiayaan bisnis. Para pelaku menggunakan instrumen finansial yang sama dengan yang dipakai di tahap sebelumnya dengan tujuan untuk membuat asal-usul uang tersebut tampak sah. Uang ini bisa digunakan secara langsung, diinvestasikan, atau bahkan untuk mendanai kegiatan kriminal kembali tanpa menimbulkan kecurigaan (Laowo, Sebastian, 2022).

## *Anti Money Laundering*

*Anti-money laundering* (AML) adalah serangkaian tindakan, kebijakan, dan prosedur yang dirancang untuk mencegah pencucian uang. AML sering digunakan di bidang keuangan dan hukum untuk meneliti, mencegah, dan memerangi aktivitas pencucian uang dan pendanaan teroris. (Reznik et al., 2023). Sistem *Anti-money laundering* memungkinkan lembaga keuangan untuk menerapkan prosedur otomatis guna mengawasi aktivitas keuangan nasabah secara *real-time* sehingga perilaku transaksi dapat dianalisis untuk mengidentifikasi pola transaksi keuangan yang mencurigakan. Hal ini dapat membantu dalam mendeteksi kemungkinan kegiatan ilegal dalam *fraud detection*. Proses otomatis memungkinkan percepatan dalam identifikasi risiko *money laundering* sehingga bank dapat segera mengambil tindakan bila ada indikasi aktivitas mencurigakan dalam transaksi keuangan (Kannan et al., 2022).

Dalam penerapannya AML memiliki 4 komponen, yaitu:



**Gambar 2.2** Tahapan penanganan AML

Sumber pencucian uang dapat berasal dari berbagai jenis transaksi keuangan yang dirancang untuk menyembunyikan asal-usul dana ilegal. Misalnya, *Deposits* (setoran) bisa digunakan untuk memasukkan uang tunai ke dalam sistem perbankan secara bertahap. *Wire Transfer* memungkinkan *transfer* dana antar rekening atau negara dengan cepat dan keamanan tinggi serta sering digunakan untuk memindahkan uang secara global tanpa menarik perhatian. *Wire transfer* merupakan salah satu jenis *funds transfer*. *Funds Transfer* (transfer dana) mencakup pemindahan dana antar akun atau lembaga keuangan, biasanya digunakan dalam kegiatan sehari-hari. Oleh karena itu, sering kali digunakan untuk mengaburkan jejak uang. *Currency Exchange* (pertukaran mata uang) melibatkan pertukaran mata uang yang dapat digunakan untuk menyamarkan asal usul uang melalui konversi mata uang yang berbeda. *Real Estate Deals* (transaksi properti) dapat digunakan untuk membeli atau menjual aset dengan nilai tinggi sehingga menyamarkan uang ilegal sebagai hasil dari investasi. *Security Transaction* (transaksi sekuritas) melibatkan perdagangan saham atau surat berharga yang dapat digunakan untuk mencampurkan uang ilegal dengan keuntungan dari investasi yang sah. Semua metode ini bertujuan untuk menyembunyikan sumber dana asli dan membuat uang yang diperoleh secara ilegal tampak sah (Esoimeme, 2020).

Analisis berbasis komponen seperti *Customer Identification* (Identifikasi Pelanggan) dan *Know Your Customer* (Kenali Pelanggan Anda) bertujuan untuk memverifikasi identitas pelanggan dan memahami profil risiko mereka guna mencegah kegiatan ilegal seperti pencucian uang. Proses *Customer Identification* memastikan bahwa informasi yang diberikan oleh pelanggan, seperti nama dan alamat telah valid dan sesuai dengan dokumen resmi. Sementara itu, *Know Your Customer* mencakup pemeriksaan latar belakang lebih mendalam untuk menilai potensi risiko dari setiap individu termasuk pemeriksaan terhadap sumber dana dan transaksi yang dilakukan. *Screening on suspicious designation* *nationalists* (pemeriksaan terhadap individu dengan penunjukan nasional yang mencurigakan) melibatkan memeriksa daftar hitam atau basis data internasional untuk mendeteksi orang-orang yang mungkin terlibat dalam kegiatan ilegal atau dianggap sebagai risiko tinggi. Dengan mengintegrasikan kedua komponen ini, lembaga keuangan dapat lebih efektif dalam mendeteksi dan mencegah pencucian uang dan kegiatan keuangan ilegal lainnya (Irwan et al., 2022).

Salah satu komponen penting dalam kerangka kerja AML adalah pelaporan transaksi mencurigakan (*Suspicious Transaction Reporting*) yang berperan krusial dalam mengidentifikasi dan menangani aktivitas keuangan yang berpotensi melibatkan pencucian uang. *Suspicious Transaction Reporting* (STR) adalah mekanisme pelaporan yang wajib dilakukan oleh lembaga keuangan ketika mereka mengidentifikasi transaksi yang dianggap mencurigakan (Shah, 2023). Transaksi ini mungkin melibatkan upaya untuk menyamarkan asal-usul dana yang ilegal atau mencerminkan pola perilaku yang tidak biasa yang menunjukkan kemungkinan adanya aktivitas pencucian uang. Proses pelaporan ini merupakan bagian penting dari strategi AML untuk mendeteksi dan mencegah pencucian uang.

Lembaga keuangan menggunakan berbagai indikator untuk mengidentifikasi transaksi mencurigakan. Beberapa contoh kriteria yang dapat memicu pelaporan STR meliputi (Irwan et al., 2022):

1. Transaksi dengan jumlah yang signifikan dan tidak sejalan dengan profil keuangan nasabah dapat menimbulkan kecurigaan. Misalnya, jika seorang nasabah yang biasanya melakukan transaksi kecil tiba-tiba men*transfer* atau menerima sejumlah besar uang tanpa penjelasan yang jelas, hal ini dapat dianggap mencurigakan.
2. Aktivitas yang menunjukkan pola transaksi yang tidak konsisten dengan kebiasaan normal nasabah, seperti peningkatan volume transaksi secara tiba-tiba dan frekuensi transaksi yang tinggi atau transaksi yang tampaknya tidak terhubung dengan kegiatan bisnis atau pribadi nasabah dapat menjadi indikasi pencucian uang.
3. Salah satu taktik yang umum digunakan oleh pencuci uang adalah memecah transaksi besar menjadi beberapa transaksi kecil untuk menghindari deteksi batas pelaporan. Hal ini bertujuan untuk menghindari batas minimum yang memicu pelaporan otomatis kepada pihak berwenang.
4. Transaksi yang melibatkan negara atau wilayah yang dikenal memiliki risiko tinggi untuk pencucian uang atau pendanaan terorisme juga bisa memicu pelaporan STR. Lembaga keuangan harus waspada terhadap transaksi yang melibatkan individu dari negara-negara tersebut.
5. Jika seorang nasabah tiba-tiba mengubah cara atau tujuan transaksinya tanpa alasan yang jelas atau jika ada perubahan signifikan dalam aktivitas rekening yang tidak dapat dijelaskan oleh nasabah, hal ini dapat mengindikasikan upaya untuk menyembunyikan sumber dana.

Di Indonesia, tindak pidana ini diatur dalam UU No. 8 Tahun 2010 tentang Pencegahan dan Pemberantasan Tindak Pidana Pencucian Uang (Denniagi, 2021). Adapun perbuatan-perbuatan yang menjadi tindak pidana pencucian uang menurut UU No. 8/2010 adalah sebagai berikut:

1. Menempatkan, men*transfer*, mengalihkan membelanjakan, membayarkan, menghibahkan, menitipkan, membawa ke luar negeri, mengubah bentuk, menukarkan dengan mata uang atau surat berharga atau perbuatan lain atas harta kekayaan yang diketahuinya atau patut diduganya merupakan hasil tindak pidana dengan tujuan menyembunyikan atau menyamarkan asal usul harta kekayaan.
2. Menyembunyikan atau menyamarkan asal usul, sumber, lokasi, peruntukan, pengalihan hak-hak, atau kepemilikan yang sebenarnya atas harta kekayaan yang diketahuinya atau patut diduganya merupakan hasil tindak pidana.
3. Menerima, menguasai penempatan, pen*transfer*an, pembayaran, hibah, sumbangan, penitipan, penukaran, atau menggunakan harta kekayaan yang diketahuinya atau patut diduganya merupakan hasil tindak pidana.

Informasi yang dikumpulkan melalui proses pemantauan harus diperiksa dan divalidasi secara mendalam pada berbagai tahap dan tingkatan. Ini berarti bahwa data yang dikumpulkan dari aktivitas pemantauan tidak hanya perlu dianalisis untuk mendeteksi pola atau anomali, tetapi juga harus diverifikasi keakuratannya di berbagai titik proses. Pada tingkat awal, informasi harus diperiksa untuk memastikan bahwa data yang masuk adalah akurat dan relevan. Selanjutnya, data yang sudah teridentifikasi sebagai mencurigakan atau penting harus melalui proses validasi lebih lanjut untuk memastikan tidak ada kesalahan atau kebingungan dalam interpretasi. Proses ini penting untuk memastikan bahwa keputusan yang diambil berdasarkan data tersebut adalah tepat dan dapat diandalkan, serta untuk menghindari kesalahan yang dapat menyebabkan kesimpulan yang salah atau tindakan yang tidak tepat.

*Financial Intelligence Units* (FIU) bertugas mengumpulkan dan menganalisis data transaksi keuangan untuk mendeteksi aktivitas yang mencurigakan, serta menyebarluaskan informasi kepada pihak berwenang untuk investigasi lebih lanjut (Sultana, 2020). *Regulatory Authorities* dibuat untuk menetapkan aturan dan kebijakan pencegahan pencucian uang, mengawasi kepatuhan lembaga keuangan terhadap peraturan, dan memberikan sanksi jika diperlukan yang berguna dalam kelancaran perekonomian (Lubis, 2023). Sementara itu, *Enforcement Agencies* fokus pada penyelidikan dan penuntutan kasus pencucian uang dan kejahatan finansial lainnya, menggunakan bukti yang dikumpulkan oleh FIU dan hasil pengawasan dari otoritas regulasi untuk memastikan tindakan hukum yang efektif (Anatolii et al., 2024). Ketiga komponen ini bekerja secara sinergis untuk mencegah dan menangani kejahatan finansial dengan cara yang terkoordinasi dan terintegrasi.

## *Machine Learning*

*Machine Learning* adalah sebuah teknik yang secara otomatis melakukan analisis terkait data dengan memanfaatkan metode perhitungan statistik yang digunakan untuk menangani data dalam skala besar dengan menciptakan model prediktif yang digunakan di berbagai bidang (Marmolejo-Ramos et al., 2022). *Machine learning* merupakan sebuah media yang dipandang sebagai aset yang dapat menghubungkan *data surge* dalam dunia data kontemporer dengan suatu permasalahan agar menghasilkan suatu solusi atau keputusan yang akurat berdasarkan data untuk menghasilkan keunggulan kompetitif terhadap proses analisis data historis (Raghavendran & Elragal, 2023).

## *Fraud Detection*

*Fraud Detection* adalah sistem pemantauan yang berfungsi untuk menentukan apakah suatu transaksi keuangan mengandung aktivitas penipuan atau tidak. *Fraud detection* telah mengalami perkembangan yang signifikan dalam prosesnya, hal ini untuk mengimbangi penipu yang telah menggunakan metode canggih untuk melakukan penipuan. Berdasarkan permasalahan tersebut lembaga keuangan mencari solusi terbaik dalam *fraud detection,* dimana *machine learning* merupakan inovasi teknologi baru yang digunakan oleh lembaga keuangan dalam menghadapi tantangan *fraud detection* yang semakin berkembang (L. K. Vishwamitra, 2024).

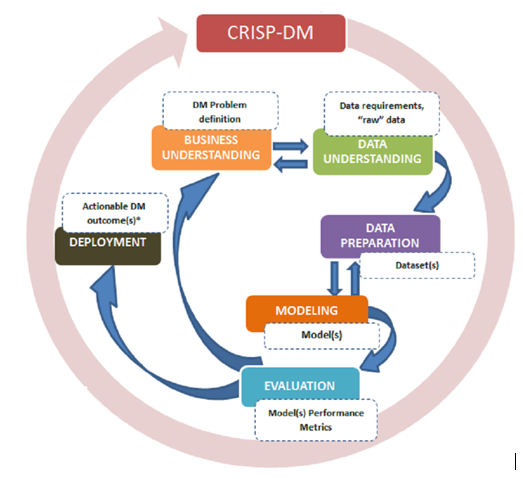
## Data Sintetis

Data sintetis adalah suatu data buatan yang dihasilkan melalui proses simulasi kecerdasan buatan untuk menciptakan data baru yang dirancang agar memiliki sifat dan pola yang mirip dengan data asli (Lopez-Rojas & Barneaud, 2019). PaySim merupakan salah satu contoh data sintetis yang dihasilkan dengan proses simulasi kecerdasan buatan melalui simulator keuangan canggih yang dirancang untuk mensimulasikan transaksi keuangan mobile berdasarkan data transaksi keuangan asli.

*Dataset* PaySim dikembangkan untuk mengatasi permasalahan akibat sulitnya akses terhadap data transaksi keuangan di wilayah Afrika karena dibatasi oleh regulasi. Disisi lain, data transaksi keuangan sangat dibutuhkan oleh para peneliti untuk mengembagkan suatu model yang dapat melakukan *fraud detection* atau mendeteksi transaksi keuangan yang mencurigakan, sehingga munculah sebuah inovasi yang dikembangkan oleh para peneliti dari Blekinge Institute of Technology (Swedia) yang bekerja sama dengan Norwegian University of Science and Technology (Norwegia) untuk menghasilkan data sintetis. Proses penelitian ini menggunakan data asli dari suatu layanan transaksi keuangan mobile salah satu perusahaan telekomunikasi di negara Afrika untuk membangun model simulasi kecerdasan buatan yang dapat menciptakan data sintetis yang memiliki kemiripan dengan data asli untuk kepentingan penelitian dalam bidang *fraud detection* dengan *machine learning* (Lopez-Rojas et al., 2016)​.

## CRIPS-DM *(Cross-Industry Standard Process for Data Mining)*

CRIPS-DM adalah metode pengembangan standar dan komprehensif yang digunakan dalam proses *data mining* (penambangan data) yang terdiri dari 6 tahap utama: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment* (Elkabalawy et al., 2024).

****

Gambar 2.3 Tahapan Metode CRISP-DM (Plotnikova et al., 2020)

Metode CRIPS-DM banyak digunakan dalam proyek analisis data karena memiliki kerangka kerja yang terstruktur dan fleksibel. Setiap prosesnya bisa dilakukan secara iteratif (Niakšu, 2015), sehingga menjadikannya sangat cocok untuk dalam proyek berbasis data untuk mendapatkan hasil yang maksimal. ​

### *Business Understanding*

*Business Understanding* adalah memahami secara mendalam ruang lingkup, tujuan, dan batasan yang ada dalam konteks bisnis (Elkabalawy et al., 2024). Tahap ini berfokus pada pemahaman tujuan proyek dari sudut pandang bisnis, mengubah masalah bisnis menjadi masalah *data science* serta merancang rencana awal untuk mencapai tujuan tersebut (Christoforaki & Beyan, 2024).

### *Data Understanding*

*Data Understanding* adalah fase mengeksplor data (*Exploratory Data Analysis*) dengan melakukan identifikasi masalah potensial untuk mendapatkan wawasan awal dan hipotesis hubungan antar data dengan tujuan memperoleh informasi rinci. Tahapan EDA sering digunakan untuk menemukan pola melalui visualisasi dan analisis statistik dengan mengidentifikasi trend dan korelasi yang muncul. Penilaian kualitas awal data sangat penting untuk memastikan bahwa data tersebut andal dan sesuai untuk membangun model prediksi yang akurat (Elkabalawy et al., 2024).

### *Data Preparation atau Data Preprocessing*

*Data preprocessing* adalah tahap krusial dalam tugas klasifikasi data yang melibatkan serangkaian langkah untuk meningkatkan kualitas data sehingga algoritma klasifikasi dapat berfungsi dengan lebih efisien dan akurat. Fase ini meliputi transformasi data mentah ke dalam bentuk yang lebih mudah dipahami oleh algoritma klasifikasi (Pimpa & Eiamkanitchat, 2024). *Preprocessing* mencakup langkah-langkah seperti pengumpulan, deskripsi, eksplorasi, dan verifikasi kualitas data untuk mengoptimalkan pengelompokan fitur. *Preprocessing* yang baik dapat memperbaiki akurasi algoritma klasifikasi dan mengurangi kesalahan hasil klasifikasi.

Kesemua proses *framework preprocessing* tersebut akan diproses melalui tahapan:

Gambar 2.4 Proses *Preprocessing*

* 1. *Load Data:*

Tahap ini adalah melakukan pengambilan atau pemuatan data dari sumber penyimpanan, seperti *database*, file CSV, atau API, ke dalam lingkungan analisis.

* 1. *Inspect Data:*

*Inspect data* berarti memeriksa dan menganalisis data untuk memahami strukturnya, kualitasnya, dan pola-pola yang ada sebelum melakukan analisis lebih lanjut atau membangun model.

* 1. *Data Cleansing*

*Data Cleansing* adalah proses yang penting dalam memastikan bahwa data siap digunakan untuk analisis. Proses ini melibatkan berbagai langkah untuk memperbaiki dan menyusun data agar mencapai kualitas yang optimal (Desai et al., 2024). Tujuan utama dari *data cleansing* adalah untuk memastikan data yang digunakan akurat, konsisten, lengkap, dan dapat diandalkan. Ini mencakup menghapus entri duplikat, menangani nilai yang hilang dengan cara yang tepat, serta memperbaiki nilai yang tidak valid. Selain itu, *data cleansing* juga melibatkan penyeragaman format data dan mendeteksi serta menangani *outlier* yang mungkin tidak representatif. Dengan data yang bersih dan terorganisir dengan baik, analisis menjadi lebih akurat dan model *machine learning* dapat memberikan hasil yang lebih efektif.

1. *Distribution*

*Distribution* dalam statistik dan analisis data mengacu pada cara data tersebar di seluruh rentang nilai. Ini menjelaskan pola penyebaran data dan frekuensi kemunculan nilai-nilai tertentu. Pemahaman pada distribusi bisa membantu dalam mengetahui bagaimana data terdistribusi dan hubungan antara berbagai nilai dalam *dataset*. Distribusi memberikan gambaran tentang bagaimana data terorganisir dan seberapa sering setiap nilai muncul, sehingga dapat membantu dalam menganalisis dan menarik kesimpulan dari data tersebut (Tufail et al., 2023).

* 1. *Enrich Data:*

*Enrich data* berarti menambahkan informasi tambahan atau detail baru ke dataset untuk meningkatkan kualitas dan kedalaman data yang ada, sehingga menghasilkan wawasan yang lebih lengkap dan akurat.

* 1. *Scaling*

*Scaling* adalah proses yang dilakukan untuk mengubah rentang nilai data ke dalam skala tertentu, seperti dari 0 hingga 1 atau -1 hingga 1. Tujuan dari *scaling* adalah agar semua fitur atau variabel dalam dataset memiliki pengaruh yang setara saat digunakan dalam model *machine learning.* Dengan melakukan *scaling,* kita memastikan bahwa tidak ada fitur yang dominan hanya karena skala nilainya yang lebih besar atau lebih kecil dibandingkan fitur lainnya. Ini membantu model untuk belajar dari data secara lebih efektif dan menghindari bias yang mungkin timbul akibat perbedaan skala antar fitur (Lokanan, 2023).

* 1. *Encoding*

*Encoding* adalah proses mengubah data yang bersifat kategorikal atau berbasis teks menjadi format numerik agar bisa diproses oleh model *machine learning*. Karena banyak algoritma *machine learning* hanya bisa bekerja dengan angka, *encoding* mengonversi kategori atau label yang berbentuk teks menjadi angka, memungkinkan model untuk memahami dan menganalisis data tersebut (Zhu et al., 2024).

