

**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 2024**

# 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 : 99 Desember 9999

# 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 di hadapan 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 : ....nama... ( )

Penguji II : ....nama... ( )

Ditetapkan di : Jakarta

Ketua Program Studi : M. Bahrul Ulum, S.Kom, M.Kom

Tanggal : 99 Desember 9999

# 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:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

beserta perangkat yang ada (apabila diperlukan). Dengan Hak Bebas Royalti Noneksklusif 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 : 99 Desember 9999

Yang menyatakan

(.........Nama Penulis.........)

# KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa yang telah memberikan kesehatan, karunia serta hidayah-nya sehingga penulis dapat menyelesaikan Seminar Proposal 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 Seminar Proposal 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 proposal skripsi ini, terutama kepada:

1. Allah SWT atas segala Rahmat dan karunia-nya.
2. Ibu Dr. Vitri Tundjungsari, ST, M.Sc, selaku Dekan Fakultas Ilmu Komputer Universitas Esa Unggul.
3. Bapak M. Bahrul Ulum, S.Kom, M.Kom selaku Kepala Program Studi Teknik Informatika dan selaku Dosen Pembimbing yang telah memberikan bimbingan, arahan, dorongan, dan semangat kepada penulis. dalam penyusunan Proposal Tugas Akhir ini.
4. Bapak Jefry Sunupurwa Asri, S.Kom, M.Kom selaku selaku Dosen Pembimbing Akademik saya.
5. Bapak Romy Ardianto, M.Sc selaku selaku Mentor dan Data Scientist Kementrian Kelautan dan Perikanan Republik Indonesia.
6. Kedua Orang Tua dan keluarga penulis yang selalu memberikan doa, dukungan moril, materiil dan selalu memberikan motivasi yang luar biasa.
7. Seluruh Dosen dan Staff Fakultas Ilmu Komputer yang tidak bisa disebutkan satu persatu namanya.
8. Sahabat serta teman-teman seperjuangan terutama Ananda Dwi Rizkyta, Shefia Anggraeni, dan Bonifasius Dandy Krisnanda yang selalu memberikan doa, dukungan, saran, dan bantuannya dalam menyelesaikan Proposal Tugas Akhir ini.

Penulis menyadari bahwa proyek seminar proposal 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 Seminar Proposal ini dapat bermanfaat bagi semua pihak, khususnya bagi penulis dan para pembaca pada umumnya.

Jakarta, 23 Januari 7839

Imam Mulyana

(2020-0801-090)

# ABSTRAK

Judul :

Nama :

Program Studi :

Abstrak berisi ringkasan hal-hal yang telah dikerjakan pada saat penelitian. Abstrak proposal harus memuat secara ringkas kondisi/situasi problem/prospek yang dihadapi/diungkapkan, orientasi solusi/manfaat yang menjadi gagasan, apa yang dikerjakan pada saat tugas akhir, kerangka berpikir (dg perspektif) dan metode/teknik dalam mencapai solusi/manfaat yang digunakan, sumber/referensi akademis (konsep/konteks), hasil yang dicapai dan evaluasi kajian, serta kesimpulan dari hasil kajian tugas akhir.

Kata kunci :

**ABSTRACT**

Title :

Name :

Study Program :

.............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. .............. ..............

Keywords :