1. *Standardization*

*Standardization* adalah metode untuk mengubah fitur data sehingga distribusinya memiliki rata-rata 0 dan deviasi standar 1. Ini dilakukan dengan mengurangi rata-rata dari setiap nilai data dan membaginya dengan deviasi standar. Proses ini membuat data memiliki skala yang konsisten, memudahkan model *machine learning* untuk memproses data dengan lebih baik dan meningkatkan akurasi analisis serta prediksi. Dengan *standardization*, data dari berbagai fitur menjadi sebanding sehingga tidak ada fitur yang mendominasi hanya karena rentang nilainya yang berbeda (Ortiz et al., 2024).

* 1. *Identify Feature:*

*Identify feature* berarti menentukan dan memilih fitur atau variabel penting dari *dataset* yang akan digunakan dalam analisis atau model *machine learning* untuk memprediksi hasil.

* 1. *Resampling*

*Resampling* adalah metode yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam sebuah *dataset*. Ketidakseimbangan kelas terjadi ketika jumlah contoh dari satu kelas jauh lebih banyak atau lebih sedikit dibandingkan dengan kelas lainnya. *Resampling* membantu menyeimbangkan distribusi kelas dengan cara menambah atau mengurangi jumlah contoh dalam kelas-kelas tersebut. Ada dua pendekatan utama dalam *resampling*: *oversampling* yang menambah jumlah data pada kelas yang kurang representatif, dan *undersampling* yang mengurangi jumlah data pada kelas yang lebih dominan. Teknik ini memastikan bahwa model *machine learning* dapat belajar dari data dengan cara yang lebih proporsional dan akurat (Gurcan & Soylu, 2024). Metode *undersampling* cocok untuk ketidakseimbangan data yang ekstrim dan sumber daya komputasi yang terbatas (Muraru et al., 2024; Yang et al., 2024).

1. *Correlation Matrix*

*Correlation Matrix* adalah tabel yang menunjukkan seberapa besar hubungan atau keterkaitan antara setiap pasangan fitur dalam sebuah *dataset*. Tabel ini mengukur sejauh mana satu fitur berubah seiring dengan perubahan fitur lainnya. Nilai dalam matrik ini mencerminkan kekuatan dan arah hubungan tersebut, dengan nilai antara -1 dan 1, di mana -1 menunjukkan hubungan negatif yang sempurna, 1 menunjukkan hubungan positif yang sempurna, dan 0 menunjukkan tidak adanya hubungan. *Matrix* ini membantu dalam memahami hubungan antar fitur dan dapat digunakan untuk analisis lebih lanjut atau pemilihan fitur dalam model *machine learning* (Prakash et al., 2021)*.*

1. *Feature Engineering*

*Feature Engineering* adalah proses mengembangkan dan memodifikasi fitur dalam *dataset* untuk meningkatkan kinerja model *machine learning.* Ini melibatkan penciptaan fitur baru yang relevan atau penyempurnaan fitur yang sudah ada dengan tujuan membuat data lebih informatif dan sesuai dengan kebutuhan model. *Feature engineering* bisa meliputi berbagai teknik, seperti menggabungkan beberapa fitur menjadi satu, mengubah skala atau format data, dan mengekstrak informasi baru dari data yang ada. Dengan melakukan *feature engineering* yang efektif, model dapat mempelajari pola yang lebih baik dan memberikan hasil yang lebih akurat (Sutoyo et al., 2023).

### *Data Modeling*

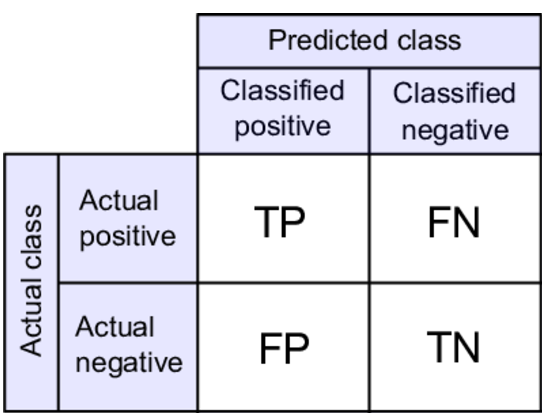
*Data modeling* adalah proses memilih dan menerapkan berbagai teknik pemodelan data dengan parameter yang sudah disesuaikan agar menghasilkan akurasi yang optimal. Setiap jenis masalah *data mining* biasanya bisa diselesaikan menggunakan beberapa metode pemodelan. Namun, penerapan teknik-teknik tersebut sering kali memerlukan pemenuhan kondisi tertentu pada data. Akibatnya, mungkin perlu kembali ke tahap pemodelan saat tahap-tahap lain dari proses berlangsung agar hasil yang diinginkan dapat dicapai dengan lebih baik (Nafisah Nurul Hakim, 2020).

Selanjutnya tahapan *data modeling* akan diproses melalui skema berikut:

Gambar 2.5 Proses *Data Modeling*

1. *Select Algorithms* adalah tahap dalam proses pengembangan *model machine learning* di mana dilakukan pemilihan algoritma yang akan digunakan untuk membangun model prediktif (Lokanan, 2023). Pemilihan algoritma bergantung pada jenis masalah yang ingin diselesaikan, seperti klasifikasi atau regresi, serta karakteristik data yang digunakan. Berikut adalah penjelasan mengenai tiga algoritma populer yang sering digunakan dalam masalah klasifikasi pada *fraud detection* (Afriyie et al., 2023)(Alwadain et al., 2023):
   1. *Decision Tree* adalah sebuah algoritma *machine learning* yang digunakan untuk membuat model prediksi berdasarkan struktur pohon. Model ini membagi data ke dalam cabang-cabang berdasarkan fitur-fitur tertentu untuk menghasilkan keputusan atau prediksi. Dalam *decision tree*, setiap simpul internal mewakili tes atau keputusan terhadap fitur tertentu, setiap cabang mewakili hasil dari tes tersebut, dan setiap daun mewakili hasil akhir atau kelas keputusan. *Decision tree* bekerja dengan cara membagi data menjadi *subset* yang lebih kecil dengan memilih fitur yang paling informatif, sehingga model ini dapat membuat keputusan yang jelas dan mudah diinterpretasikan. Algoritma ini cocok untuk data yang bersifat diskrit maupun kontinu dan sering digunakan dalam masalah klasifikasi dan regresi (Ostonov & Moshkov, 2024).
2. *Random Forest* adalah algoritma *machine learning* yang merupakan kumpulan dari banyak pohon keputusan (*Decision Trees*) yang digabungkan *(ensemble learning)* dan bekerja secara paralel *(bagging)* untuk meningkatkan akurasi dan stabilitas model prediksi*.* Dalam *random forest,* beberapa pohon keputusan dibangun secara independen dengan menggunakan *subset* acak dari data pelatihan dan *subset* acak dari fitur pada setiap pembagian pohon. Hasil prediksi dari setiap pohon kemudian digabungkan, biasanya dengan cara *voting* untuk klasifikasi atau rata-rata untuk regresi untuk menghasilkan prediksi akhir. Pendekatan ini membantu mengurangi risiko *overfitting* yang sering terjadi pada model pohon keputusan *(Decision Tree)* individu dan meningkatkan kinerja model secara keseluruhan. *Random forest* dikenal karena kemampuannya yang baik dalam menangani data yang kompleks dan memiliki banyak fitur (Chen et al., 2024).
3. *XGBoost (Extreme Gradient Boosting)* adalah algoritma *machine learning* yang sangat efisien dan populer untuk masalah klasifikasi dan regresi. *XGBoost* merupakan implementasi dari teknik *boosting*, dimana beberapa model *machine learning*, khususnya pohon keputusan digabungkan secara bertahap *(boosting)* untuk meningkatkan performa model. Proses pelatihan dengan *boosting* ini melibatkan pelatihan model berturut-turut, di mana setiap model baru berusaha memperbaiki kesalahan yang dibuat oleh model sebelumnya. *XGBoost* menonjol karena kecepatan dan efisiensinya dalam hal waktu pelatihan, serta kemampuannya untuk menangani data yang tidak seimbang, fitur yang hilang, dan outlier. Selain itu, *XGBoost* dilengkapi dengan berbagai fitur regulasi untuk mencegah *overfitting*, seperti pengaturan pelatihan dan *pruning* pohon, serta mendukung berbagai teknik optimasi untuk mempercepat proses pelatihan. Karena kemampuannya yang unggul dalam menangani berbagai jenis data dan masalah, *XGBoost* sering digunakan dalam kompetisi *data science* dan aplikasi dunia nyata (Zhang et al., 2024).
4. *Train & Optimize Models* adalah tahap dalam pengembangan *model machine learning* di mana model dilatih menggunakan data pelatihan dan kemudian dioptimalkan untuk meningkatkan kinerjanya (de Pedro-Carracedo et al., 2023). Proses ini dimulai dengan memberikan data pelatihan kepada model yang telah dipilih, seperti *Decision Tree, Random Forest,* atau *XGBoost* untuk membangun hubungan antara fitur *input* dan target *output*. Setelah model dilatih, langkah optimasi dilakukan dengan menyesuaikan berbagai *hyperparameter*, seperti kedalaman pohon pada *Decision Tree,* jumlah pohon pada *Random Forest*, atau tingkat pembelajaran (*learning rate*) pada *XGBoost* untuk meningkatkan akurasi dan mengurangi *overfitting*. Tujuannya adalah untuk menghasilkan model yang tidak hanya akurat tetapi juga dapat diandalkan ketika digunakan pada data baru yang tidak terlihat sebelumnya. Salah satu teknik *hyperparameter* adalah *Grid Search CV*, yaitu metode optimasi *hyperparameter tuning* dalam *machine learning* yang digunakan untuk menemukan kombinasi parameter terbaik bagi suatu model. Prosesnya bekerja dengan melakukan pencarian pada ruang parameter yang telah ditentukan sebelumnya. Setiap kombinasi parameter diuji menggunakan teknik *cross-validation* (CV) untuk menilai performa model dengan metrik tertentu seperti akurasi, F1-score, atau AUC-ROC. *Grid Search CV* memastikan bahwa model bekerja dengan pengaturan parameter yang optimal (Asif et al., 2023).
   1. *Validate Performance* adalah tahap dalam pengembangan *model machine learning* di mana kinerja model yang telah dilatih diuji dan dievaluasi menggunakan data yang belum pernah dilihat oleh model sebelumnya. Tujuannya adalah untuk menilai seberapa baik model dapat memprediksi hasil yang benar pada data baru yang mencerminkan kemampuan generalisasi model (Kalboussi et al., 2024).

### *Data Evaluation*

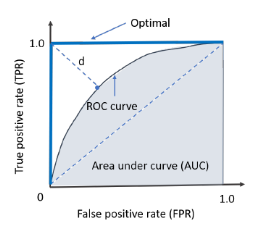


Gambar 2.6 *Confusion Matrik* (Richter-Laskowska et al., 2022)

*Confusion matrix* adalah salah satu metode yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan mengukur seberapa baik model tersebut memprediksi kelas data. *Matrix* ini terdiri dari empat komponen utama: *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) yang membantu menilai keakuratan prediksi model. Berdasarkan komponen-komponen tersebut, berbagai metrik seperti *accuracy, precision, f1-Score, specifity* dan *recall (sensitivity)* dapat dihitung untuk menilai performa model secara keseluruhan, sehingga memberikan gambaran yang jelas tentang seberapa efektif model dalam mengklasifikasikan data (Handayani & Charis Fauzan, 2024).

Gambar 2.7 Rumus Evaluasi (Hasan & Wagler, 2024)

1. *Precision* mengukur sejauh mana hasil prediksi yang diberikan oleh model saling konsisten atau mendekati satu sama lain, mencerminkan tingkat konsistensi model dalam memprediksi data.
2. *Accuracy* mengukur seberapa dekat hasil prediksi model dengan nilai sebenarnya, menunjukkan sejauh mana prediksi model mendekati nilai referensi atau kenyataan.
3. *Recall (sensitivity)* menilai kemampuan model dalam mendeteksi sampel positif yang sebenarnya, atau seberapa baik model mengidentifikasi seluruh data yang termasuk dalam kelas positif.
4. *Specificity* mengukur kemampuan model untuk mengidentifikasi dengan benar sampel negatif atau non-positif.
5. *F1 Score* adalah metrik yang digunakan untuk mengevaluasi kinerja model klasifikasi, terutama dalam kasus di mana distribusi kelas tidak merata, dengan menggabungkan *precision* dan *recall* menjadi satu metrik harmonis.



Gambar 2.8 Kurva ROC AUC (Liu et al., 2020)

Dan ROC AUC *(Receiver Operating Characteristic - Area Under the Curve)* adalah ukuran kinerja model yang menunjukkan seberapa baik model memisahkan kelas positif dan negatif dengan menghitung luas di bawah kurva yang memplot hubungan antara tingkat false positif dan sensitivitas *(recall)* (de Pedro-Carracedo et al., 2023).

Serta *Cross-validation* adalah teknik evaluasi yang digunakan untuk menilai kemampuan model prediktif dalam menggeneralisasi data baru atau data dunia nyata. Proses ini membagi dataset menjadi beberapa *subset* *(folds),* di mana model dilatih pada sebagian data *(training set)* dan diuji pada bagian lainnya *(validation set)* secara bergantian. Tujuannya adalah untuk mengukur performa model secara konsisten dan menghindari masalah *overfitting*, yaitu kondisi ketika model sangat baik dalam mempelajari data pelatihan namun gagal memprediksi data baru karena terlalu kompleks dan menangkap *noise* yang tidak relevan. *Cross-validation* juga mencegah *underfitting*, yaitu kondisi ketika model terlalu sederhana sehingga gagal menangkap pola dalam data. Teknik ini membantu mendapatkan estimasi performa model yang lebih akurat, memastikan keseimbangan antara bias dan *varians* model (Mbiva & Correa, 2024).

### *Deployment*

Pada tahap ini, model yang telah dibuat diterapkan ke dalam media yang memungkinkan pengguna memahami cara kerjanya. Proses *deployment* dapat dilakukan dalam berbagai bentuk, mulai dari penyusunan laporan hingga implementasi sistem diserver produksi atau perangkat lokal yang memungkinkan interaksi langsung antara pengguna dan sistem. Tujuannya adalah agar pengguna dapat menggunakan dan merasakan manfaat dari model tersebut secara praktis (Nafisah Nurul Hakim, 2020).

## Tools

### Python

Python yang sangat populer saat ini pertama kali dikembangkan oleh Guido van Rossum di Stichting Mathematisch Centrum (CWI) di Amsterdam pada tahun 1991. Bahasa pemrograman ini terinspirasi dari ABC dan berbeda dari bahasa lainnya karena pengembangannya melibatkan banyak programmer, peneliti, dan pengguna dari berbagai disiplin tidak terbatas pada bidang IT, karena Python bersifat *open source*. Python menggunakan *interpreter* untuk menjalankan kodenya secara langsung dan mendukung berbagai *platform* seperti Windows dan Linux. Bahasa ini menggabungkan berbagai paradigma pemrograman, termasuk prosedural seperti C, berorientasi objek seperti Java, dan fungsional seperti Lisp, yang membuatnya fleksibel untuk berbagai jenis proyek (Rahman et al., 2023).

### Numpy

NumPy *(Numerical Python)* adalah salah satu pustaka Python yang banyak digunakan untuk melakukan komputasi numerik. Pustaka ini menyediakan dukungan untuk bekerja dengan array serta memungkinkan operasi array dilakukan secara efisien, selain menyediakan berbagai fungsi matematika yang dioptimalkan untuk kinerja yang cepat. NumPy juga menjadi fondasi bagi banyak pustaka lain yang berfokus pada ilmu pengetahuan dan komputasi numerik dalam ekosistem Python (Dasawarsa et al., 2023).

### Pandas

Pandas adalah pustaka *open source* berlisensi BSD yang menyediakan struktur khusus untuk analisis data dalam Python. Pustaka ini digunakan untuk membersihkan data mentah sehingga siap dianalisis, dan mampu menangani tugas-tugas dasar seperti penyelarasan dan penggabungan berbagai sumber data. Salah satu fitur utamanya adalah kemampuannya untuk mengelola data yang hilang. Struktur utama dalam Pandas adalah *dataframe* yang merupakan kumpulan kolom dengan nama dan tipe data tertentu, mirip dengan tabel dalam basis data di mana setiap baris merepresentasikan satu entitas atau contoh (Putra & Toba, 2020).

### Matplotlib

Matplotlib adalah pustaka Python yang sangat fleksibel dan kuat untuk membuat visualisasi data dalam berbagai bentuk, termasuk *scatter plot, line chart, bar chart,* dan *pie chart,* serta grafik yang lebih kompleks. Matplotlib memungkinkan pengguna melakukan kustomisasi mendalam pada grafik, seperti menambahkan judul, label sumbu, legenda, hingga mengatur skala, warna, dan gaya garis. Kemampuan ini menjadikan Matplotlib sangat cocok untuk menyajikan data secara visual dalam berbagai konteks analisis, dari eksplorasi data awal hingga presentasi yang lebih formal dan kompleks (Maulida Surbakti et al., 2024).

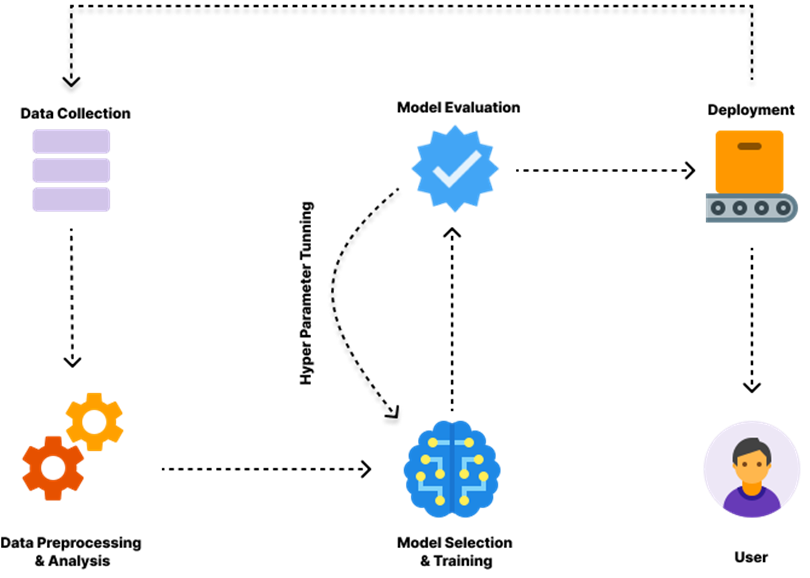
### Scikit Learn

Modul utama yang digunakan dalam fase pemodelan ini adalah scikit-learn, sebuah pustaka yang memungkinkan pengembangan algoritma *machine learning* berbasis Python. Scikit-learn menyediakan berbagai alat dan metode untuk pengolahan data, pemodelan prediktif, serta evaluasi model. Pustaka ini dirancang agar mudah digunakan, mendukung banyak teknik *machine learning* populer, seperti klasifikasi, regresi, dan *clustering*. Scikit-learn juga didistribusikan di bawah lisensi BSD 3-Clause yang memungkinkan penggunaannya secara bebas dalam berbagai proyek, termasuk yang bersifat komersial (Nafisah Nurul Hakim, 2020).

# METODE

## Rencana Penelitian

Rencana penelitian adalah panduan terstruktur yang menguraikan langkah-langkah yang akan diambil untuk menjalankan sebuah penelitian. Rencana ini menjelaskan tujuan penelitian, metode yang digunakan, serta tahapan yang diperlukan untuk mencapai hasil yang diinginkan. Rencana penelitian berfungsi sebagai peta jalan yang membantu peneliti agar proses penelitian berjalan sistematis, efisien, dan fokus pada tujuan utama.



Gambar 3.1 Proses Pengembangan M*achine Learning*

Secara garis besar, pengembangan model *machine learning* akan melewati *data collection* untuk mengumpulkan data yang akan digunakan*, data preprocessing* dan *analysis* untuk memahami atau mengeksplor data dan membersihkan permasalahan data, melakukan pelatihan model berdasarkan algoritma yang telah dipilih, mengevaluasi kinerja model untuk mempertimbangkan kemampuan model agar bisa diterapkan ke dalam sistem, dan *deployment* kedalam sistem atau hanya dalam penjelasan hasil dari pembuatan model. Berikut adalah rencana penelitian yang akan dilakukan dalam pengembangan model *machine learning* dengan menggunakan metode CRISP-DM:

1. *Business Understanding*

Industri keuangan menghadapi tantangan serius dalam *fraud detection*, termasuk pencucian uang yang semakin kompleks. Hal ini diperkuat dengan Indonesia menempati peringkat 85 dalam Index Basel AML 2023. Rendahnya pelaporan terhadap transaksi keuangan yang mencurigakan dan keterbatasan akses data transaksi keuangan yang bersifat rahasia menjadi permasalahan utama dalam pengembangan model *fraud detection* berbasis *machine learning*. Penelitian ini bertujuan untuk mengatasi keterbatasan tersebut dengan menggunakan *dataset* sintetis PaySim yang mensimulasikan transaksi keuangan *mobile* untuk menciptakan model *machine learning*. Model ini dirancang untuk mendeteksi transaksi keuangan yang mencurigakan secara cepat dan akurat, sehingga dapat membantu lembaga keuangan dalam meningkatkan keamanan sistem, meminimalkan kerugian finansial, serta mendukung kebijakan *anti-fraud* yang lebih efektif. Keberhasilan model akan diukur melalui metrik evaluasi yang diterapkan pada *machine learning.*

1. *Data Understanding*

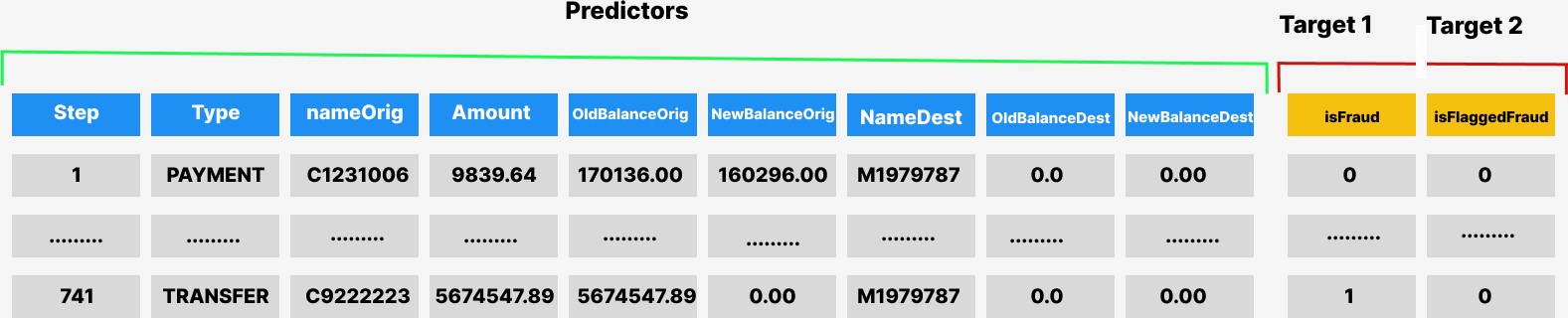
*Exploratory Data Analysis* (EDA) meliputi analisis data sintetis untuk memahami struktur data, pola transaksi, dan karakteristik umum dari transaksi *fraud*. EDA dapat membantu mengenali distribusi data, hubungan antar variabel, mengidentifikasi nilai yang hilang, *outlier*, atau inkonsistensi, dan memungkinkan analisis korelasi untuk menentukan fitur yang relevan terhadap target.

1. *Data Preparation atau Preprocessing*

Setelah memahami struktur data, pola transaksi, dan karakteristik umum dari transaksi *fraud* maka dilakukanlah penanganan permasalahan yang telah diidentifikasi sebelumnya dengan membersihkan data, menangani nilai yang hilang atau outlier, *handling imbalance data* dan melakukan *feature engineering* untuk menghasilkan fitur yang relevan bagi model. Data juga akan dibagi menjadi data latih *(train)* dan data uji *(test).*

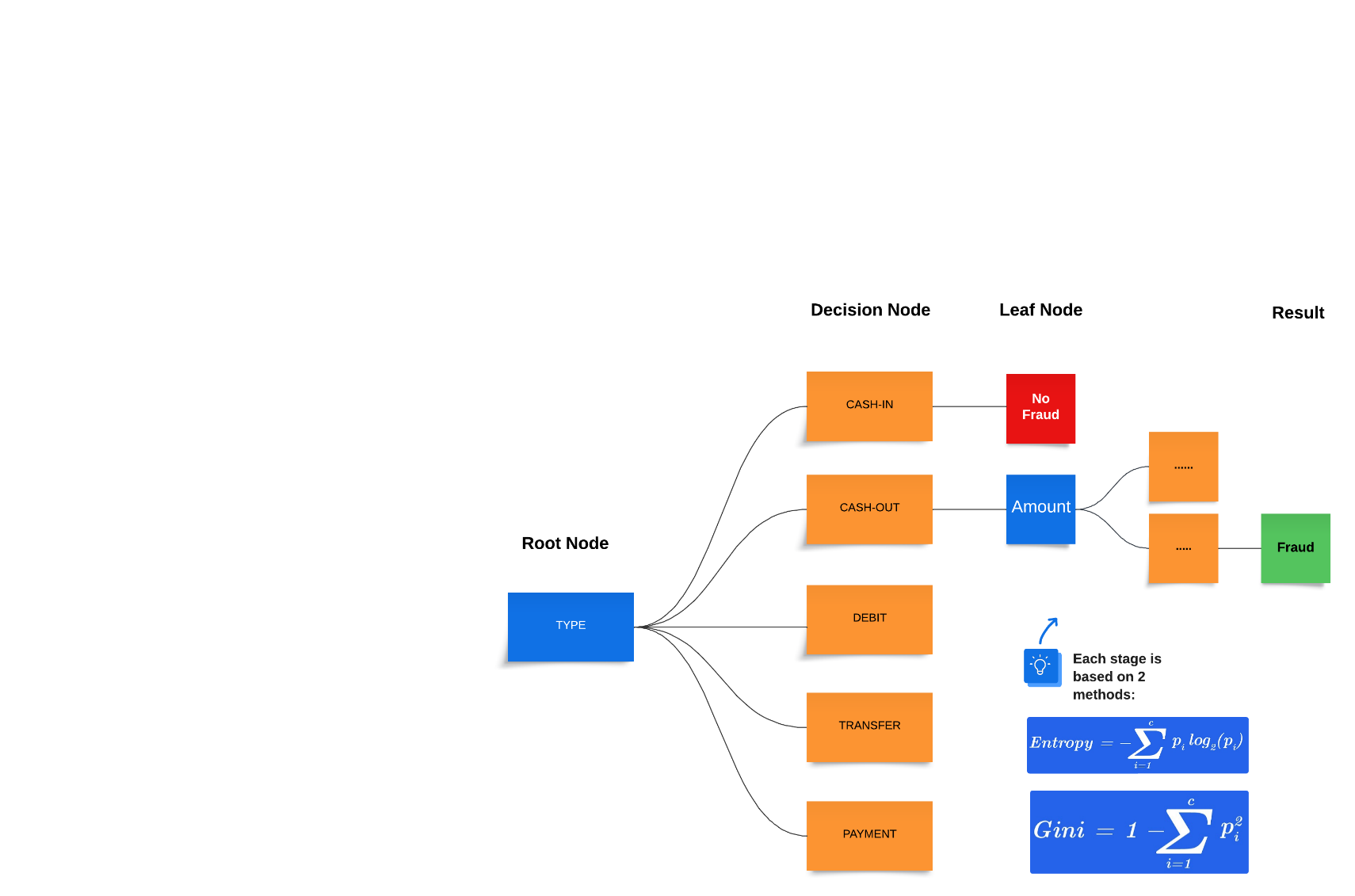
1. *Data Modeling*

Membangun beberapa model *machine learning* dan melatihnya pada data latih. Algoritma yang digunakan seperti *Decision Tree, Random Forest,* dan *XGBoost* akan dibandingkan untuk memilih model terbaik. Adapun gambaran cara kerja dari algoritma *Decision Tree* adalah sebagai berikut:



Gambar 3.2 Pembagian Dataset

Proses diawali dengan mengimpor seluruh dataset yang akan digunakan untuk melatih model. *Dataset* ini berisi berbagai atribut atau karakteristik yang digunakan sebagai input atau *predictors*, serta label atau target yang menjadi hasil yang ingin diprediksi. *Decision Tree* akan membagi dataset ini secara bertahap berdasarkan atribut-atribut yang paling relevan untuk membentuk struktur pohon keputusan.



Gambar 3.3 Cara kerja *Decision Tree*

Pada setiap node dalam *Decision Tree*, algoritma akan memilih fitur yang paling relevan untuk memisahkan data ke dalam *subset*-*subset* yang lebih kecil. Pemilihan fitur ini didasarkan pada kriteria tertentu seperti *Gini Impurity* dan *Entropy* yang digunakan untuk mengukur seberapa baik fitur tersebut memisahkan data. Dalam *library* Scikit-Learn versi algoritma *decision tree* yang digunakan adalah versi CART *(Classification and Regression Tree)*, dimana kriteria pemilihan fitur *Gini Impurity* digunakan sebagai perhitungan matematis dalam kinerja model *decision tree*. Tujuan dari proses ini adalah untuk memaksimalkan pemisahan data yang bersih berdasarkan kelas atau nilai target.

Setelah fitur yang paling relevan dipilih, data akan dibagi berdasarkan nilai fitur tersebut. Proses ini diulang secara terus-menerus pada setiap subset pohon keputusan yang dihasilkan hingga kondisi berhenti yang ditetapkan telah terpenuhi. Jika kondisi sudah terpenuhi, node terakhir pada pohon *(leaf node)* akan memberikan hasil akhir dari prediksi atau keputusan. Dimana kelas yang paling sering muncul dalam data yang menjadi hasil prediksi. Adapun gambarannya sebagai berikut:

1. Hitung *Gini Index* awal (*Gini* sebelum splitting)

Misalkan data yang digunakan terdapat 100 data dengan 2 kelas, yaitu *fraud* dan *non fraud* dengan masing-masing kelas 60 data dan 40 data.

1. Probabilitas tiap kelas:
2. Hitung Gini Index

Impuritas dataset awal adalah 0,48

1. Hitung *Gini Index* untuk setiap Fitur (*Splitting*)

Splitting dilakukan pada semua fitur mulai dari *step, type, amount, nameOrig, oldbalanceOrig, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFraud* dengan nilai pemisah atau *treshold* tergantung pada tipe data (kategorikal atau numerik) dan untuk *subset* yang terbentuk setelah pembagian dilakukan perhitungan. Misalkan fitur *type* dibagi dalam 2 *subset*, yaitu subset kiri dengan 70 data yang terdiri dari 50 kelas *fraud* dan 20 kelas *non fraud* dan *subset* kanan dengan 30 data yang terdiri dari 10 kelas *fraud* dan 20 kelas *non fraud.*

* + 1. Hitung *Gini Index* *Subset* Kiri:
    2. Hitung *Gini Index* *Subset* Kanan:

67

* + 1. Hitung *Gini Index split* untuk fitur *type*

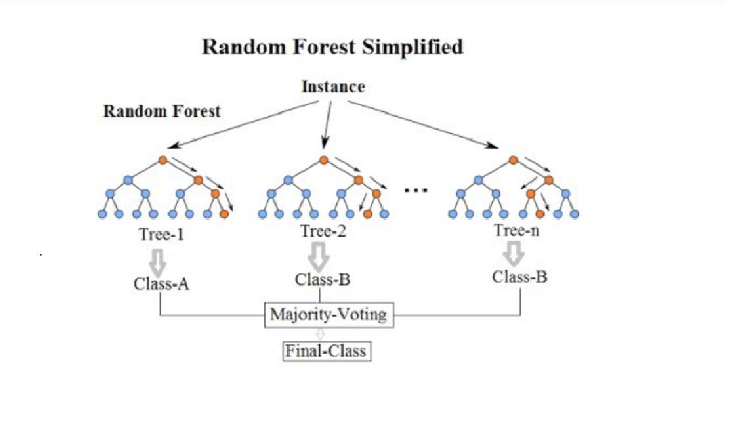
1. Ulangi Perhitungan untuk semua fitur

Lakukan langkah serupa pada fitur *step, amount, nameOrig, oldbalanceOrig, newbalanceOrig, nameDest, oldbalanceDest,* dan *newbalanceDest*, lalu *Gini Index* terkecil untuk dipilih sebagai *root node.*

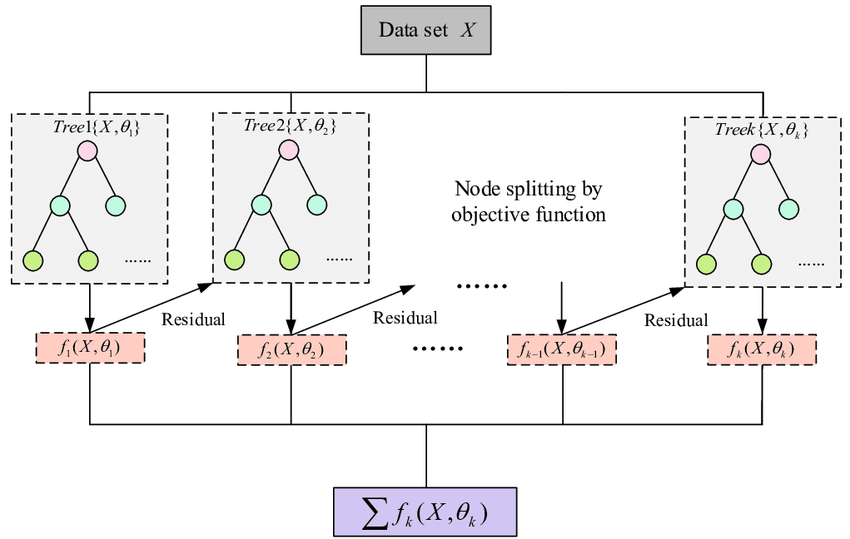
1. Pembentukan node *subtree*

Setelah fitur type dipilih sebagai *root node*, *dataset* dibagi menjadi dua *subset*, yaitu *subset* kiri dan *subset* kanan. Proses ini diulang untuk setiap *subset* hingga salah satu dari dua kondisi berikut terpenuhi:

* + 1. Seluruh data dalam *subset* berasal dari satu kelas (impuritas = 0)
    2. Tidak ada lagi fitur yang dapat digunakan untuk membagi (membentuk *leaf node*).



Gambar 3.4 Cara Kerja *Random Forest* (Azhari et al., 2019)



Gambar 3.5 Cara Kerja *XGBoost* (Guo et al., 2020)

Secara keseluruhan konsep algoritma *Random Forest* dan *XGBoost* adalah berbasis *ensemble learning*, yaitu menggabungkan beberapa model dasar *Decision Tree* untuk memaksimalkan hasil prediksi. Namun perbedaan terletak pada pendekatan yang digunakan untuk mencapai hasil akhir, jika *Random Forest* menggunakan pendekatan berdasarkan *bagging*, yaitu membuat banyak *Decision Tree* secara independen, lalu menggabungkan hasilnya menggunakan *voting* untuk memilih hasil mayoritas setiap pohon sebagai hasil akhir, sedangkan *XGBoost* menggunakan pendekatan berdasarkan *boosting*, yaitu membangun pohon keputusan secara berurutan, dimana setiap pohon mencoba memperbaiki kesalahan prediksi *(residual)* dari pohon sebelumnya, langkah ini diulangi beberapa kali hingga *error* minimum.

1. *Data Evaluation*

Mengevaluasi model menggunakan data uji dengan metrik yang ditentukan. Untuk melakukan evaluasi tersebut, penelitian ini menggunakan metrik evaluasi seperti *Accuracy, Precision, Recall, F1-Score, croos-validation* dan *ROC-AUC*. Jika hasil evaluasi tidak memadai, dilakukan iterasi ulang dengan penyesuaian parameter atau teknik lain.

1. *Deployment*

Setelah model terbaik terpilih, model akan disiapkan untuk diimplementasikan dalam sistem *fraud detection* atau dalam bentuk laporan hasil kinerja model.

## Objek Penelitian

Objek penelitian dalam studi ini adalah *fraud detection* dengan menggunakan data sintetis dari transaksi keuangan *mobile* yang dihasilkan melalui simulator PaySim. Data ini meniru pola transaksi keuangan asli yang memungkinkan peneliti untuk melakukan *fraud detection* tanpa harus mengakses data transaksi keuangan asli yang bersifat rahasia. Data sintetis ini dihasilkan berdasarkan sampel data transaksi asli dari layanan telekomunikasi yang memiliki data keuangan *mobile* yang beroperasi di Afrika. Simulator PaySim mensimulasikan perilaku pengguna semirip mungkin dengan data asli sehingga data yang dihasilkan memiliki karakteristik yang mirip dengan data asli.

## Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini dilakukan dengan menggunakan data sekunder hasil simulasi yang dikembangkan dalam penelitian sebelumnya oleh Lopez-Rojas, Edgar Alonsom Elmir, Ahmad Axelsson, dan Stefan. Penelitian tersebut menghasilkan data sintetis melalui simulator PaySim yang dirancang untuk meniru pola dan karakteristik transaksi keuangan *mobile* berdasarkan data asli. Data sintetis ini diciptakan untuk mengatasi keterbatasan akses terhadap data transaksi keuangan asli yang dilindungi oleh regulasi. Dengan demikian, data hasil simulasi ini digunakan sebagai alternatif untuk mendukung pengembangan model *fraud detection* menggunakan *machine learning*.