# DAFTAR ISI

Halaman

HALAMAN PERNYATAAN KEASLIAN ii

HALAMAN PENGESAHAN TUGAS AKHIR iii

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH iv

KATA PENGANTAR vi

ABSTRAK viii

DAFTAR ISI x

DAFTAR TABEL xii

DAFTAR GAMBAR xiii

DAFTAR SIMBOL xiv

BAB 1 PENDAHULUAN 15

1.1 Latar Belakang 15

1.2 Identifikasi Masalah 17

1.3 Tujuan Tugas Akhir 18

1.4 Manfaat Tugas Akhir 18

1.5 Lingkup Tugas Akhir 19

1.6 Kerangka Berpikir 19

1.7 Sistematika Penulisan Tugas Akhir 20

BAB 2 TINJAUAN PUSTAKA 22

2.1 Penelitian Sebelumnya 22

*2.2* *Money Laundering* 27

*2.3* *Anti Money Laundering* 29

*2.4* *Machine Learning* 33

*2.5* *Fraud Detection* 34

2.6 Data Sintetis 34

*2.7* CRIPS-DM *(Cross-Industry Standard Process for Data Mining)* 35

*2.7.1* *Business Understanding* 36

*2.7.2* *Data Understanding* 36

*2.7.3* *Data Preparation atau Data Preprocessing* 36

*2.7.4* *Data Modeling* 41

*2.7.5* *Data Evaluation* 44

*2.7.6* *Deployment* 46

2.8 Tools 46

2.8.1 Python 46

2.8.2 Numpy 46

2.8.3 Pandas 47

2.8.4 Matplotlib 47

2.8.5 Scikit Learn 47

BAB 3 METODE 49

3.1 Rencana Penelitian 49

3.2 Objek Penelitian 53

3.3 Teknik Pengumpulan Data 53

3.4 Pendekatan Penelitian 57

3.5 Instrumen Penelitian 57

BAB 4 HASIL DAN PEMBAHASAN 59

4.1 Data Hasil Penelitian 59

4.2 Pembahasan Hasil Penelitian 59

4.3 Analisis Data 59

4.4 Perbandingan dengan Penelitian Lain 59

BAB 5 KESIMPULAN DAN SARAN 60

5.1 Kesimpulan 60

5.2 Saran 60

DAFTAR REFERENSI 61

Lampiran 1 Daftar Riwayat Hidup 65

Lampiran 2 Asdf 66

Lampiran 3 Xyzxyz 67

# DAFTAR TABEL

Halaman

[Tabel 1 Data varibel PaySim 54](#_Toc182834732)

# DAFTAR GAMBAR

Halaman

[Gambar 1 Sebaran Negara Risiko AML 15](#_Toc182835606)

[Gambar 2 Kerang Berpikir Penelitian 20](#_Toc182835607)

[Gambar 3 *Money Laundering Cycle* 28](#_Toc182835608)

[Gambar 4 Tahapan penanganan AML 30](#_Toc182835609)

[Gambar 5 Tahapan Metode CRISP-DM 35](#_Toc182835610)

[Gambar 6 Proses *Preprocessing* 37](#_Toc182835611)

[Gambar 7 Proses *Data Modeling* 42](#_Toc182835612)

[Gambar 8 *Confusion Matrik* dan Rumus Evaluasi 45](#_Toc182835613)

[Gambar 9 Kurva ROC AUC 46](#_Toc182835614)

[Gambar 10 Proses pengembangan *machine learning* 49](#_Toc182835615)

[Gambar 11 Pembagian dataset 51](#_Toc182835616)

[Gambar 12 Cara kerja *decision tree* 51](#_Toc182835617)

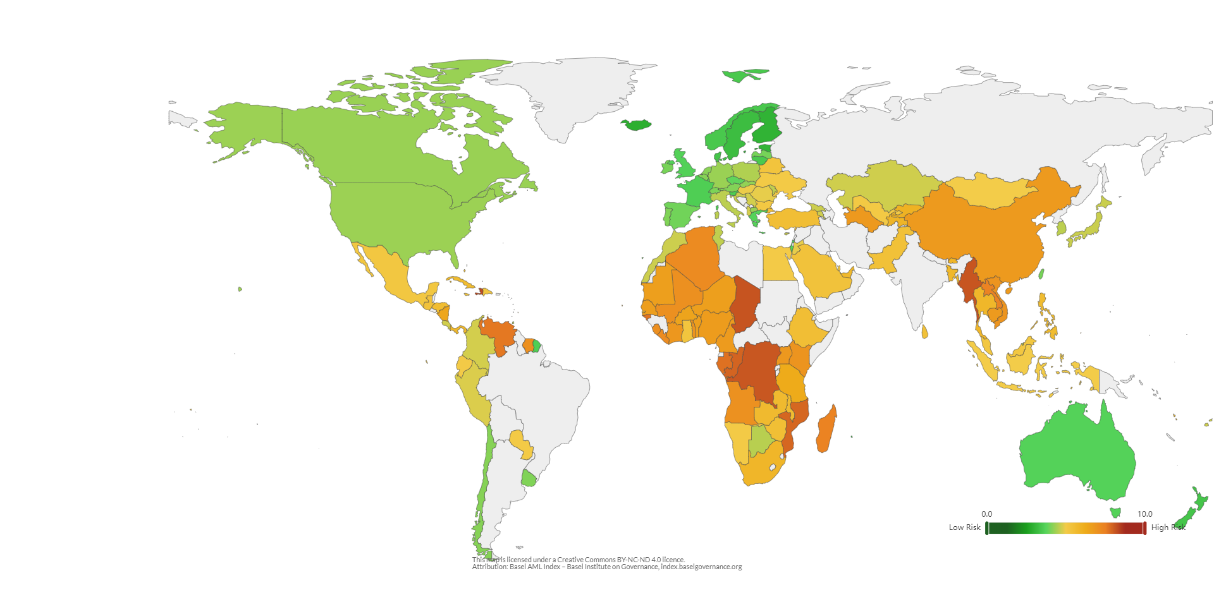
[Gambar 13 Cara kerja *random forest* 52](#_Toc182835618)

# DAFTAR SIMBOL

# 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 dan Indonesia berada pada rank 85 dengan skor 5.01.



Gambar 1 Sebaran Negara Risiko AML

(Sumber: www. index.baselgovernance.org)