Data asli ini berasal dari catatan keuangan selama satu bulan dari sebuah layanan keuangan *mobile* yang digunakan di Afrika. Data tersebut diberikan oleh sebuah perusahaan multinasional yang menyediakan layanan keuangan seluler dan beroperasi di lebih dari 14 negara di seluruh dunia. Untuk keperluan analisis dataset asli ini kemudian diperkecil menjadi seperempat dari ukuran aslinya dan dibuat dalam bentuk *dataset* sintetis. Dataset ini mencakup total 6.362.620 sampel transaksi keuangan *mobile*, dimana 8.213 di antaranya diidentifikasi sebagai transaksi penipuan, sedangkan 6.354.407 transaksi lainnya diklasifikasikan sebagai transaksi non-penipuan. Adapun fitur-fitur yang terdapat dalam dataset tersebut adalah sebagai berikut:

Tabel 3.1 Data variabel PaySim

|  |  |  |
| --- | --- | --- |
| No | Nama Fitur | Deskripsi |
| 1 | *step* | Merepresentasikan unit waktu di dunia nyata. Dalam kasus ini, 1 step = 1 jam waktu. Total step = 744 (simulasi 30 hari) |
| 2 | *type* | Jenis transaksi, yaitu *CASH-IN, CASH-OUT, DEBIT, PAYMENT,* dan *TRANSFER* |
| 3 | *nameOrig* | Pelanggan yang memulai transaksi |
| 4 | *amount* | Jumlah transaksi yang dilakukan |
| 5 | *oldbalanceOrg* | Saldo awal pengirim sebelum transaksi dilakukan. |
| 6 | *newbalanceOrig* | Saldo baru pengirim setelah transaksi dilakukan |
| 7 | *nameDest* | Pelanggan yang menjadi penerima transaksi |
| 8 | *oldbalanceDest* | Saldo awal penerima sebelum transaksi dilakukan |
| 9 | *newbalanceDest* | Saldo baru penerima setelah transaksi dilakukan |
| 10 | *isFraud* | Mengkategorikan transaksi: 1 = penipuan, 0 = bukan penipuan |
| 11 | *isFlaggedFraud* | mengkategorikan upaya ilegal, jika men*transfer* lebih dari 200.000 dalam satu transaksi maka dalam pengawasan (1) dan jika dibawahnya maka bukan pengawasan (0) |

1. *step*: Dalam konteks simulasi PaySim, "*step*" mewakili satuan waktu yang digunakan untuk mengukur durasi simulasi. Setiap "*step*" dalam simulasi ini setara dengan 1 jam di dunia nyata. Jadi, jika simulasi memiliki total 744 *step*, ini berarti simulasi tersebut berlangsung selama 744 jam, atau sekitar 30 hari (karena 30 hari × 24 jam = 720 jam, ditambah beberapa jam ekstra untuk menutupi perbedaan dalam data). Setiap *step* mencatat transaksi yang terjadi selama 1 jam tersebut.
2. *type*: Ini merujuk pada kategori transaksi yang terjadi dalam simulasi. Ada beberapa jenis transaksi, yaitu:
3. *CASH-IN* adalah proses di mana saldo akun bertambah ketika pengguna menyetor uang tunai kepada seorang pedagang. Ini berarti pengguna menambahkan uang ke akun *mobile* mereka dengan membayar tunai.
4. *CASH-OUT* adalah kebalikan dari *CASH-IN*, yaitu proses penarikan uang tunai dari seorang pedagang yang mengurangi saldo akun pengguna. Dengan kata lain, pengguna mengeluarkan uang dari akun *mobile* mereka dalam bentuk tunai.
5. *Debit* serupa dengan *CASH-OUT*, namun dalam hal ini, uang dikirim dari layanan uang *mobile* ke akun bank. Ini juga mengurangi saldo akun *mobile* pengguna, tetapi uangnya dipindahkan ke bank, bukan ditarik sebagai tunai.
6. *PAYMENT* adalah proses pembayaran untuk barang atau jasa kepada seorang pedagang yang mengurangi saldo akun pengguna dan menambah saldo akun penerima (pedagang). Ini adalah transaksi yang terjadi saat membeli sesuatu atau membayar layanan.
7. *TRANSFER* adalah proses pengiriman uang dari satu pengguna layanan ke pengguna lain melalui *platform* uang *mobile*. Ini memungkinkan pengguna untuk men*transfer* dana langsung ke akun *mobile* orang lain, tanpa melalui perantara seperti bank atau pedagang.
8. *amount*: Ini adalah jumlah uang yang terlibat dalam setiap transaksi, dan nilainya dinyatakan dalam mata uang lokal. Misalnya, jika transaksi adalah 100.000, ini berarti jumlah uang yang terlibat dalam transaksi tersebut adalah 100.000 unit mata uang lokal.
9. *nameOrig*: Ini adalah identitas atau nama pelanggan yang memulai transaksi. "*nameOrig*" menunjukkan siapa yang memulai transaksi, seperti orang atau entitas yang mengirim uang dalam *transfer* atau yang menarik uang dalam kasus *CASH-OUT*.
10. *oldbalanceOrg*: Ini adalah saldo awal yang dimiliki oleh pelanggan sebelum transaksi terjadi. Misalnya, jika pelanggan memiliki 500.000 sebelum transaksi dan mereka melakukan transaksi senilai 100.000, maka "*oldbalanceOrg*" akan mencatat 500.000.
11. *newbalanceOrig*: Ini adalah saldo baru yang dimiliki oleh pelanggan setelah transaksi selesai. Menggunakan contoh sebelumnya, setelah transaksi senilai 100.000, saldo baru pelanggan akan menjadi 400.000, dan ini akan dicatat sebagai "*newbalanceOrig*".
12. *nameDest*: Ini adalah identitas atau nama pelanggan yang menerima transaksi. "*nameDest*" menunjukkan siapa yang menerima uang dalam sebuah *transfer*, atau akun yang diisi dengan dana dalam kasus *CASH-IN*.
13. *oldbalanceDest*: Ini adalah saldo awal dari penerima sebelum transaksi terjadi. Untuk pelanggan yang namanya dimulai dengan huruf "M" (yang mengacu pada pedagang), informasi ini tidak tersedia. Hal ini mungkin karena data untuk pedagang diperlakukan secara berbeda atau tidak sepenuhnya dicatat.
14. *newbalanceDest*: Ini adalah saldo baru dari penerima setelah transaksi selesai. Seperti "*oldbalanceDest*" informasi ini juga tidak tersedia untuk pelanggan dengan nama yang dimulai dengan huruf "M" (Pedagang).
15. *isFraud*: Ini adalah indikator apakah transaksi tersebut merupakan bagian dari kegiatan penipuan atau tidak dalam simulasi. Dalam *dataset* ini, penipuan biasanya melibatkan agen yang berusaha mengambil alih akun pelanggan, kemudian mencoba mengosongkan dana dari akun tersebut dengan men*transfer* uang ke akun lain dan akhirnya mencairkan uang tersebut dari sistem.
16. *isFlaggedFraud*: Ini adalah indikator apakah transaksi tersebut ditandai sebagai upaya penipuan yang potensial oleh model bisnis yang digunakan dalam simulasi. Simulasi ini dirancang untuk mendeteksi dan menandai *transfer* besar yang mencurigakan, terutama jika transaksi tersebut melibatkan *transfer* uang lebih dari 200.000 dalam satu transaksi. Jika transaksi melebihi ambang batas ini, maka akan ditandai sebagai "*isFlaggedFraud*" menandakan bahwa transaksi tersebut mungkin ilegal dan perlu diperiksa lebih lanjut.

## Pendekatan Penelitian

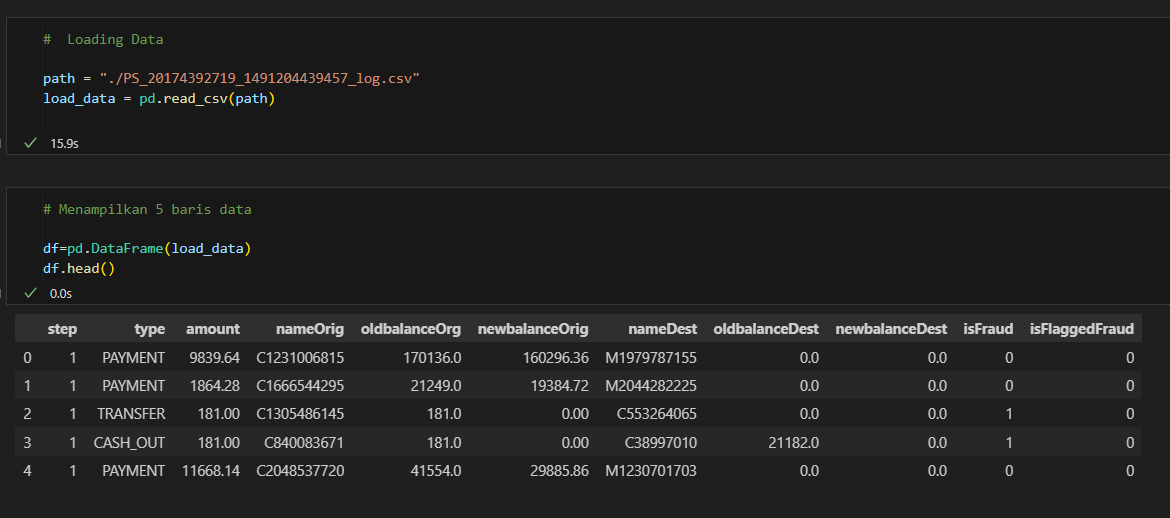
Penelitian ini mengadopsi pendekatan kuantitatif dengan metode eksperimen untuk merancang dan menguji model *machine learning* dalam deteksi penipuan pada data transaksi keuangan. Dengan menggunakan studi kasus, penelitian ini berfokus pada analisis data sintetis dari transaksi keuangan *mobile* yang dikembangkan untuk meniru pola transaksi keuangan asli. Pendekatan ini memungkinkan pengujian model *fraud detection* secara akurat tanpa melibatkan data asli, sehingga melindungi privasi namun tetap menghadirkan skenario transaksi yang realistis.

## Instrumen Penelitian

Instrumen utama dalam penelitian ini adalah algoritma *machine learning* yang cocok untuk *fraud detection.* Berdasarkan studi literatur yang telah dilakukan terdapat 3 model *machine learning* yang banyak diterapkan untuk mendeteksi *fraud* pada transaksi keuangan yaitu *Decision Tree, Random Forest,* dan *XGBoost.* Untuk memastikan efektivitas dan menentukan model yang akurat, penelitian ini menggunakan metrik evaluasi seperti *Accuracy, Precision, Recall, F1-Score,* dan *ROC-AUC* yang memberikan gambaran komprehensif mengenai kinerja model dalam mengidentifikasi transaksi keuangan mencurigakan dengan tingkat keandalan yang tinggi.

# HASIL DAN PEMBAHASAN

## *Load Data*

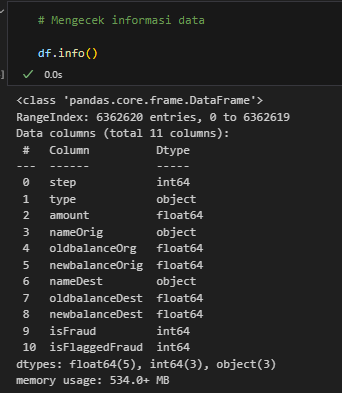


Gambar 4.1 Loading dataset

Gambar 4.1 menunjukkan data terkait *fraud* dalam format CSV dimuat ke dalam lingkungan program menggunakan fungsi *read\_csv* dari *library Pandas* di Visual Studio Code, lalu diubah menjadi *Pandas DataFrame* untuk memudahkan manipulasi dan analisis. Setelah data berhasil diimpor, dilakukan eksplorasi awal dengan menampilkan lima baris pertama dari *dataset* yang terdiri dari 11 kolom, yaitu *step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFraud*, dan *isFlaggedFraud*. Proses ini memberikan gambaran awal tentang struktur dan jenis data sebagai dasar untuk analisis lebih lanjut.

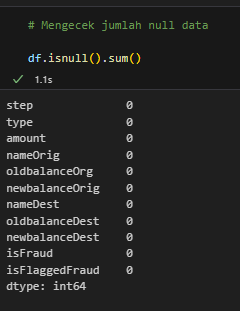
## *Preprocessing Data* dan Analisis Data

Sebelum melanjutkan ke tahap *preprocessing* dan analisis data, dilakukan pengecekan awal terhadap informasi *dataset* untuk memahami karakteristiknya secara keseluruhan. Pengecekan ini meliputi jumlah baris dan kolom data, tipe data pada setiap kolom, jumlah memori yang digunakan, serta distribusi tipe data dalam *dataset*. Informasi ini diperoleh menggunakan fungsi info() dari *library Pandas* yang memberikan detail penting mengenai struktur *dataset*.

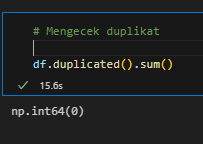


Gambar 4.2 Mengecek informasi data

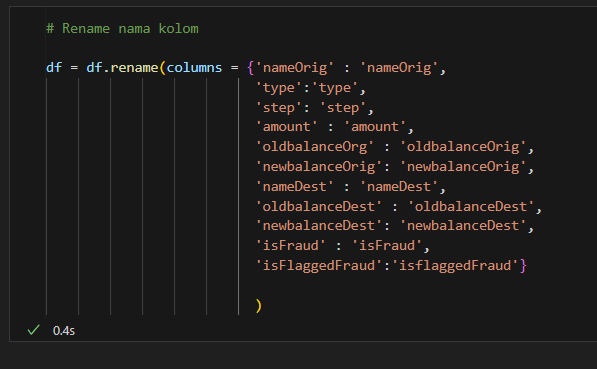
Berdasarkan gambar 4.2 hasil pengecekan informasi data, *dataset* ini memiliki 6.362.620 baris dan 11 kolom dengan total penggunaan memori sebesar 534.0 MB. Kolom-kolom dalam *dataset* terbagi menjadi tiga jenis tipe data: tipe *int64* untuk nilai numerik integer yang mencakup kolom *step, isFraud,* dan *isFlaggedFraud,* tipe *object* untuk data teks seperti pada kolom *type, nameOrig,* dan *nameDest,* serta tipe *float64* untuk nilai numerik desimal yang mencakup kolom *amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest,* dan *newbalanceDest.*



Gambar 4.3 Mengecek nilai null



Gambar 4.4 Mengecek duplikat data



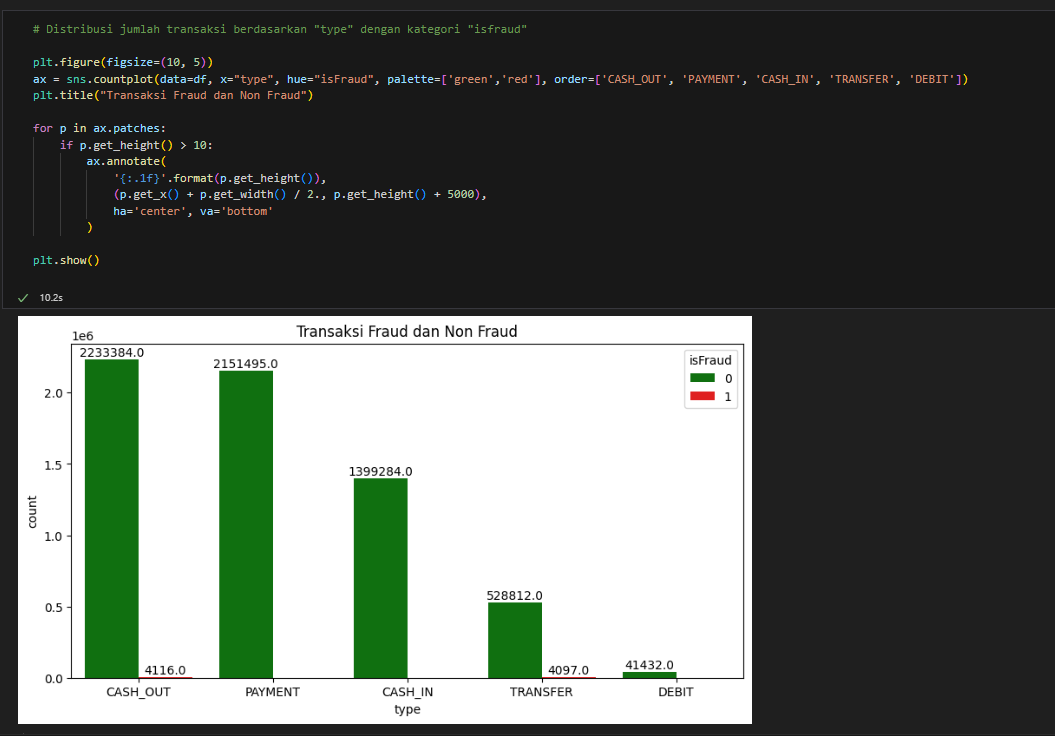
Gambar 4.5 Rename kolom

Setelah memastikan informasi pada *dataset*, pada gambar 4.3, 4.4, dan 4.5 proses *preprocessing* awal dilakukan untuk meningkatkan kualitas data. Proses ini meliputi pengecekan dan penanganan data kosong *(null),* penghapusan data duplikat, pengubahan nama kolom agar lebih deskriptif, dan standarisasi format data.



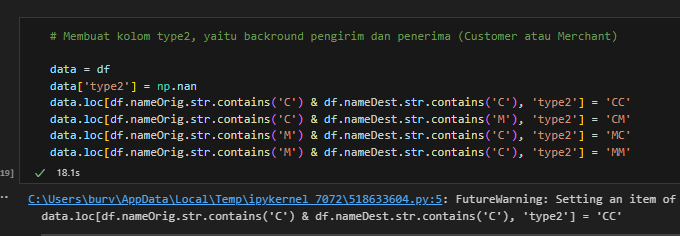
Gambar 4.7 Jumlah *transaksi fraud* dan *non fraud*

Barplot pada gambar 4.7 menunjukkan jumlah transaksi yang dikategorikan sebagai *fraud* dan *non-fraud*. Terlihat bahwa jumlah transaksi *non-fraud* (ditandai dengan angka 0) jauh lebih besar dibandingkan transaksi *fraud* (ditandai dengan angka 1). Dari data pada grafik, terdapat sekitar 6.354.407 transaksi *non-fraud,* sedangkan hanya sekitar 8.213 transaksi yang terindikasi sebagai *fraud*. Ketimpangan ini menunjukkan bahwa kasus *fraud* dalam transaksi sangat jarang terjadi dibandingkan dengan transaksi yang sah dan terjadinya *imbalance data* yang bisa menjadi tantangan dalam mendeteksi *fraud* secara efektif.



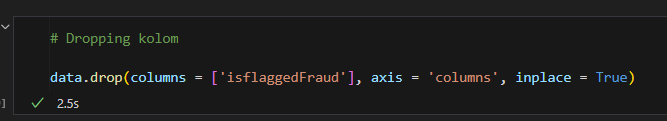
Gambar 4.8 Distribusi transaksi *fraud* dan *non fraud* berdasarkan type

Grafik pada gambar 4.8 diatas menunjukkan distribusi transaksi berdasarkan jenisnya serta apakah transaksi tersebut termasuk *fraud* atau bukan. Jenis transaksi *cash\_out* dan *transfer* memiliki jumlah kasus *fraud* yang lebih tinggi dibandingkan jenis lainnya, meskipun secara keseluruhan jumlah transaksi *non-fraud* jauh lebih dominan untuk semua jenis. Transaksi *seperti payment, cash\_in,* dan *debit* memiliki jumlah kasus *fraud* yang sangat kecil atau bahkan tidak ada, menunjukkan bahwa jenis transaksi ini cenderung lebih aman dari *fraud*. Hal ini mengindikasikan bahwa fokus analisis *fraud* sebaiknya lebih diarahkan pada transaksi jenis *cash\_out* dan *transfer,* mengingat risiko yang lebih tinggi.



Gambar 4.9 *Feature Engineering*

Gambar 4.9 diatas menggambarkan bagaimana kode program digunakan untuk membuat kolom baru bernama *type2* yang menggambarkan latar belakang pengirim dan penerima dalam transaksi, yaitu apakah berasal dari *customer* (C)atau *merchant* (M). Nilai pada kolom *type2* ditentukan berdasarkan kombinasi string pada kolom *nameOrig* (pengirim) dan *nameDest* (penerima). Kombinasi ini menghasilkan empat kategori: CC *(customer ke customer),* CM *(customer ke merchant),* MC *(merchant ke customer),* dan MM *(merchant ke merchant).* Proses ini dilakukan dengan menggunakan kondisi logika pada *nameOrig* dan *nameDest,* lalu menetapkan nilai sesuai kategori yang relevan.



Gambar 4.10 *Droping* kolom

Gambar 4.10 menunjukkan fungsi *drop* dari Pandas untuk menghapus kolom *isFlaggedFraud* dari *dataframe* *data.* Parameter *inplace=True* memastikan perubahan dilakukan langsung pada *dataframe* asli tanpa membuat salinan baru. Setelah dilakukan *feature engineering* dan *droping* kolom atau *feature* maka variabel baru yang akan digunakan dalam proses selanjutnya adalah sebagai berikut:

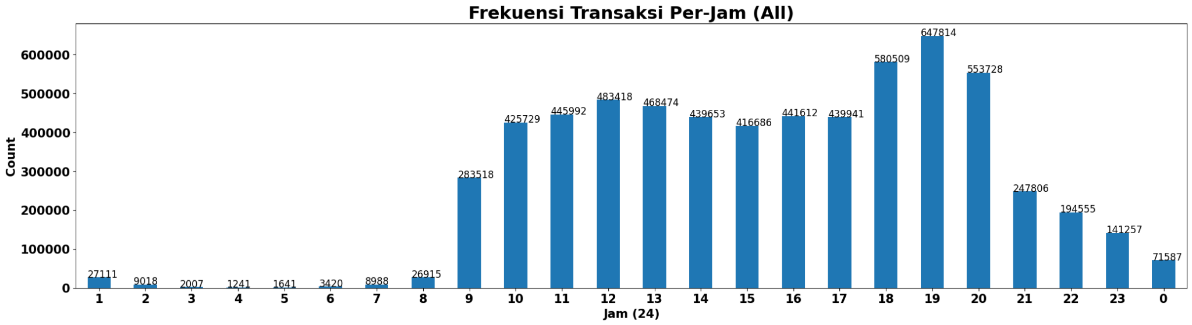
Tabel 4.1 Variabel baru dataset PaySim

|  |  |  |
| --- | --- | --- |
| No | Nama Fitur | Deskripsi |
| 1 | *step* | Merepresentasikan unit waktu di dunia nyata. Dalam kasus ini, 1 step = 1 jam waktu. Total step = 744 (simulasi 30 hari) |
| 2 | *type* | Jenis transaksi, yaitu *CASH-IN, CASH-OUT, DEBIT, PAYMENT,* dan *TRANSFER* |
| 3 | *type2* | Sub Jenis transaksi, yaitu CC *(cutomer to customer)* dan CM *(cutomer to merchant)* |
| 4 | *nameOrig* | Pelanggan yang memulai transaksi |
| 5 | *amount* | Jumlah transaksi yang dilakukan |
| 6 | *oldbalanceOrg* | Saldo awal pengirim sebelum transaksi dilakukan. |
| 7 | *newbalanceOrig* | Saldo baru pengirim setelah transaksi dilakukan |
| 8 | *nameDest* | Pelanggan yang menjadi penerima transaksi |
| 9 | *oldbalanceDest* | Saldo awal penerima sebelum transaksi dilakukan |
| 10 | *newbalanceDest* | Saldo baru penerima setelah transaksi dilakukan |
| 11 | *isFraud* | Mengkategorikan transaksi: 1 = penipuan, 0 = bukan penipuan |



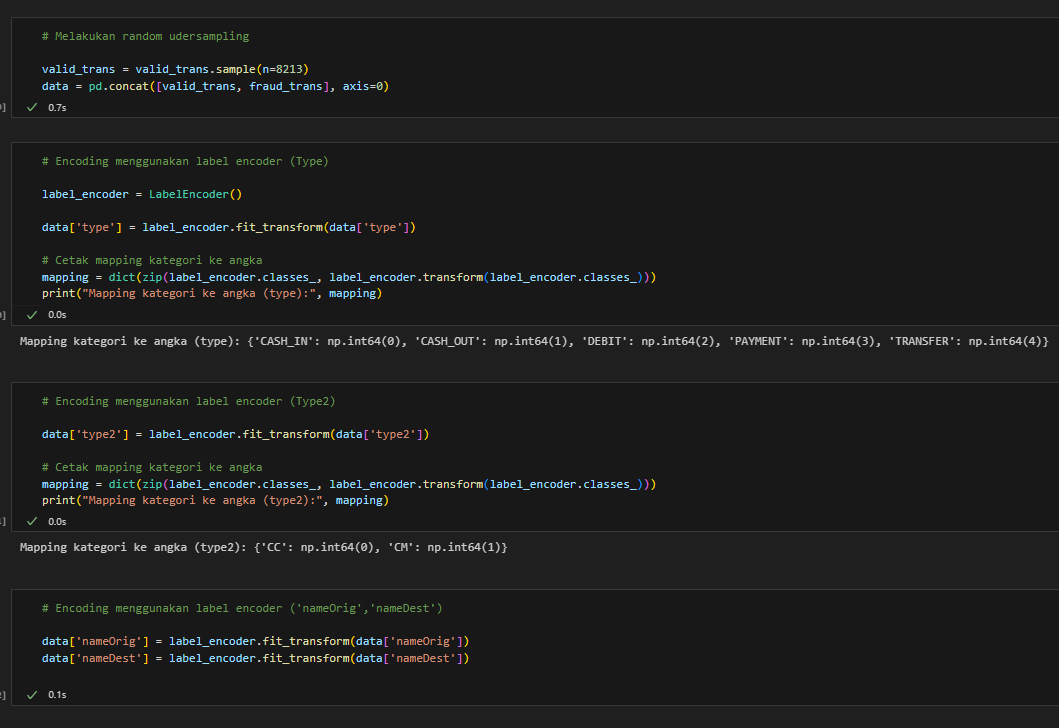
Gambar 4.11 Distribusi transaksi *fraud* dan *non fraud* berdasarkan type2

Gambar 4.11 menunjukkan distribusi jumlah transaksi berdasarkan jenis transaksi *Customer to Customer* (CC) dan *Customer to Merchant* (CM), dengan kategori transaksi *fraud* (*isFraud*). Dari hasil visualisasi, transaksi CC memiliki jumlah yang jauh lebih tinggi dibandingkan CM, dengan sekitar 4.202.912 transaksi *non-fraud* dan hanya 8.213 *transaksi fraud*. Sementara itu, transaksi CM memiliki sekitar 2.151.495 transaksi *non-fraud* tanpa indikasi transaksi fraud yang signifikan. Hal ini menunjukkan bahwa transaksi antara *customer* (CC) lebih umum terjadi, tetapi juga lebih rentan terhadap *fraud* dibandingkan dengan transaksi *customer to merchant* (CM).



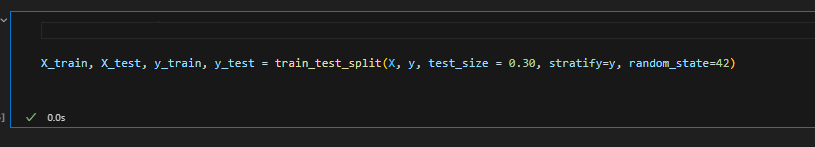
Gambar 4.12 Frekuensi Transaksi dalam 24 Jam

Grafik pada gambar 4.12 menunjukkan distribusi frekuensi transaksi per jam dalam format 24 jam. Aktivitas transaksi mulai meningkat signifikan pada pukul 9 pagi dan mencapai puncaknya pada pukul 19 malam, dengan jumlah transaksi tertinggi sekitar 647.814 transaksi. Setelah pukul 19, jumlah transaksi mulai menurun secara bertahap hingga mencapai titik terendah pada pukul 3-5 dini hari. Pola ini mengindikasikan bahwa transaksi paling aktif terjadi pada jam kerja hingga awal malam, sedangkan dini hari adalah periode dengan aktivitas transaksi paling rendah. Hal ini dapat mencerminkan kebiasaan pengguna dalam melakukan transaksi selama waktu produktif.



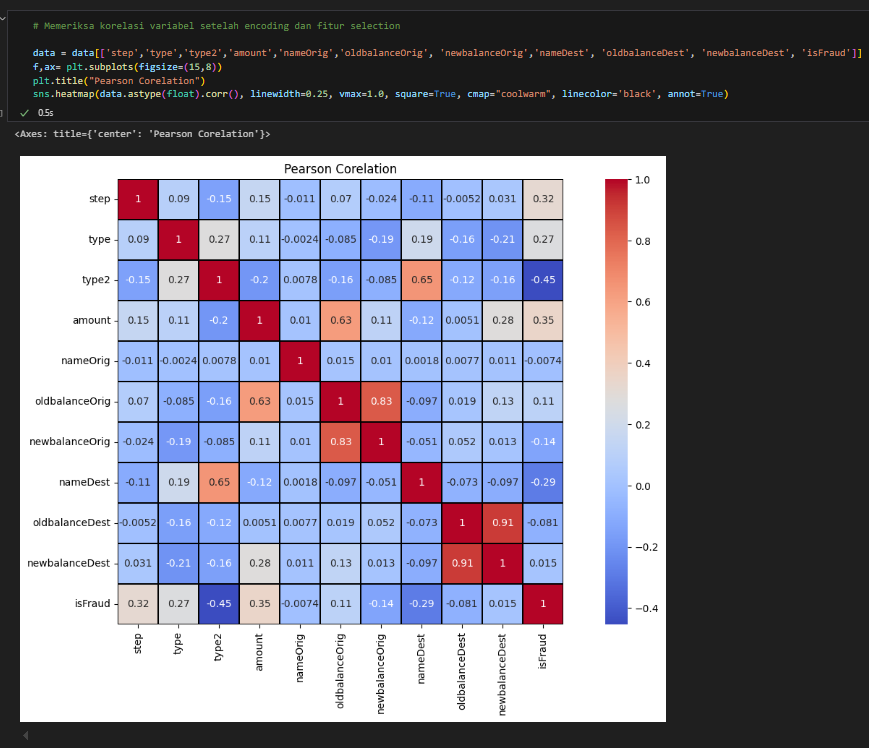
Gambar 4.13 Melakukan *resampling* dan *label encoding*

Gambar 4.13 yang ditampilkan melakukan beberapa tahap pemrosesan data. Pertama, dilakukan *random undersampling* dengan memilih sejumlah transaksi valid (sebanyak 8.213) untuk menyeimbangkan dataset dengan transaksi *fraud*. Selanjutnya, dilakukan *label encoding* untuk mengonversi data kategorikal menjadi numerik, termasuk kolom "type" yang berisi jenis transaksi (*CASH\_IN, CASH\_OUT, DEBIT, PAYMENT, TRANSFER*), "type2" yang membedakan transaksi antara *Customer to Customer (CC)* dan *Customer to Merchant (CM)*, serta *encoding* untuk kolom "nameOrig" dan "nameDest" yang merepresentasikan ID pengirim dan penerima. Proses ini penting untuk memungkinkan algoritma *machine learning* memahami data kategorikal dalam bentuk numerik, sehingga model dapat lebih efektif dalam mendeteksi pola transaksi *fraud.*



Gambar 4.14 *Splitting data*

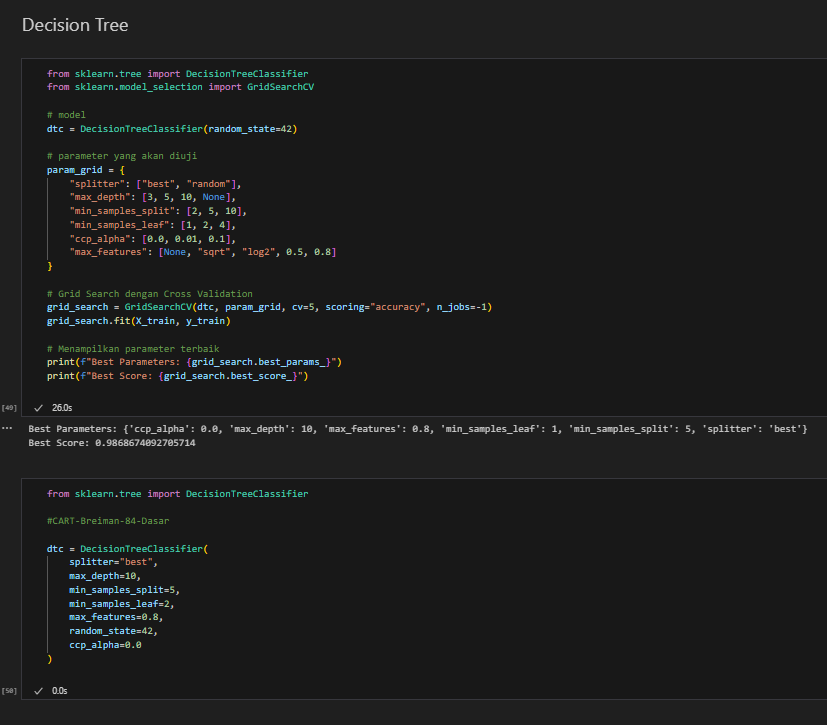
Gambar 4.14 menunjukkan cara membagi dataset menjadi data pelatihan (70%) dan data pengujian (30%) menggunakan fungsi *train\_test\_split* dari *scikit-learn,* dengan memastikan distribusi label tetap proporsional melalui parameter *stratify=y.* Parameter *random\_state=42* digunakan untuk memastikan pembagian data konsisten setiap kali kode dijalankan.



Gambar 4.15 *Corelation Matrik*

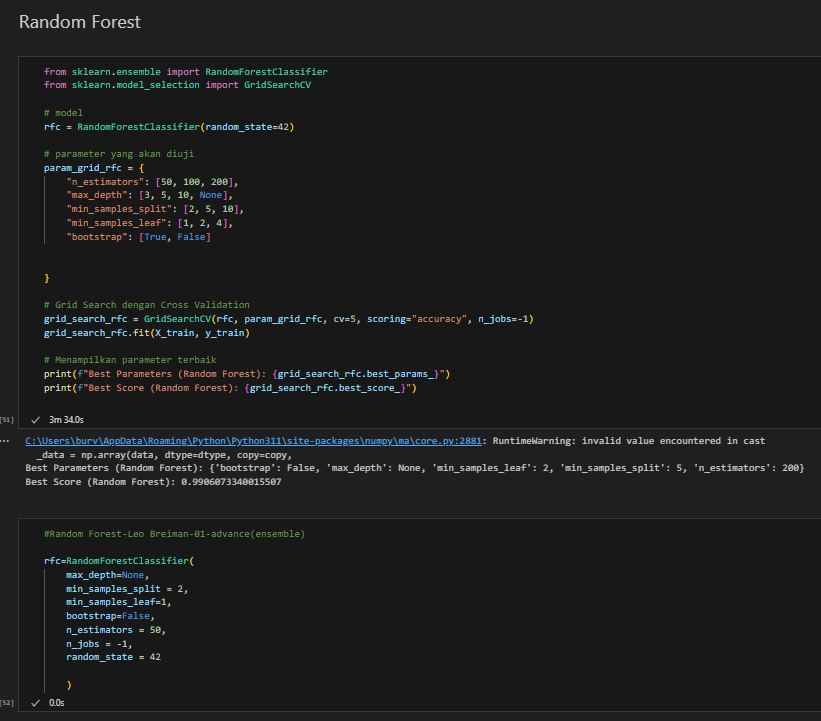
Gambar 4.15 menunjukkan korelasi antara berbagai fitur dalam dataset setelah proses *encoding* dan seleksi fitur menggunakan *Pearson Correlation*. Warna merah menunjukkan korelasi positif yang kuat, sedangkan warna biru menunjukkan korelasi negatif yang lebih kuat. Dari hasil visualisasi, terlihat bahwa fitur *oldbalanceOrig* dan *newbalanceOrig* memiliki korelasi yang sangat tinggi (0.83), demikian juga *oldbalanceDest* dan *newbalanceDest* (0.91), menunjukkan bahwa saldo awal dan saldo baru saling terkait erat. Fitur *type2* memiliki korelasi negatif yang cukup kuat dengan *isFraud* (-0.45), yang menunjukkan bahwa jenis transaksi berperan penting dalam menentukan apakah suatu transaksi merupakan *fraud* atau *non fraud*. Selain itu, *amount* memiliki korelasi 0.35 dengan *isFraud* yang berarti jumlah transaksi juga memiliki pengaruh terhadap kemungkinan *fraud*.

## *Modeling*



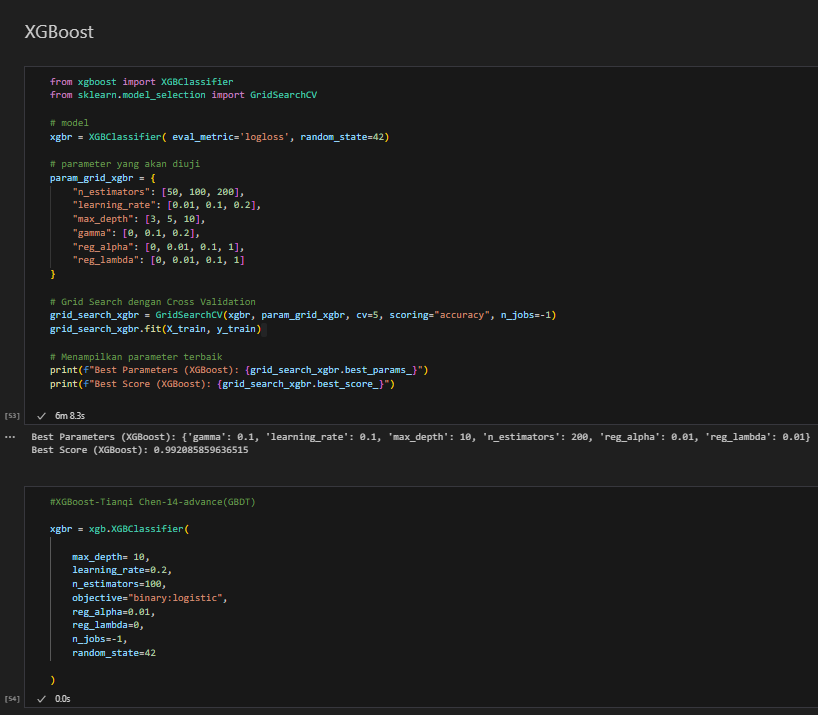
Gambar 4.16 *Modelling Decision Tree*

Gambar 4.16 menunjukkan proses pembuatan model *Decision Tree Classifier* dengan optimasi parameter menggunakan *Grid Search Cross Validation*. Pertama, model *DecisionTreeClassifier* dibuat dengan parameter *random\_state=42* untuk memastikan hasil yang konsisten. Kemudian, dilakukan pencarian parameter terbaik (*splitter*, *max\_depth*, *min\_samples\_split*, *min\_samples\_leaf*, *ccp\_alpha*, dan *max\_features*) dengan *GridSearchCV* yang menguji berbagai kombinasi parameter dan memilih yang memberikan akurasi tertinggi. Hasilnya menunjukkan bahwa parameter terbaik mencakup *max\_depth=10*, *max\_features=0.8*, *min\_samples\_leaf=1*, *min\_samples\_split=5*, dan *splitter='best'*, dengan akurasi mencapai 98.86%. Setelah parameter optimal ditemukan, model *Decision Tree* dikonfigurasi ulang dengan parameter tersebut untuk meningkatkan performa dalam mendeteksi transaksi *fraud* secara lebih efektif.



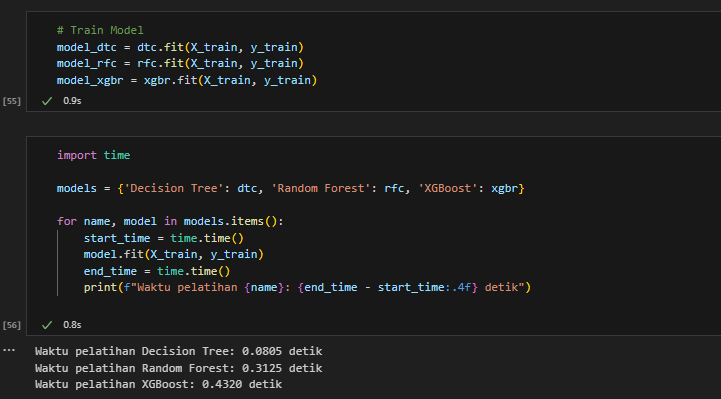
Gambar 4.17 *Modeling Random Forest*

Gambar 4.17 menunjukkan proses pelatihan model *Random Forest Classifier* dengan optimasi parameter menggunakan *Grid Search Cross Validation*. Model *RandomForestClassifier* dibuat dengan *random\_state=42* untuk memastikan hasil yang optimal. *GridSearchCV* digunakan untuk mencari kombinasi parameter terbaik, termasuk jumlah estimator (*n\_estimators*), kedalaman maksimum (*max\_depth*), jumlah sampel minimum untuk split (*min\_samples\_split*), jumlah sampel minimum pada daun (*min\_samples\_leaf*), serta penggunaan *bootstrap*. Hasil pencarian menunjukkan bahwa kombinasi terbaik adalah *n\_estimators=200*, *max\_depth=None*, *min\_samples\_leaf=2*, *min\_samples\_split=5*, dan *bootstrap=False*, dengan akurasi terbaik mencapai 99.06%. Model kemudian dikonfigurasi ulang menggunakan parameter optimal ini untuk meningkatkan kinerja dalam deteksi transaksi fraud.