Sedangkan menurut *Financial Action Task Force* (FATF) (*Executive Summary 1.*, 2023), pada laporan tahun 2023 Indonesia 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 laundry* 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 & 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 memberi tahu 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 penelitian ini menggunakan algoritma *machine learning,* seperti *logistic regression, random forest,* dan *stochastic gradient descent* dengan dataset dari PaySim. Hasilnya menunjukkan bahwa *random forest* adalah model paling akurat dengan tingkat akurasi 89% (Lokanan, 2023). Selanjutnya Naveed, dimana penelitian ini menggunakan model *Graph Convolutional Networks* (GCN) dengan metode *semi-supervised learning* untuk mendeteksi *fraud detection* menggunakan data sintetis dari AMLSim. Hasilnya model GCN mampu mencapai akurasi 77-79% dalam mendeteksi *fraud detection* padatransaksi keuangan yang mencurigakan (Naveed et al., 2023). Penelitian dari Hajek menunjukan hasil yang paling optimal, dimana penelitian ini menemukan bahwa metode *XGBoost* mampu mencapai tingkat akurasi *fraud detection* sebesar 99,98% dengan menggunakan data sintetis PaySim (Hajek et al., 2023).

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 dirumuskan masalah yang dibahas sebagai berikut:

1. Data keuangan sering kali bersifat rahasia dan sulit diakses sehingga menghambat penelitian dan pengembangan model *fraud detection* atau transaksi keuangan mencurigakan berbasis *machine learning*;
2. Metode tradisional kurang efektif dalam menilai risiko dan mendeteksi *fraud* atau transaksi keuangan mencurigakan secara cepat karena kompleksitas data transaksi keuangan dan proses yang dilakukan.
3. Pemilihan model *machine learning* yang tepat menjadi hal yang krusial dalam memberikan hasil yang akurat dan sesuai dengan kebutuhan dalam pengembangan *fraud detection.*

## Tujuan Tugas Akhir

Dalam penelitian ini terdapat beberapa tujuan yang ingin dicapai pada penelitian ini 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.
3. Mengevaluasi efektivitas model *machine learning* dalam mendeteksi aktivitas *fraud* secara cepat dan akurat dengan membandingkan hasil model.
4. 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.

## Manfaat Tugas Akhir

Dengan adanya penelitian ini di harapkan 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.

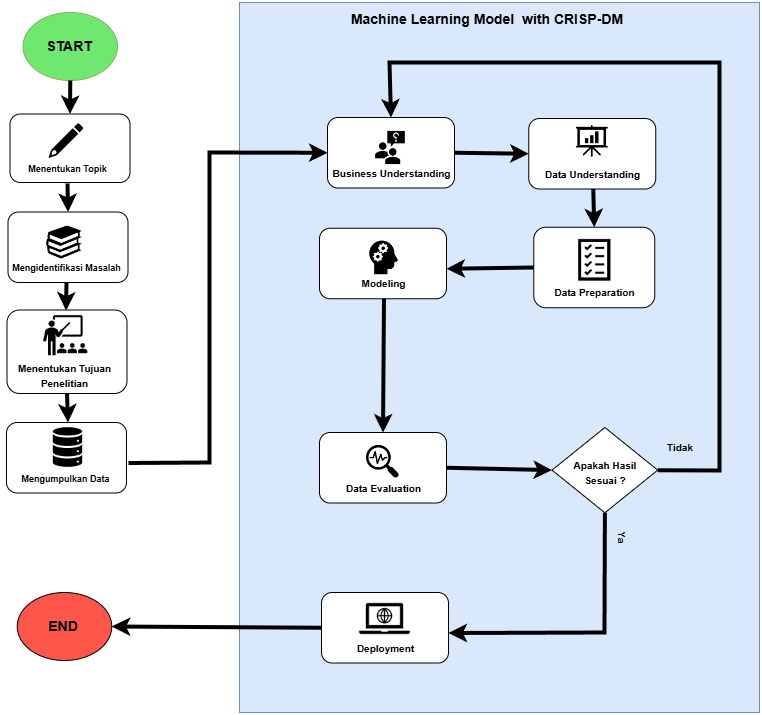
## Lingkup Tugas Akhir

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



Gambar 2 Kerang 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, rumusan, batasan, tujuan 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 dibidang 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:

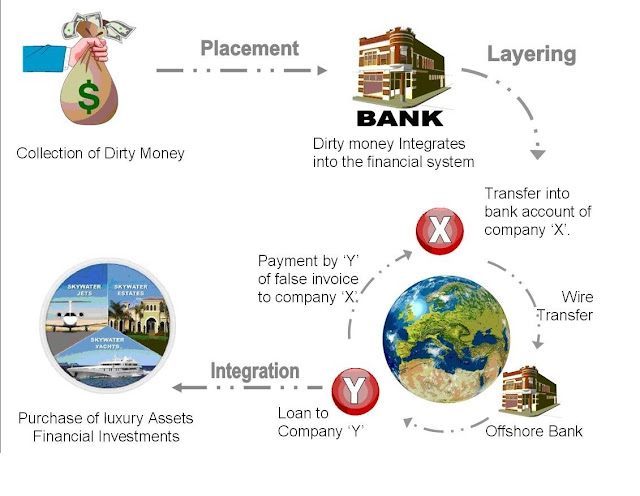
Tabel 1 Penelitian Sebelumnya

|  |  |  |  |
| --- | --- | --- | --- |
| 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 *stochastic gradient descent* untuk mendeteksi penipuan dalam transaksi uang elektronik dengan menggunakan data yang dihasilkan dari PaySim dan teknik SMOTE-ENN digunakan untuk menangani ketidakseimbangan kelas. Hasilnya menunjukkan bahwa *random forest* adalah model paling akurat dengan tingkat akurasi 89% dan MCC 0,78 sehingga menjadikannya model yang efektif dalam mendeteksi aktivitas penipuan 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 penipuan. Hasilnya menunjukkan bahwa *XGBoost* memberikan kinerja terbaik dengan *accuracy* 99,9% dan *F1-score* 0,999 dalam mendeteksi transaksi penipuan. |
| 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 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 ketidakseimbangan 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 penipuan keuangan. |
| 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 penipuan, 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%. Di sisi 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 penipuan pada transaksi uang 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 mentransfer 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 berbasis agen untuk menghasilkan dataset sintetis dari data keuangan anonim yang menyerupai dinamika dan statistik transaksi asli, kemudian digunakan untuk mempelajari skenario penipuan dan mengevaluasi kontrol penipuan. Hasil simulasi menunjukkan bahwa kontrol penipuan yang lebih ketat mengurangi penipuan yang tidak terdeteksi atau *False Negatives*, tetapi meningkatkan jumlah *False Positives*, sementara kontrol yang lebih longgar memiliki efek sebaliknya |
| 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 mobileasli, 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 3 *Money Laundering Cycle*

(Sumber: www. antikorupsi.org)

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).

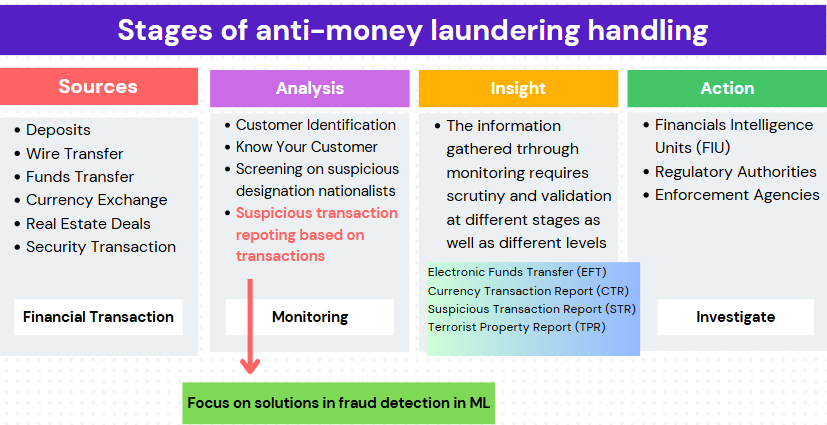
1. *Integration*

*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 4 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 sering digunakan untuk memindahkan uang secara global tanpa menarik perhatian. *Funds Transfer* (transfer dana) mencakup pemindahan dana antar akun atau lembaga keuangan, 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.

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:

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 mentransfer 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. Praktik ini, yang dikenal sebagai *structuring* atau *smurfing,* 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, mentransfer, 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, pentransferan, 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 menghasikan 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 peosesnya, 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).

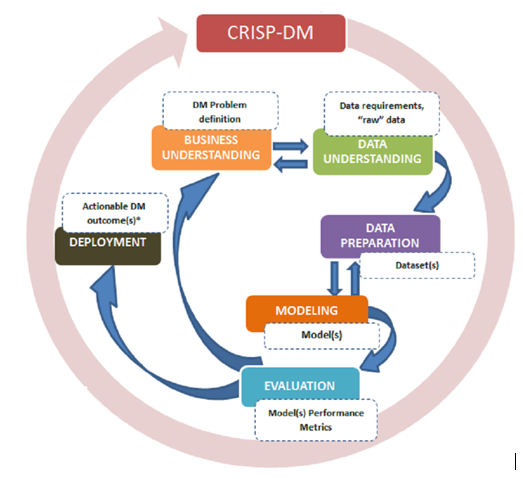
## Data Sintetis

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 diwilayah 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 5 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 dilaku kan 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 memproleh informasi rinci. Tahapan EDA sering digunakan untuk menemukan pola melalui visualisasi dan analisis statistik dengan mengidentifikasi tren 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 6 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. Proses ini juga membantu menghindari biaya tambahan akibat kesalahan dan memastikan kepatuhan terhadap standar dan peraturan yang berlaku.

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.

* 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.

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.

* 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 proposional dan akurat.

1. *Linearity*

*Linearity* mengacu pada adanya hubungan langsung dan proporsional antara fitur-fitur dalam dataset. Dimana perubahan pada satu fitur akan menghasilkan perubahan yang sebanding pada fitur lainnya. Dalam konteks analisis data hubungan linier berarti bahwa fitur-fitur dapat diprediksi atau dijelaskan dengan menggunakan persamaan garis lurus. Ini penting untuk beberapa algoritma dan model yang mengandalkan asumsi bahwa hubungan antar fitur mengikuti pola linier untuk membuat prediksi atau analisis yang akurat.

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 matrix 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.*

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.

### *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 seringkali 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).

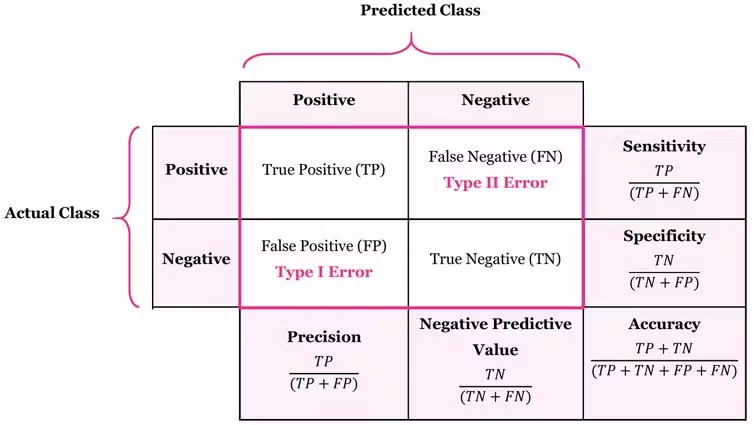
Selanjutnya tahapan *data modeling* akan diproses melalui skema berikut:

Gambar 7 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 besrsifat 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. 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 optimisasi 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.
   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. Proses validasi biasanya dilakukan dengan membagi dataset menjadi data latih *(train),* data validasi *(validation),* dan data uji *(test).* Metrik evaluasi seperti *accuracy, precision, recall,* dan *AUC (Area Under the Curve)* digunakan untuk mengukur kinerja model. Validasi kinerja membantu dalam mengidentifikasi apakah model *overfitting, underfitting,* atau sudah optimal, serta memberikan gambaran tentang seberapa baik model akan bekerja dalam situasi dunia nyata.

### *Data Evaluation*

*Confusion matrix* adalah salah satu metode yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan mengukur seberapa baik model tersebut memprediksi kelas data.

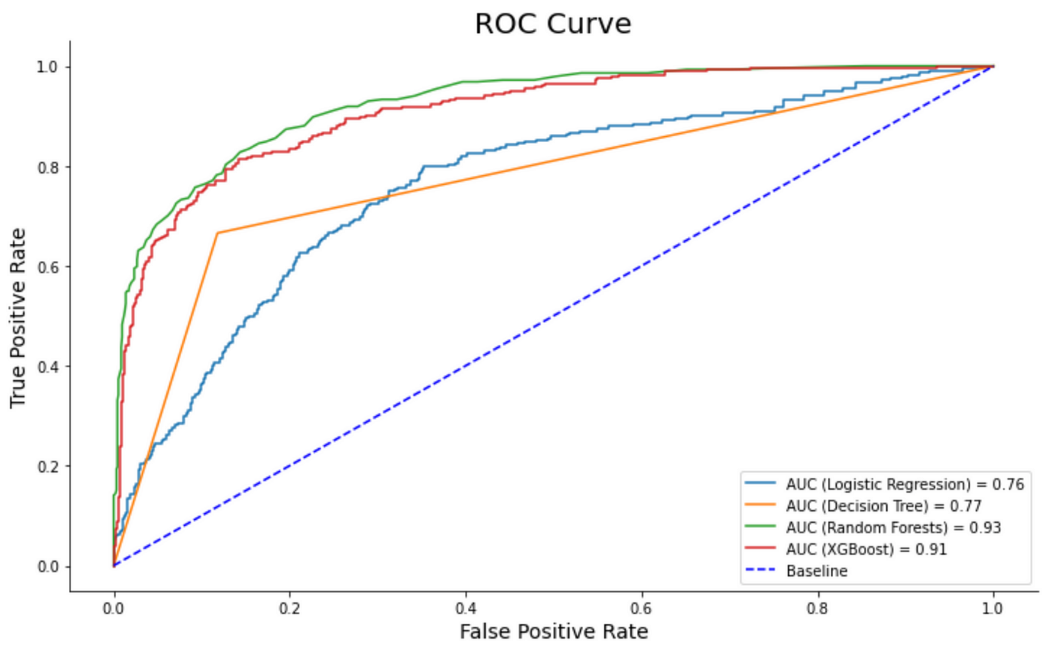


Gambar 8 *Confusion Matrik* dan Rumus Evaluasi

(Sumber: [www.researchgate.net](http://www.researchgate.net))

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).