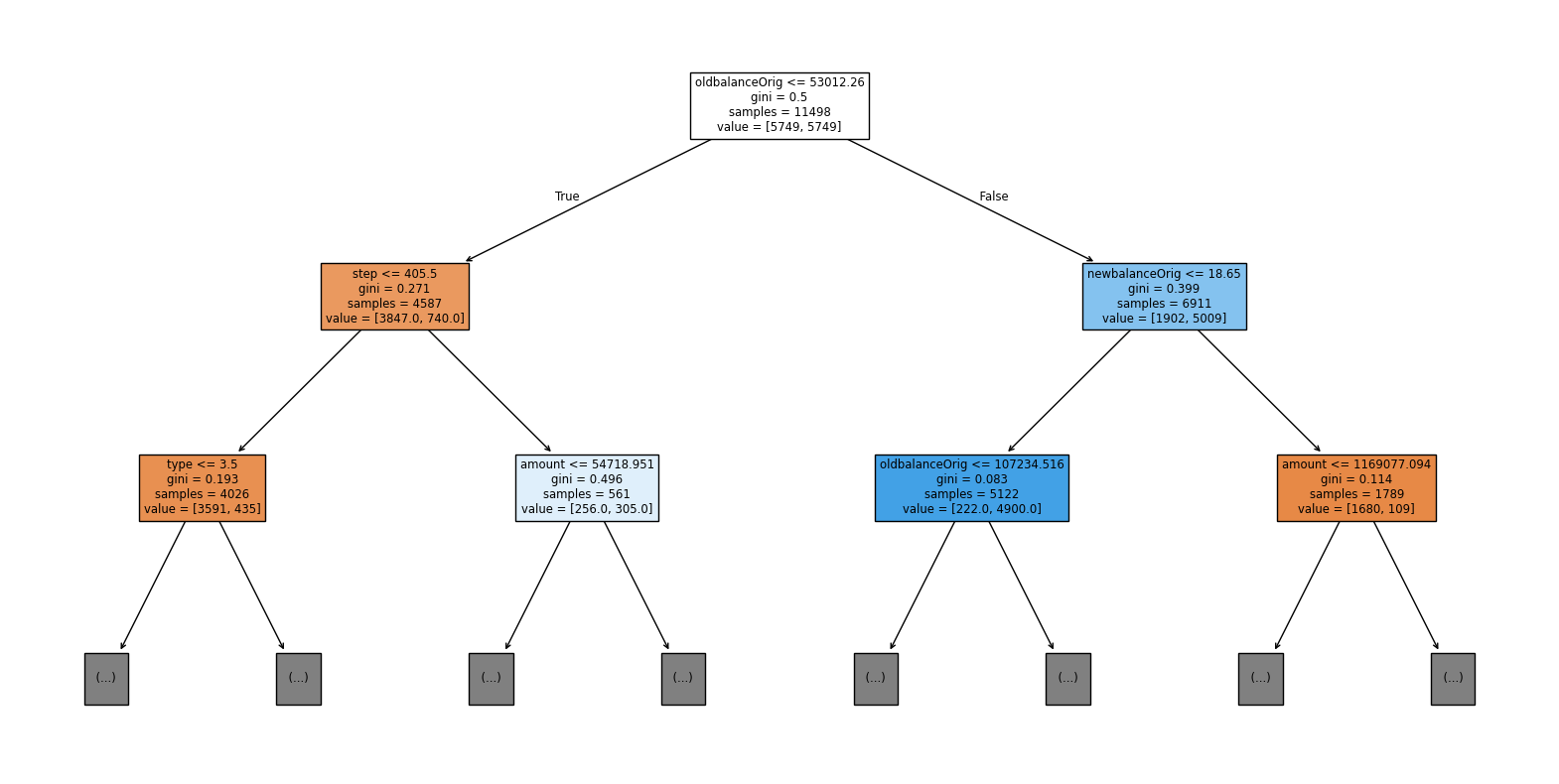
Gambar 4.18 *Modeling XGBoost*

Gambar 4.18 menunjukkan implementasi model *XGBoost Classifier* dengan optimasi parameter menggunakan *Grid Search Cross Validation*. Model *XGBClassifier* dibuat dengan *eval\_metric='logloss'* dan *random\_state=42* untuk memastikan hasil yang dapat optimal. Proses *tuning* dilakukan dengan *GridSearchCV* untuk mencari parameter optimal, termasuk jumlah estimator (*n\_estimators*), *learning rate*, kedalaman maksimum (*max\_depth*), serta regulasi *gamma*, *reg\_alpha*, dan *reg\_lambda*. Hasil pencarian menunjukkan parameter terbaik adalah *gamma=0.1*, *learning\_rate=0.1*, *max\_depth=10*, *n\_estimators=200*, *reg\_alpha=0.01*, dan *reg\_lambda=0.01*, dengan akurasi terbaik mencapai 99.29%. Model kemudian dikonfigurasi ulang dengan parameter optimal ini untuk meningkatkan performa dalam mendeteksi transaksi *fraud* secara lebih akurat dan efisien.



Gambar 4.19 Waktu pelatihan model

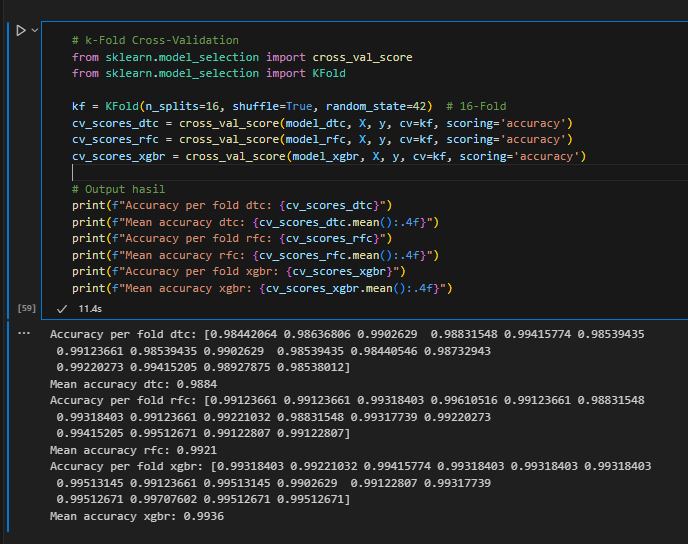
Gambar 4.19 menunjukkan model yang digunakan: *Decision Tree (dtc), Random Forest (rfc),* dan *XGBoost (xgbr)* menggunakan dataset X\_train dan y\_train. Bagian pertama kode melakukan pelatihan model secara langsung tanpa mencatat waktu eksekusi. Bagian kedua menggunakan modul time untuk mengukur waktu pelatihan masing-masing model dengan mencatat waktu sebelum dan sesudah pemanggilan fit(), lalu menghitung selisihnya. Hasil eksekusi menunjukkan bahwa *Decision Tree* memiliki waktu pelatihan tercepat (0.0805 detik), diikuti oleh *Random Forest* (0.3125 detik), dan *XGBoost* yang memiliki waktu pelatihan terlama (0.4320 detik). Hal ini menunjukkan bahwa model yang lebih kompleks cenderung membutuhkan waktu pelatihan lebih lama dibandingkan model yang lebih sederhana.



Gambar 4.20 Visualisasi pohon model *decision tree*

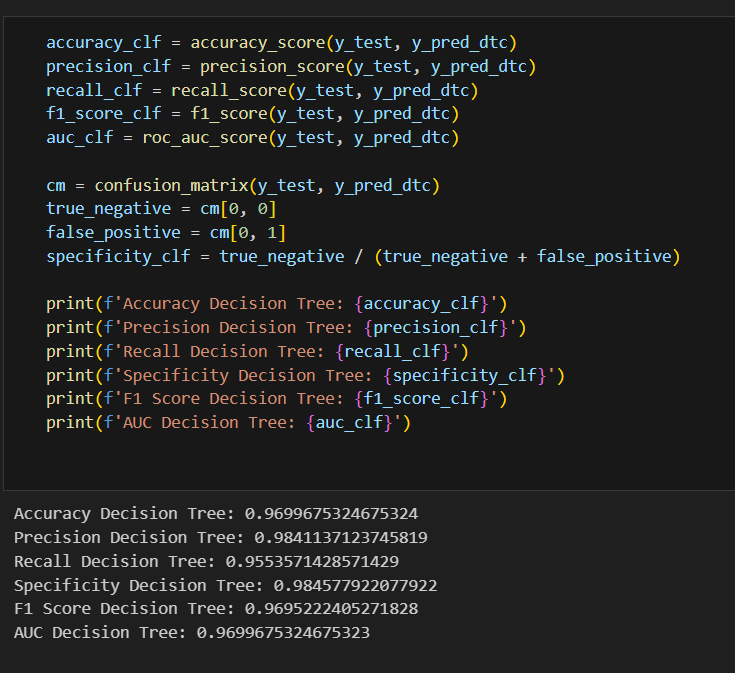
Gambar 4.20 merupakan visualisasi pohon keputusan (*Decision Tree*) yang digunakan untuk mendeteksi transaksi *fraud* berdasarkan fitur-fitur seperti *step, type, type2, amount, nameOrig, oldbalanceOrig, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest,* dan *isFraud*. Pohon keputusan ini bekerja dengan membagi dataset berdasarkan nilai fitur-fitur tersebut menggunakan kriteria *Gini impurity* untuk menentukan seberapa murni pembagian kelas dalam tiap simpul (*node*). Simpul akar dimulai dengan fitur *oldbalanceOrig*, yang membagi data menjadi dua cabang utama. Semakin ke bawah, setiap node membagi data lebih lanjut berdasarkan fitur yang paling relevan hingga mencapai daun (*leaf nodes*), yang menunjukkan hasil klasifikasi akhir. Warna dalam diagram menunjukkan proporsi kelas dalam setiap node, dengan warna biru yang lebih dominan menunjukkan data *non-fraud* dan warna oranye yang lebih dominan menunjukkan data *fraud*. Model ini dapat digunakan untuk memahami pola transaksi yang mencurigakan dan membantu dalam deteksi *fraud* yang lebih efektif.

## *Evaluation*



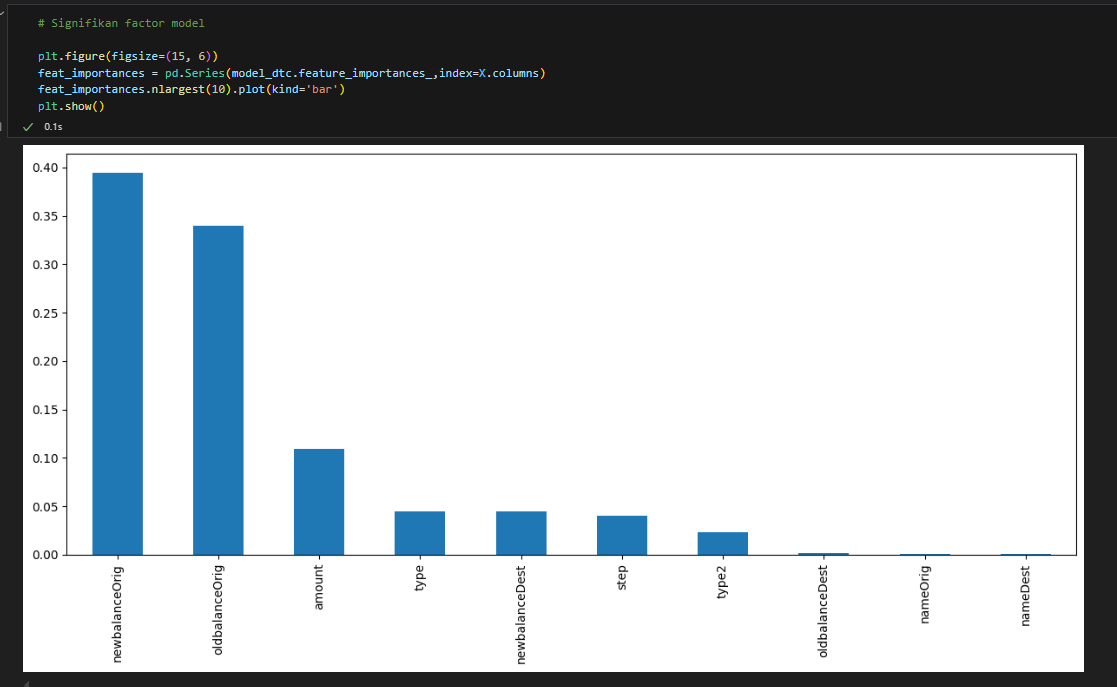
Gambar 4.21 *Cross-Validation*

Pada gambar 4.21 menunjukkan penerapan *k-Fold Cross-Validation* dengan 16 *fold* (*n\_splits=16*) untuk mengevaluasi performa tiga model klasifikasi: *Decision Tree Classifier (DTC)*, *Random Forest Classifier (RFC)*, dan *XGBoost Classifier (XGBR)*. Teknik ini membagi dataset menjadi 16 bagian, melatih model pada 15 bagian, dan mengujinya pada 1 bagian, lalu mengulangi proses tersebut hingga semua bagian telah digunakan sebagai data uji. Hasil evaluasi menunjukkan bahwa *Decision Tree Classifier* memiliki rata-rata akurasi 0.9854, *Random Forest Classifier* memiliki rata-rata akurasi 0.9931, dan *XGBoost Classifier* mencapai akurasi tertinggi sebesar 0.9936. Ini menunjukkan bahwa *XGBoost* memiliki performa terbaik dalam mendeteksi *fraud*, diikuti oleh *Random Forest* dan *Decision Tree*. Teknik *k-Fold Cross-Validation* memastikan evaluasi model lebih akurat dengan mengurangi kemungkinan bias dari pembagian dataset.



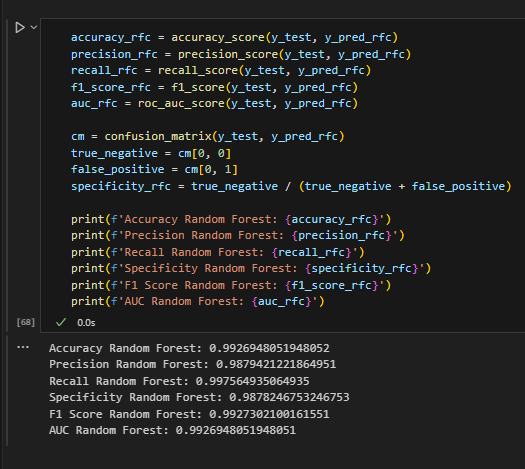
Gambar 4.22 Evaluasi model *decision tree*

Gambar 4.22 adalah hasil mengevaluasi performa model *Decision Tree Classifier* menggunakan metrik evaluasi utama*: accuracy, precision, recall, F1-score, specificity,* dan *AUC (Area Under Curve)*. Hasil pengujian menunjukkan bahwa model memiliki *accuracy* sebesar 0.9878, *precision* 0.9831, *recall* 0.9926, *F1-score* 0.9878, *specifity* 0.9829, dan *AUC* 0.9878. Nilai *recall* yang tinggi menunjukkan bahwa model sangat baik dalam menangkap transaksi *fraud,* sementara *specifity* yang tinggi menunjukkan model juga dapat mengidentifikasi transaksi *non-fraud* dengan baik.



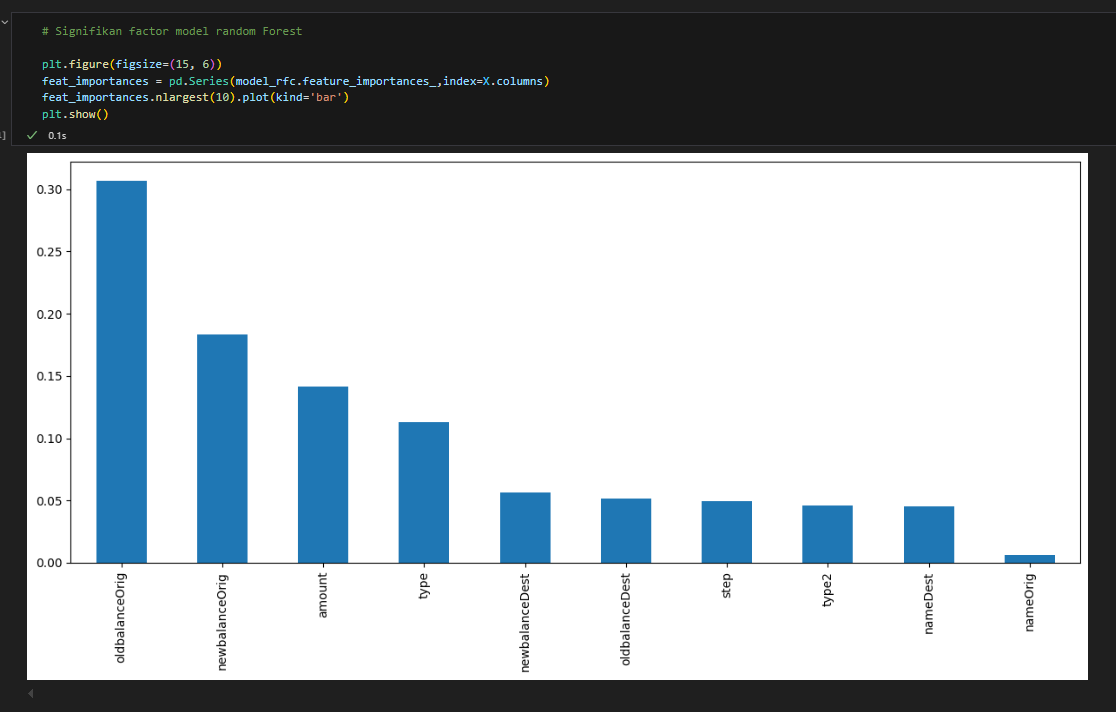
Gambar 4.23 Signifikan faktor *Decision Tree*

Gambar 4.23 menunjukkan faktor-faktor terpenting dalam mendeteksi transaksi fraud berdasarkan model *Decision Tree Classifier*. Dari hasil analisis, fitur *newbalanceOrig* dan *oldbalanceOrig* memiliki pengaruh terbesar dalam klasifikasi menunjukkan bahwa saldo awal dan saldo setelah transaksi sangat penting dalam mendeteksi aktivitas mencurigakan. Oleh karenanya, banyak kasus *fraud* melibatkan perpindahan saldo yang tidak wajar atau saldo pengirim yang berkurang drastis setelah transaksi. Selain itu, fitur *amount* (jumlah transaksi) juga memiliki kontribusi yang signifikan, mengindikasikan bahwa nominal transaksi bisa menjadi indikator penting dalam identifikasi *fraud.* Fitur lain seperti *type*, *newbalanceDest*, dan *step* juga memberikan kontribusi tetapi dengan bobot yang lebih kecil. Dari sudut pandang sistem deteksi *fraud*, model ini menunjukkan bahwa perubahan saldo akun pengirim sebelum dan sesudah transaksi merupakan sinyal utama dalam mengidentifikasi anomali yang berpotensi sebagai tindakan *fraud.*

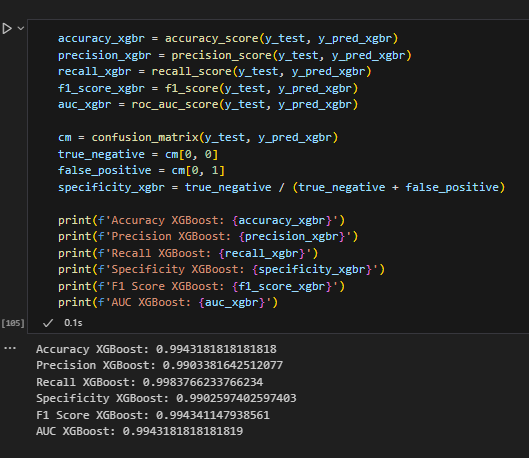


Gambar 4.24 Evaluasi model *Random Forest*

Gambar 4.24 mengevaluasi performa model *Random Forest Classifier* menggunakan berbagai metrik seperti *accuracy, precision, recall, specificity, F1-score,* dan *AUC.* Hasil evaluasi menunjukkan bahwa model memiliki *accuracy* 0.9926, *precision* 0.9879, *recall* 0.9975, *specificity* 0.9878, *F1-score* 0.9927, dan *AUC* 0.9926.