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 presisi dan *recall* menjadi satu metrik harmonis.



Gambar 9 Kurva ROC AUC

(Sumber: [www.towardsdatascience.com](http://www.towardsdatascience.com))

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 positif palsu dan sensitivitas *(recall)* (de Pedro-Carracedo et al., 2023).

### *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 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).

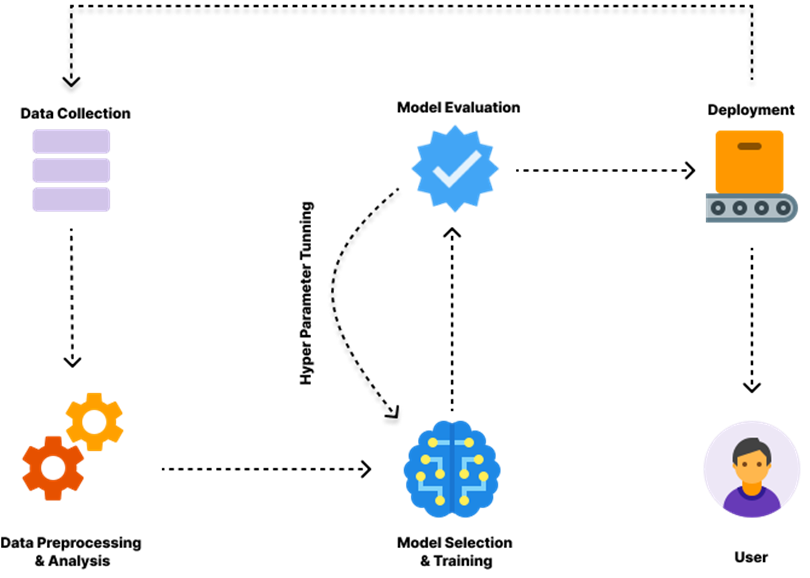
### Scikit Learn

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 10 Proses Pengembangan M*achine Learning*

Secara garis besar, pengembangan model *machine learning* akan melewati *data collection* untuk menumpulkan data yang akan digunakan*, data preprocessing* dan *analysis* untuk memahami atau mengekplor 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*

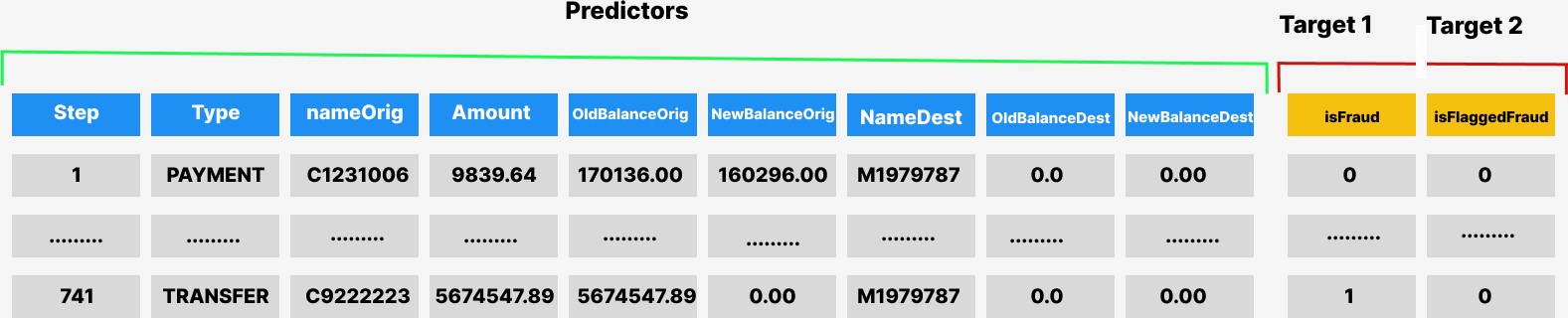
*Exploratory Data Analysis* (EDA) meliputi analisis data sintetis untuk memahami struktur data, pola transaksi, dan karakteristik umum dari transaksi *fraud*. EDA dapat membantu mengenali distribusi data, hubungan antar variabel, mengidentifikasi nilai yang hilang, outlier, atau inkonsistensi, dan memungkinkan 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),* data validasi *(validation),* dan data uji *(test).*

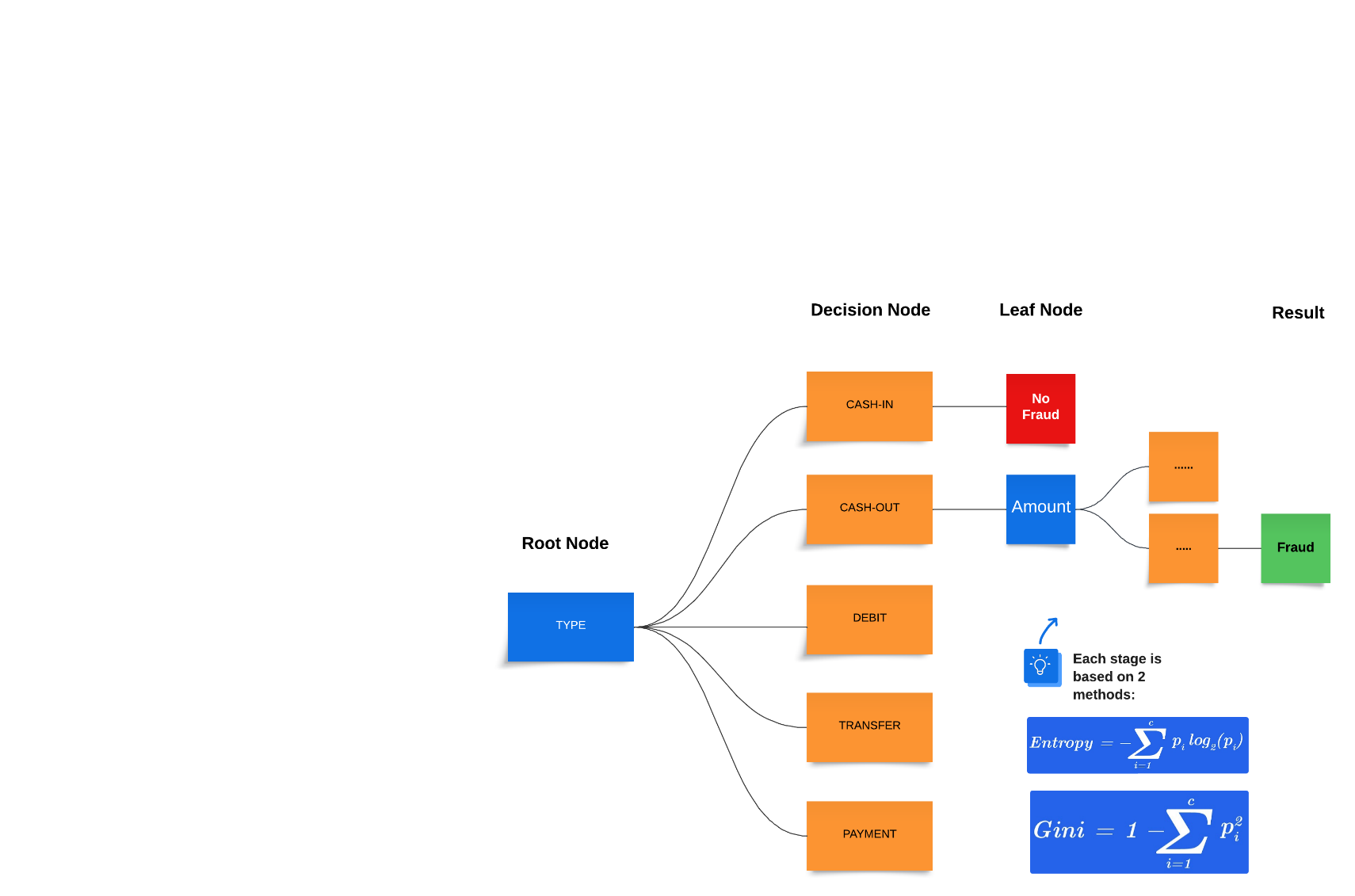
1. *Data Modeling*

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 11 Pembagian Dataset

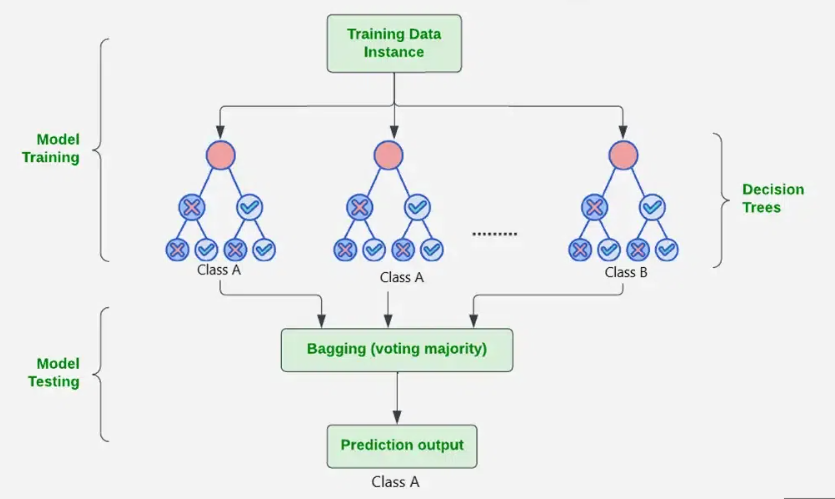
Proses diawali dengan mengimpor seluruh dataset yang akan digunakan untuk melatih model. Dataset ini berisi berbagai atribut atau karakteristik yang digunakan sebagai input atau *predictors*, serta label atau target yang menjadi hasil yang ingin diprediksi. *Decision Tree* akan membagi dataset ini secara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