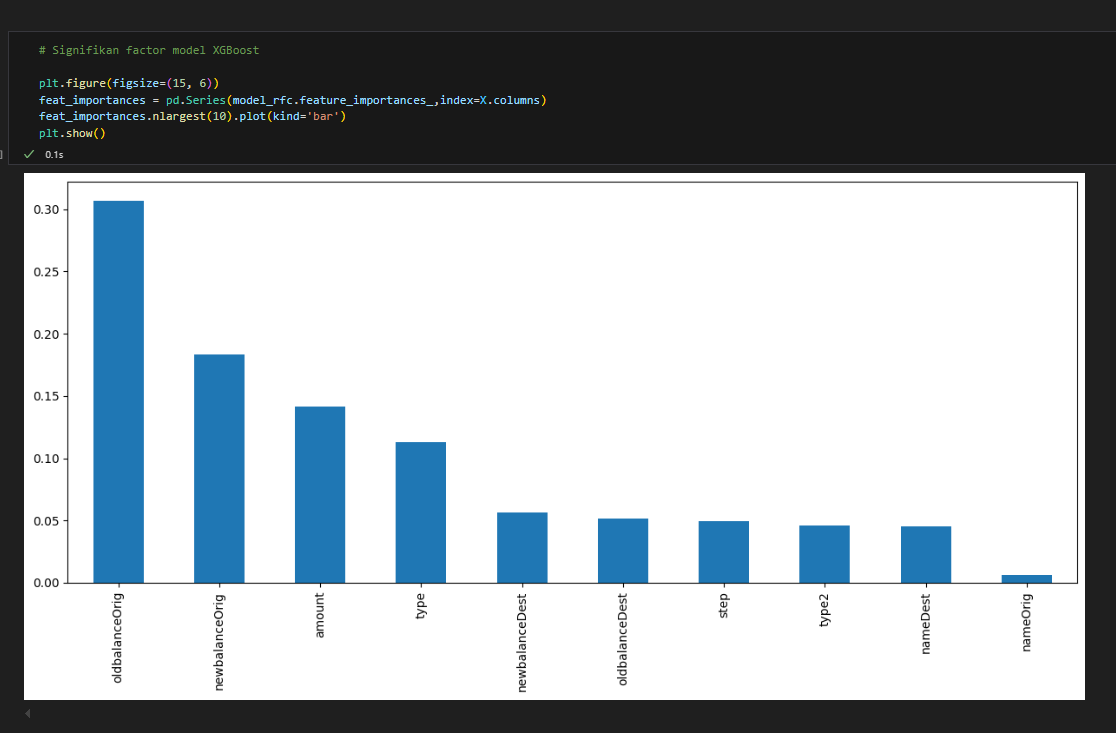
Gambar 4.25 Signifikan faktor *Random Forest*

Gambar 4.25 menampilkan faktor-faktor paling signifikan dalam mendeteksi transaksi *fraud* berdasarkan model *Random Forest Classifier*. Dari perspektif sistem *fraud detection*, fitur *oldbalanceOrig* (saldo awal pengirim) dan *newbalanceOrig* (saldo akhir pengirim) memiliki pengaruh terbesar dalam klasifikasi transaksi mencurigakan. Hal ini menunjukkan bahwa perubahan saldo sebelum dan sesudah transaksi merupakan indikator utama dalam mengidentifikasi anomali keuangan. Selain itu, jumlah transaksi (*amount*) juga memainkan peran penting, karena *fraud* sering kali terjadi dalam transaksi dengan nilai yang tidak wajar. Faktor lain seperti *type* (jenis transaksi), *newbalanceDest* (saldo akhir penerima), dan *step* (waktu transaksi) juga berkontribusi, tetapi dengan bobot yang lebih kecil. Dari perspektif sistem deteksi *fraud*, fokus utama harus diberikan pada analisis pergerakan saldo pengirim dan pola transaksi, karena perubahan signifikan pada saldo dapat mengindikasikan aktivitas mencurigakan yang perlu ditandai untuk pemeriksaan lebih lanjut.



Gambar 4.26 Evaluasi model *XGBoost*

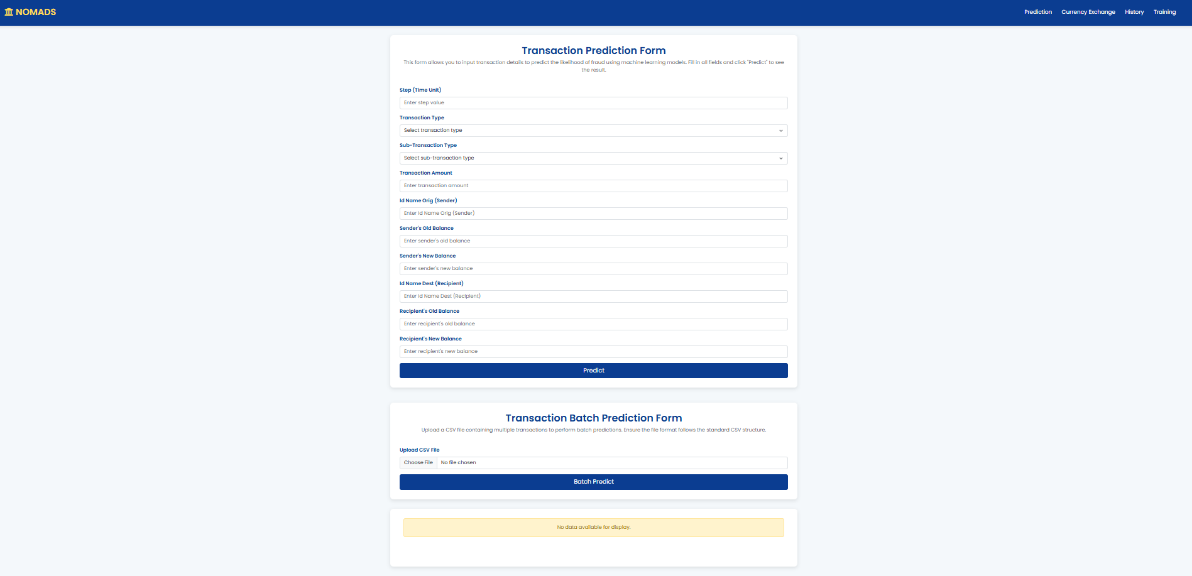
Gambar 4.26 adalah hasil mengevaluasi performa model *XGBoost Classifier* menggunakan metrik *accuracy, precision, recall,* *specificity*, *F1-score*, dan *AUC*. Hasil evaluasi menunjukkan bahwa model memiliki *accuracy* 0.9943, *precision* 0.9934, *recall* 0.9938, *specificity* 0.9902, *F1-score* 0.9943, dan *AUC* 0.9943. Dengan nilai *accuracy* dan *F1-score* yang sangat tinggi, model ini menunjukkan kemampuan yang sangat baik dalam mengidentifikasi transaksi *fraud* dan *non-fraud* secara seimbang.



Gambar 4.27 Signifikan faktor *XGBoost*

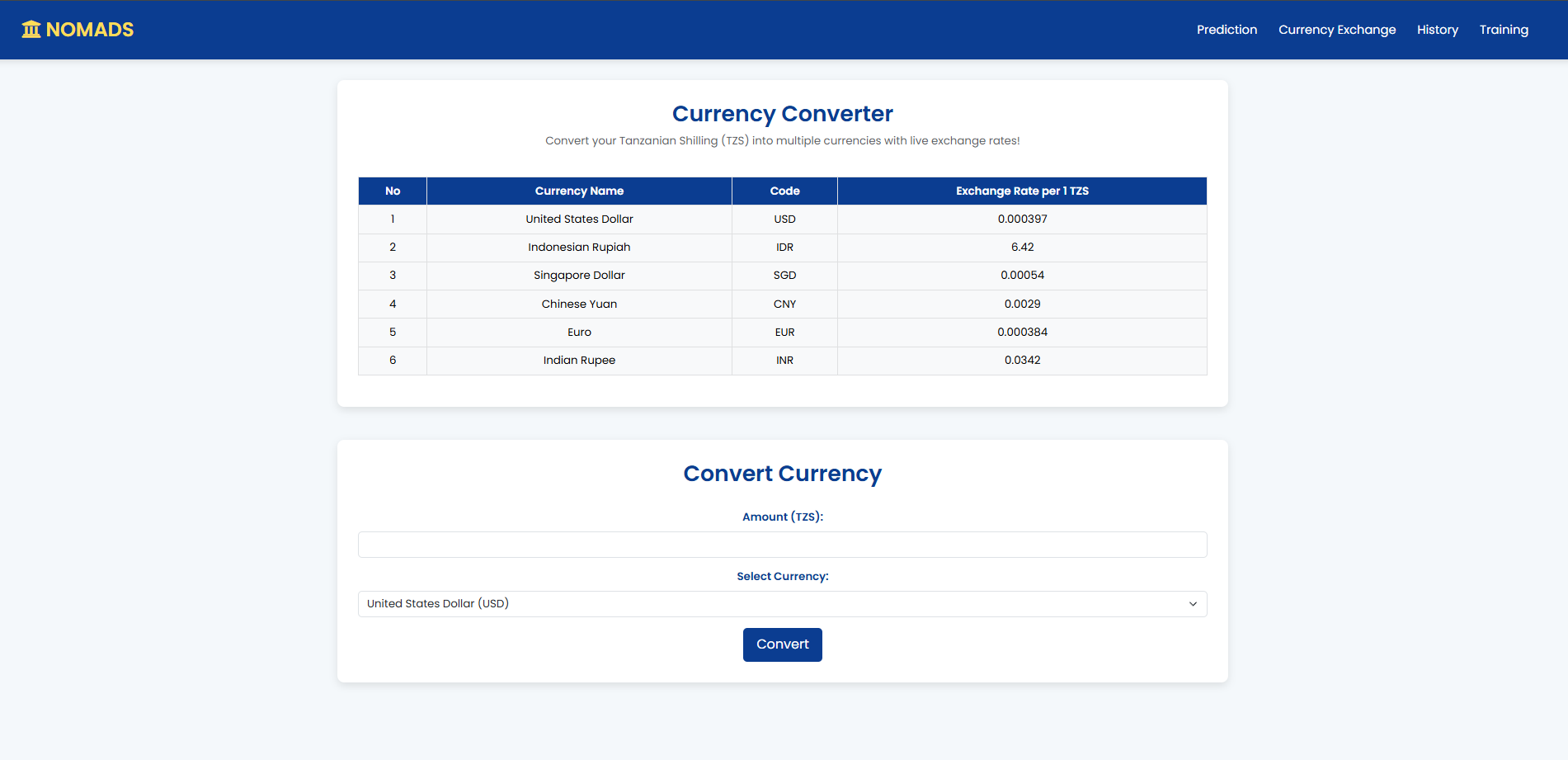
Gambar 4.27 menunjukkan faktor-faktor utama yang mempengaruhi deteksi transaksi *fraud* berdasarkan model *XGBoost Classifier*. Dari perspektif sistem deteksi *fraud*, fitur *oldbalanceOrig* (saldo awal pengirim) memiliki kontribusi terbesar yang menunjukkan bahwa perubahan saldo sebelum transaksi adalah indikator kunci untuk mendeteksi aktivitas mencurigakan. Fitur *newbalanceOrig* (saldo akhir pengirim) juga memiliki peran signifikan, karena saldo yang tidak berubah setelah transaksi bisa menjadi indikasi transaksi ilegal atau anomali. Selain itu, jumlah transaksi (*amount*) juga memainkan peran penting, mengingat transaksi *fraud* seringkali melibatkan jumlah uang yang besar atau tidak wajar. Fitur lain seperti *type* (jenis transaksi), *newbalanceDest* (saldo akhir penerima), dan *step* (waktu transaksi) juga memberikan kontribusi dalam mengidentifikasi pola *fraud.* Dari sudut pandang sistem keamanan keuangan, analisis fitur ini sangat membantu dalam mengoptimalkan model deteksi *fraud* dengan fokus pada pola perubahan saldo dan nominal transaksi yang mencurigakan sehingga dapat meningkatkan akurasi dalam menangkap anomali keuangan.

## *Deployment*



Gambar 4.28 *Page prediction*

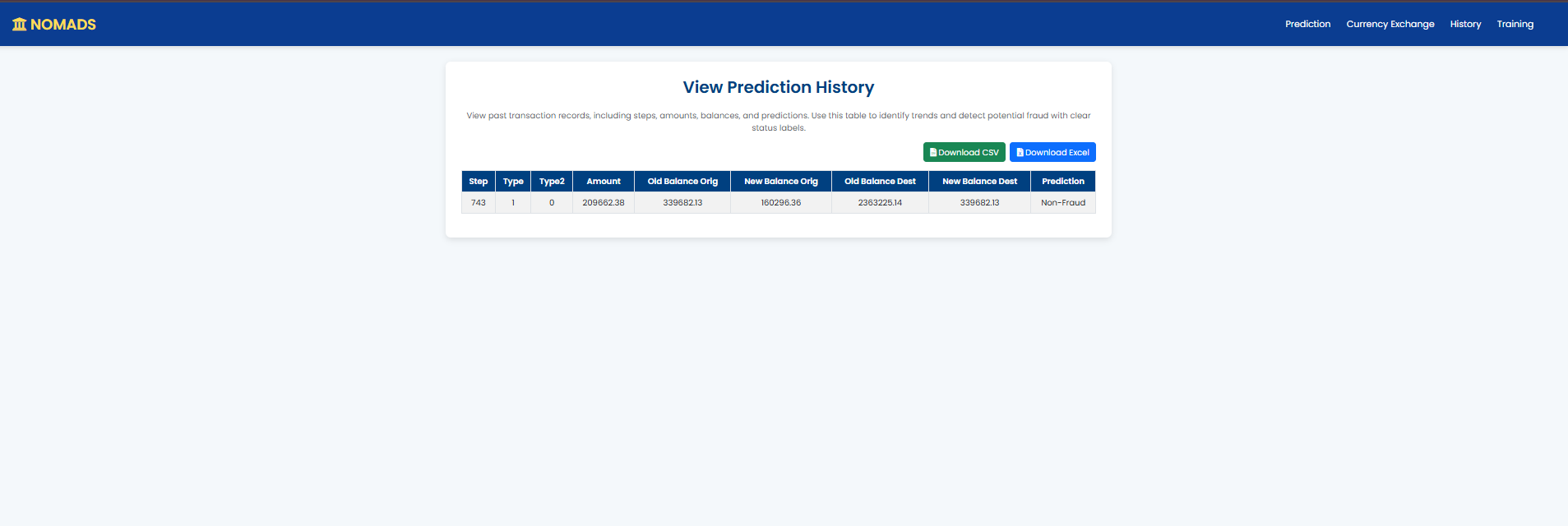
Gambar 4.28 menunjukkan Formulir *Transaction Prediction Form* pada gambar memungkinkan pengguna untuk memasukkan detail transaksi guna memprediksi kemungkinan terjadinya *fraud* menggunakan model *machine learning*. Elemen-elemen yang harus diisi mencakup *Step (Time step)* sebagai tahapan waktu transaksi, *Transaction Type* yang menunjukkan jenis transaksi, serta *Sub-Transaction Type* untuk memberikan rincian lebih spesifik. *Transaction Amount* mengacu pada nominal transaksi yang dilakukan. *Id Name Orig (Sender)* mengacu pada id pengirim. Selain itu, terdapat informasi terkait saldo, termasuk *Sender’s Old Balance* (saldo awal pengirim), *Id Name Dest (Recipient)* mengacu pada id penerima *Sender’s New Balance* (saldo akhir pengirim setelah transaksi), *Recipient’s Old Balance* (saldo awal penerima), dan *Recipient’s New Balance* (saldo akhir penerima setelah transaksi). Sementara itu, *Transaction Batch Prediction Form* memungkinkan pengguna untuk melakukan prediksi *fraud* secara massal dengan mengunggah file CSV yang berisi data beberapa transaksi. File yang diunggah harus mengikuti struktur standar seperti pada formulir individual, dan pengguna dapat melakukan prediksi dengan menekan tombol *Batch Predict*.



Gambar 4.29 *Page Currency Exchange*

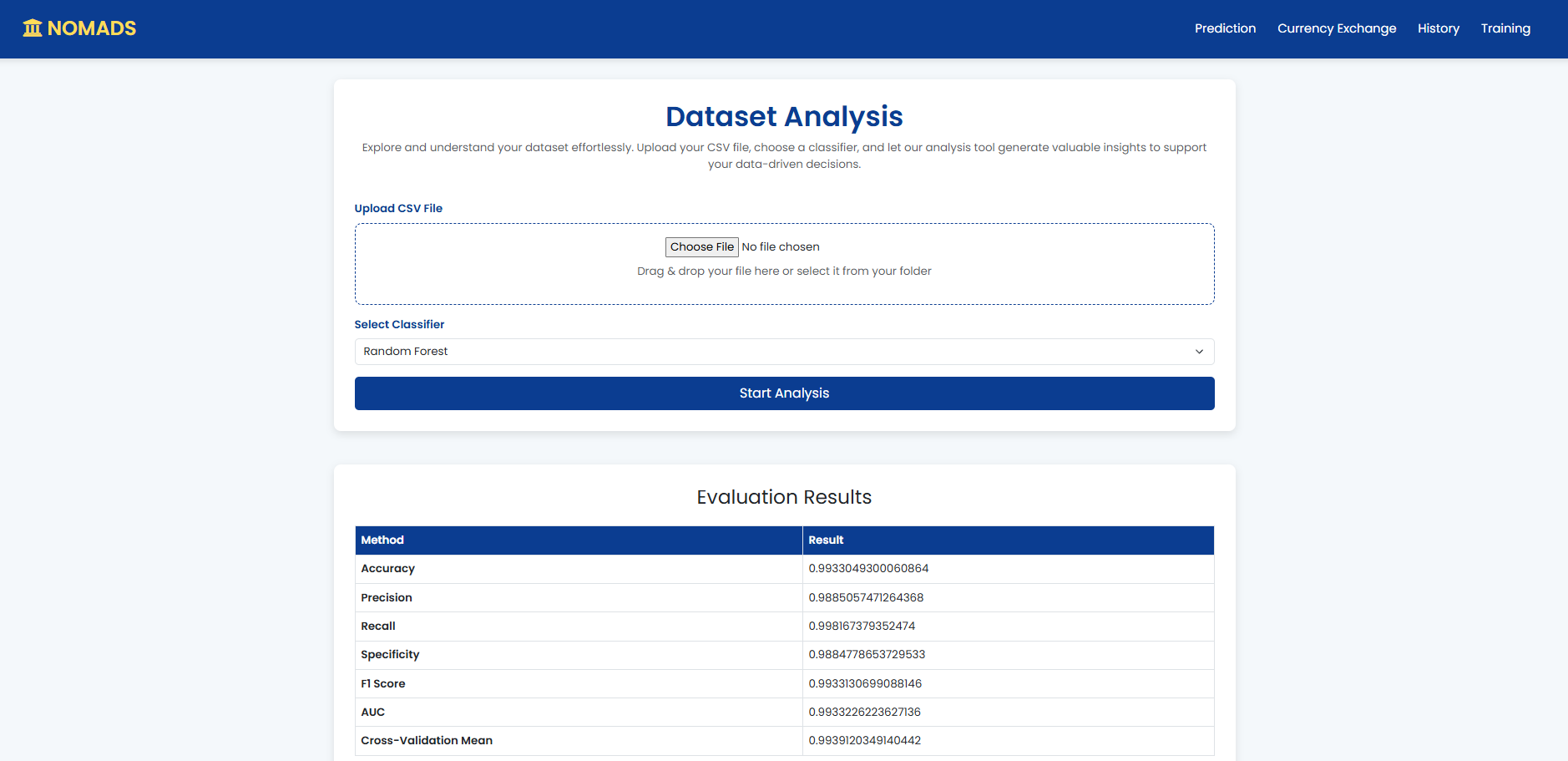
Gambar 4.29 memperlihatkan halaman konversi mata uang dengan dua fitur utama. Bagian pertama adalah tabel yang menampilkan nilai tukar dari mata uang *Tanzanian Shilling* (TZS) ke beberapa mata uang lainnya seperti *United States Dollar* (USD), *Indonesian Rupiah* (IDR), *Singapore Dollar* (SGD), *Chinese Yuan* (CNY), *Euro* (EUR), dan *Indian Rupee* (INR). Bagian kedua adalah formulir untuk menghitung nilai konversi TZS ke mata uang tertentu berdasarkan jumlah yang dimasukkan oleh pengguna.