Gambar 12 Cara kerja *Decision Tree*

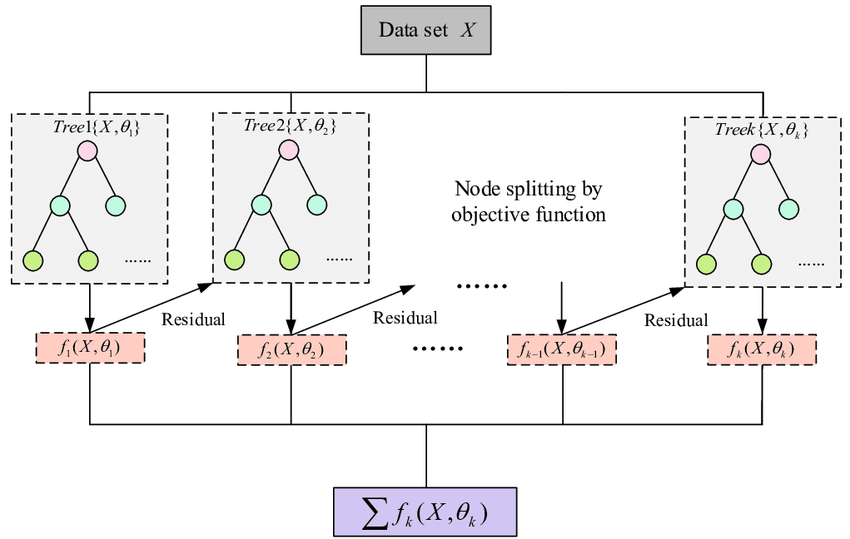
Pada setiap node dalam *Decision Tree*, algoritma akan memilih fitur yang paling relevan untuk memisahkan data ke dalam subset-subset yang lebih kecil. Pemilihan fitur ini didasarkan pada kriteria tertentu seperti *Gini Impurity* dan *Entropy* yang digunakan untuk mengukur seberapa baik fitur tersebut memisahkan data. 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.



Gambar 13 Cara Kerja *Random Forest*

(Sumber: www.medium.com)



Gambar 14 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*, 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,* 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 2 Data variabel PaySim

|  |  |
| --- | --- |
| Feature name | Description |
| step | Skala satuan waktu dengan dunia nyata. 1 step adalah 1 jam dunia nyata. |
| type | Tipe transaksi: CASH-IN, CASH-OUT, DEBIT, PAYMENT dan TRANSFER |
| amount | jumlah transaksi dalam mata uang lokal |
| nameOrig | Pelanggan atau orang yang memulai transaksi |
| oldbalanceOrg | saldo awal sebelum transaksi |
| newbalanceOrg | saldo pelanggan atau pengguna setelah transaksi. |
| nameDest | ID penerima transaksi. |
| oldbalanceDest | saldo penerima awal sebelum transaksi. |
| newbalanceDest | saldo penerima setelah transaksi. |
| isFraud | mengkategorikan transaksi, jika penipuan (1) dan bukan penipuan (0) |
| isFlaggedFraud | mengkategorikan upaya ilegal, jika mentransfer 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 mentransfer 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 mentransfer 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

{*Menjelaskan hasil penelitian (TA dan seting (cara) evaluasi), termasuk prosedur yang dijalankan, tolok ukur yang dipakai dan indikator keberhasilannya. Dari hasil evaluasi dapat diperlihatkan ketercapaian pelaksanaan TA (solusi yang ditawarkan dapat menyelesaikan persoalan yang didefinisikan pada pernyataan masalah, atau bagaimana sebuah future dapat direalisasikan/didekati)*}

## Data Hasil Penelitian

## Pembahasan Hasil Penelitian

## Analisis Data

## Perbandingan dengan Penelitian Lain

# KESIMPULAN DAN SARAN

## Kesimpulan

***Petunjuk****:*

*Berisi kesimpulan (menyatakan laporan TA sudah diselesaikan) dan pandangan alternatif (bilamana terdapat peluang saran untuk pengembangan bagi pelaksanaan penelitian lanjutan). Kesimpulan harus dapat ditarik dari awal mulai identifikasi masalah. tujuan penelitian, pembahasan, dan hasil pembahasan.*

## Saran

# 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

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

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

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

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

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

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.

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

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

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

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

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

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

# Lampiran 1 Daftar Riwayat Hidup

# Lampiran 2 Asdf

# Lampiran 3 Xyzxyz

*{...Formulir Bagian Punggung Halaman Sampul...}*

|  |  |  |  |
| --- | --- | --- | --- |
| **TUGAS AKHIR** | **NAMA :**  **NIM :** | **...........JUDUL TUGAS AKHIR...........** | **....TAHUN....** |