Bagian kedua adalah formulir untuk menghitung nilai konversi mata uanag yang memungkinkan pengguna menghitung nilai konversi dari TZS ke mata uang tertentu. Pengguna cukup memasukkan jumlah TZS yang ingin dikonversi dan memilih mata uang tujuan, lalu sistem akan langsung menampilkan hasilnya berdasarkan nilai tukar terbaru.



Gambar 4.30 *Page History*

Gambar 4.30 menampilkan riwayat prediksi transaksi. Tabel berisi informasi seperti langkah *(step),* jenis transaksi *(type),*sub jenis transaksi *(type2),* jumlah *(amount),* saldo awal dan akhir pengirim *(old balance orig, new balance orig),* saldo awal dan akhir penerima *(old balance dest, new balance dest),* serta hasil prediksi *(prediction)* apakah *fraud* atau *non-fraud.* Ada opsi untuk mengunduh riwayat ini dalam format CSV atau Excel.



Gambar 4.31 *Page Training*

Gambar 4.31 digunakan untuk analisis *dataset*, di mana pengguna dapat mengunggah file CSV dan memilih model klasifikasi seperti *Decision Tree*, *Random Forest* atau *XGBoost.* Hasil evaluasi model, termasuk metrik seperti akurasi, presisi, *recall*, dan lainnya ditampilkan dalam tabel, bersama dengan *preview dataset* dan visualisasi *confusion matrix* serta distribusi data *fraud dan non-fraud.*

# KESIMPULAN DAN SARAN

## Kesimpulan

Penelitian ini berhasil menerapkan *machine learning* dalam *fraud detection* menggunakan dataset sintetis PaySim yang menunjukkan hasil yang sangat efektif dalam mengidentifikasi transaksi keuangan mencurigakan. Model *XGBoost* memberikan performa terbaik dengan akurasi 99,29%, diikuti oleh *Random Forest* dengan akurasi 99,06%, dan *Decision Tree* dengan akurasi 98,86%, di mana fitur *oldbalanceOrig* (saldo awal pengirim), *newbalanceOrig* (saldo setelah transaksi), dan *amount* (jumlah transaksi) menjadi faktor paling signifikan dalam klasifikasi transaksi *fraud*. Fitur-fitur ini menunjukkan bahwa transaksi *fraud* sering ditandai dengan perubahan saldo yang tidak wajar, dimana saldo awal pengirim sering kali tidak berubah setelah transaksi atau mengalami penurunan drastis. Selain itu, transaksi dengan jumlah besar lebih rentan terhadap aktivitas *fraud.* Teknik *Grid Search CV* berhasil mengoptimalkan parameter model, sedangkan metode *cross-validation* memastikan generalisasi yang baik. Selain itu, pengembangan sistem berbasis *machine learning* ini memungkinkan lembaga keuangan untuk secara otomatis dan cepat mendeteksi transaksi mencurigakan, meningkatkan efektivitas pengawasan, serta memperkuat kebijakan *anti-fraud* dengan mengurangi risiko kejahatan keuangan secara lebih akurat dan efisien.

## Saran

1. Penelitian lanjutan dapat mengeksplorasi penggunaan *dataset* sintetis yang lebih besar atau lebih kompleks untuk mencerminkan berbagai pola transaksi yang lebih beragam atau menggunakan data asli yang diperoleh lewat perizinan khusus
2. Kombinasi algoritma *machine learning* dengan teknik lain seperti *deep learning* atau *Graph Neural Networks (GNN)* dan kombinasiteknik *handling imbalance data* seperti *synthetic minority oversampling technique* (SMOTE)atau *adaptive synthetic sampling* (ADASYN)
3. Uji coba model dalam jangka panjang di berbagai skenario transaksi yang dinamis dapat memberikan wawasan lebih mendalam tentang keandalan model dalam kondisi nyata.

# DAFTAR REFERENSI

Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., Ayeh, S. A., & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, *6*(January), 100163. https://doi.org/10.1016/j.dajour.2023.100163

Alwadain, A., Ali, R. F., & Muneer, A. (2023). Estimating Financial Fraud through Transaction-Level Features and Machine Learning. *Mathematics*, *11*(5). https://doi.org/10.3390/math11051184

Anatolii, D., Sergii, M., Olena, D., Dmytro, T., Oksana, R., & Katerina, H. (2024). Combating Economic Crime in Ukraine in Conditions of Sustainable Development. *Revista de Gestao Social e Ambiental*, *18*(6), 1–20. https://doi.org/10.24857/rgsa.v18n6-049

Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Algorithms*, *16*(6), 1–18. https://doi.org/10.3390/a16060308

Azhari, M., Alaoui, A., Achraoui, Z., Ettaki, B., & Zerouaoui, J. (2019). Adaptation of the random forest method: Solving the problem of Pulsar Search. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3368756.3369004

Basel. (2023). Basel AML Index 2023 : 12th Public Edition Ranking money laundering and terrorist financing risks around the world. *Basel*, *10*, 1–42. https://index.baselgovernance.org/api/uploads/Basel\_AML\_Index\_2023\_12th\_Edition\_879b07b7b2.pdf

Botchey, F. E., Qin, Z., & Hughes-Lartey, K. (2020). Mobile money fraud prediction-A cross-case analysis on the efficiency of support vector machines, gradient boosted decision trees, and Naïve Bayes algorithms. *Information (Switzerland)*, *11*(8). https://doi.org/10.3390/INFO11080383

Christoforaki, M., & Beyan, O. D. (2024). Towards an ELSA Curriculum for Data Scientists. *AI (Switzerland)*, *5*(2), 504–515. https://doi.org/10.3390/ai5020025

Dasawarsa, A. R. P., Sucipto, A., Sofyan, A., Saptaji, M. B., & Rosyani, P. (2023). Analisis Penerapan Python Dalam Perhitungan Limit. *Jurnal Matematika, Algoritma Dan Sains*, *1*(1), 69–72.

de Pedro-Carracedo, J., Clemente, J., Fuentes-Jimenez, D., Cabrera-Umpiérrez, M. F., & Gonzalez-Marcos, A. P. (2023). Photoplethysmographic Signal-Diffusive Dynamics as a Mental-Stress Physiological Indicator Using Convolutional Neural Networks. *Applied Sciences (Switzerland)*, *13*(15). https://doi.org/10.3390/app13158902

Denniagi, E. (2021). Analisis Ke-Ekonomian Pemidanaan Tindak Pidana Pencucian Uang Dalam Undang-Undang Nomor 8 Tahun 2010 Tentang Pencegahan Dan Pemberantasan Tindak Pidana Pencucian Uang. *Jurnal Lex Renaissance*, *6*(2), 246–264. https://doi.org/10.20885/jlr.vol6.iss2.art3

Desai, A. S., Bandopadhyaya, A., Ashok, A., Maneesha, & Bhagat, N. (2024). Decoding characteristics of key physical properties in silver nanoparticles by attaining centroids for cytotoxicity prediction through data cleansing. *Machine Learning: Science and Technology*, *5*(2), 1–14. https://doi.org/10.1088/2632-2153/ad51cb

Elkabalawy, M., Al-Sakkaf, A., Mohammed Abdelkader, E., & Alfalah, G. (2024). CRISP-DM-Based Data-Driven Approach for Building Energy Prediction Utilizing Indoor and Environmental Factors. *Sustainability*, *16*(17), 7249. https://doi.org/10.3390/su16177249

Esoimeme, E. E. (2020). Identifying and reducing the money laundering risks posed by individuals who have been unknowingly recruited as money rules. *Journal of Money Laundering Control*, *24*(1), 201–212. https://doi.org/10.1108/JMLC-05-2020-0053

*Executive Summary 1.* (2023). *August 2022*.

Gilmour, N. (2023). Crime scripting the criminal activities of money laundering – holistically. *Journal of Money Laundering Control*, *26*(3), 594–608. https://doi.org/10.1108/JMLC-09-2020-0109

Gilmour, P. M. (2023). Reexamining the anti-money-laundering framework: a legal critique and new approach to combating money laundering. *Journal of Financial Crime*, *30*(1), 35–47. https://doi.org/10.1108/JFC-02-2022-0041

Ginting, Y. P., Gunadi, E. M., Londe, E. N., Wijaya, L., Laura, J., Fisher, J., Jawa, P., & Sanjaya, N. T. (2023). *Pencucian Uang*. *02*(10), 862–879.

Gunawan, G. (2024). Tindak Pidana Pencucian Uang dan Kejahatan Terorisme. *Edulaw: Journal of Islamic Law and Jurisprudance*, *6*(1), 21–31.

Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J., & Gao, D. (2020). Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. *Applied Sciences (Switzerland)*, *10*(18), 1–17. https://doi.org/10.3390/APP10186593

Gurcan, F., & Soylu, A. (2024). Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers*, *16*(19). https://doi.org/10.3390/cancers16193417

Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. *Information Systems Frontiers*, *25*(5), 1985–2003. https://doi.org/10.1007/s10796-022-10346-6

Handayani, P., & Charis Fauzan, A. (2024). KLIK: Kajian Ilmiah Informatika dan Komputer Machine Learning Klasifikasi Status Gizi Balita Menggunakan Algoritma Random Forest. *Media Online)*, *4*(6), 3064–3072. https://doi.org/10.30865/klik.v4i6.1909

Hasan, M. E., & Wagler, A. (2024). New Convolutional Neural Network and Graph Convolutional Network-Based Architecture for AI Applications in Alzheimer’s Disease and Dementia-Stage Classification. *AI (Switzerland)*, *5*(1), 342–363. https://doi.org/10.3390/ai5010017

Hernandez Aros, L., Bustamante Molano, L. X., Gutierrez-Portela, F., Moreno Hernandez, J. J., & Rodríguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: a literature review. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-03606-0

Irwan, M., Rani, A., Zolkaflil, S., Nazatul, S., Syed, F., & Nazri, M. (2022). the Money Mule Red Flags in Anti-Money Laundering Transaction Monitoring Investigation. *International Journal of Business and Economy (IJBEC)*, *4*(1), 150–163. http://myjms.mohe.gov.my/index.php/ijbecJournalwebsite:http://myjms.mohe.gov.my/index.php/ijbec

Jensen, R. I. T., Ferwerda, J., Jørgensen, K. S., Jensen, E. R., Borg, M., Krogh, M. P., Jensen, J. B., & Iosifidis, A. (2023). A synthetic data set to benchmark anti-money laundering methods. *Scientific Data*, *10*(1), 1–11. https://doi.org/10.1038/s41597-023-02569-2

Kalboussi, E., Ndhaief, N., & Rezg, N. (2024). Last-Mile Optimization Using Neural Networks. *Applied Sciences (Switzerland)*, *14*(2). https://doi.org/10.3390/app14020787

Kannan, R., Reddiar, Y., Ramakrishnan, K., Eastaff, M. S., & Ramesh, S. (2022). Job characteristics of a Malaysian bank’s anti-money laundering system and its employees’ job satisfaction. *F1000Research*, *10*, 1–14. https://doi.org/10.12688/f1000research.73234.2

Karim, M. R., Hermsen, F., Chala, S. A., De Perthuis, P., & Mandal, A. (2024). Scalable Semi-Supervised Graph Learning Techniques for Anti Money Laundering. *IEEE Access*, *12*, 50012–50029. https://doi.org/10.1109/ACCESS.2024.3383784

L. K. Vishwamitra, V. D. A. (2024). Machine Learning Models for Fraud Detection: A Comprehensive Review and Empirical Analysis. *Journal of Electrical Systems*, *20*(3s), 1138–1149. https://doi.org/10.52783/jes.1427

Laowo, Sebastian, Y. (2022). Kajian Hukum Tindak Pidana Pencucian Uang (Money Laundering). *Jurnal Panah Keadilan*, *1*(1), 86.

Liu, Y., Pang, Z., Karlsson, M., & Gong, S. (2020). Anomaly detection based on machine learning in IoT-based vertical plant wall for indoor climate control. *Building and Environment*, *183*, 107212. https://doi.org/10.1016/j.buildenv.2020.107212

Lokanan, M. E. (2023). Predicting mobile money transaction fraud using machine learning algorithms. *Applied AI Letters*, *4*(2), 1–16. https://doi.org/10.1002/ail2.85

Lopez-Rojas, E. A., & Barneaud, C. (2019). Advantages of the PaySim Simulator for Improving Financial Fraud Controls. *Advances in Intelligent Systems and Computing*, *998*, 727–736. https://doi.org/10.1007/978-3-030-22868-2\_51

Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). Paysim: A financial mobile money simulator for fraud detection. *28th European Modeling and Simulation Symposium, EMSS 2016*, *c*, 249–255.

Lubis, M. R. D. (2023). Peran Otoritas Jasa Keuangan (OJK) dalam Meminimalisir Investasi Bodong yang Dipromosikan Secara Online. *Mutiara: Multidiciplinary Scientifict Journal*, *1*(7), 295–305. https://doi.org/10.57185/mutiara.v1i7.49

Marmolejo-Ramos, F., Ospina, R., García-Ceja, E., & Correa, J. C. (2022). Ingredients for Responsible Machine Learning: A Commented Review of The Hitchhiker’s Guide to Responsible Machine Learning. *Journal of Statistical Theory and Applications*, *21*(4), 175–185. https://doi.org/10.1007/s44199-022-00048-y

Maulida Surbakti, N., Talia, A., Br Perangin-Angin, C., Olivia Nainggolan, D., Devi Friskauly, N., & Ruth Br Tumorang, S. (2024). Penggunaan Bahasa Pemrograman Python dalam Pembelajaran KalkulusFungsi Dua Variabel. *Kebumian Dan Angkasa*, *2*(3), 98–107. https://doi.org/10.62383/algoritma.v2i3.67

Mbiva, S. M., & Correa, F. M. (2024). Machine Learning to Enhance the Detection of Terrorist Financing and Suspicious Transactions in Migrant Remittances. *Journal of Risk and Financial Management*, *17*(5). https://doi.org/10.3390/jrfm17050181

Muraru, M. M., Simó, Z., & Iantovics, L. B. (2024). Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods. *Applied Sciences (Switzerland)*, *14*(22), 1–23. https://doi.org/10.3390/app142210085

Nafisah Nurul Hakim. (2020). Implementasi Machine Learning pada Sistem Prediksi Kejadian dan Lokasi Patah Rel Kereta Api di Indonesia. *Jurnal Sistem Cerdas*, *3*(1), 25–35. https://doi.org/10.37396/jsc.v3i1.58

Naveed, N., Munawar, S., & Usman, A. (2023). Intelligent Anti-Money Laundering Fraud Control Using Graph-Based Machine Learning Model for the Financial Domain. *Journal of Cases on Information Technology*, *25*(1), 1–21. https://doi.org/10.4018/JCIT.316665

Niakšu, O. (2015). CRISP Data Mining Methodology Extension for Medical Domain. *Baltic J. Modern Computing*, *3*(2), 92–109.

Nobel, S. M. N., Sultana, S., Singha, S. P., Chaki, S., Mahi, M. J. N., Jan, T., Barros, A., & Whaiduzzaman, M. (2024). Unmasking Banking Fraud: Unleashing the Power of Machine Learning and Explainable AI (XAI) on Imbalanced Data. *Information (Switzerland)*, *15*(6), 1–23. https://doi.org/10.3390/info15060298

Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., Ziegenbein, C., Singhal, A., Choi, S. W., & Tewari, M. (2024). Data Preprocessing Techniques for Artificial Intelligence (AI)/Machine Learning (ML)-Readiness: Systematic Review of Wearable Sensor Data in Cancer Care (Preprint). *JMIR MHealth and UHealth*, *12*. https://doi.org/10.2196/59587

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, *6*, 1–43. https://doi.org/10.7717/PEERJ-CS.267

Prakash, B., Baboo, G. K., & Baths, V. (2021). A novel approach to learning models on eeg data using graph theory features—a comparative study. *Big Data and Cognitive Computing*, *5*(3). https://doi.org/10.3390/bdcc5030039

Putra, A., & Toba, H. (2020). Pengembangan Gudang Data Pendukung Analisis Tren Penyewaan Peralatan Katering dengan Algoritma Apriori. *Jurnal of Information System and Technology*, *1*(1), 5–14.

Raghavendran, K. R., & Elragal, A. (2023). Low-Code Machine Learning Platforms: A Fastlane to Digitalization. *Informatics*, *10*(2). https://doi.org/10.3390/informatics10020050

Rahman, S., Sembiring, A., Siregar, D., Khair, H., Gusti Prahmana, I., Puspadini, R., & Zen, M. (2023). Python : Dasar Dan Pemrograman Berorientasi Objek. In *Penerbit Tahta Media*.

Reznik, O., Utkina, M., & Bondarenko, O. (2023). Financial intelligence (monitoring) as an effective way in the field of combating money laundering. *Journal of Money Laundering Control*, *26*(1), 94–105. https://doi.org/10.1108/JMLC-09-2021-0102

Richter-Laskowska, M., Kurpas, M., & Maśka, M. (2022). *A learning by confusion approach to characterize phase transitions*. https://doi.org/10.1103/PhysRevE.108.024113

Rose, K. J. (2020). Disclosing anti-money launderers through CSR regulation – a new way to combat money laundering. *Journal of Money Laundering Control*, *23*(1), 11–25. https://doi.org/10.1108/JMLC-07-2019-0062

Shah, S. M. A. (2023). A qualitative study exploring challenges in money laundering investigations. *Journal of Money Laundering Control*, *26*(3), 523–534. https://doi.org/10.1108/JMLC-09-2019-0070

Sultana, S. (2020). Role of financial intelligence unit (FIU) in anti-money laundering quest: Comparison between FIUs of Bangladesh and India. *Journal of Money Laundering Control*, *23*(4), 931–947. https://doi.org/10.1108/JMLC-01-2020-0003

Sutoyo, R., Warnars, H. L. H. S., Isa, S. M., & Budiharto, W. (2023). Indonesian Twitter Emotion Recognition Model using Feature Engineering. *International Journal of Advanced Computer Science and Applications*, *14*(12), 1057–1065. https://doi.org/10.14569/IJACSA.2023.01412108

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. *Electronics (Switzerland)*, *12*(8). https://doi.org/10.3390/electronics12081789

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-023-00857-7

Yeni, A., & Hartati, S. (2020). Studi Literatur: Stimulasi kemampuan anak mengenal huruf melalui permainan menguraikan kata di taman kanak-kanak Alwidjar Padang. *Jurnal Pendidikan Tambusai*, *4*, 608–616. http://download.garuda.kemdikbud.go.id/article.php?article=1729966&val=13365&title=STUDI LITERATUR STIMULASI KEMAMPUAN ANAK MENGENAL HURUF MELALUI PERMAINAN MENGURAIKAN KATA DI TAMAN KANAK-KANAK ALWIDJAR PADANG

Zhu, W., Qiu, R., & Fu, Y. (2024). *Comparative Study on the Performance of Categorical Variable Encoders in Classification and Regression Tasks*. 1–19. http://arxiv.org/abs/2401.09682