GARUDA Dataset Specifications: Complete Data Structure Report

Dataset Overview

The GARUDA synthetic dataset generates **5 interconnected datasets** that provide comprehensive data for training graph neural networks to detect gambling money laundering networks. All datasets are generated in CSV format with proper headers and data types optimized for machine learning workflows.

Dataset 1: Transaction Records (`garuda_transactions.csv`)

Purpose: Core transaction log containing all financial transfers in the synthetic banking network
Expected Size: ~50,000 rows
Primary Use: Training sequence models, temporal analysis, transaction classification

Variable Specifications

Variable	Data Type	Description	Example Values
`transaction_id`	string	Unique identifier for each transaction	"TXN_000001", "TXN_000002"
`timestamp`	datetime	Transaction date and time (ISO format)	"2024-03-15 14:23:45"
`from_account`	string	Source account number	"0141234567890", "GAMBLER_12345"
`to_account`	string	Destination account number	"0089876543210", "MERCHANT_5678"
`amount`	float	Transaction amount in Indonesian Rupiah	250000.0, 5000000.0
`transaction_type`	string	Categorized transaction type	"GAMBLING_DEPOSIT", "NORMAL_PAYMENT"
`description`	string	Transaction description/memo	"TRANSFER - DEPOSIT GAME", "TRANSFER - UTILITIES"
`from_bank`	string	Source bank code	"BCA", "MANDIRI", "BRI"
`to_bank`	string	Destination bank code	"BNI", "CIMB", "DANAMON"
`is_cross_bank`	boolean	Whether transaction crosses bank boundaries	true, false
`hour`	integer	Hour of day (0-23)	14, 19, 23
`day_of_week`	integer	Day of week (0=Monday, 6=Sunday)	0, 3, 6
`is_suspicious`	boolean	Ground truth label for suspicious activity	true, false
`risk_score`	float	Calculated risk score (0-1)	0.05, 0.89, 0.95

Transaction Type Categories

- **NORMAL**: Regular legitimate transactions
- **GAMBLING_DEPOSIT**: Money sent to gambling operations
- **GAMBLING_WITHDRAWAL**: Winnings withdrawn from gambling
- **MONEY_LAUNDERING**: Proceeds being laundered through complex schemes
- **SUSPICIOUS_DEPOSIT**: Large unusual incoming transfers
- **SUSPICIOUS_WITHDRAWAL**: Rapid withdrawal of large amounts
- **FAKE_BUSINESS_INCOME**: Shell company fake revenue
- **LAUNDERING_RECEIPT**: Receiving gambling proceeds for laundering

Dataset 2: Account Profiles (`garuda_accounts.csv`)

Purpose: Account-level information with behavioral features and network role labels

Expected Size: ~10,000 rows

Primary Use: Node classification, account profiling, risk assessment

Variable Specifications

Variable	Data Type	Description	Example Values
`account_id`	integer	Unique account identifier	1, 2, 3
`account_number`	string	Bank account number	"0141234567890"
`bank`	string	Bank name	"BCA", "MANDIRI", "BRI"
`region`	string	Geographic region	"JAKARTA", "SURABAYA", "BANDUNG"
`role`	string	Network role (ground truth label)	"GAMBLING_OPERATOR", "NORMAL"
`creation_date`	date	Account opening date	"2022-05-15"
`initial_balance`	float	Starting balance in IDR	5000000.0
`is_suspicious`	boolean	Binary suspicious flag	true, false
`risk_level`	integer	Risk level (0-5, 5=highest)	0, 2, 5
`total_transactions`	integer	Total number of transactions	45, 156, 23
`total_volume_in`	float	Total incoming transaction volume	50000000.0
`total_volume_out`	float	Total outgoing transaction volume	48000000.0
`avg_transaction_amount`	float	Average transaction size	125000.0
`max_transaction_amount`	float	Largest single transaction	10000000.0
`unique_counterparties`	integer	Number of different accounts transacted with	23
`cross_bank_ratio`	float	Ratio of cross-bank transactions	0.35
`velocity_score`	float	Transaction frequency indicator	2.5
`temporal_pattern_score`	float	Unusual timing pattern indicator	0.7

Role Categories (Ground Truth Labels)

- **NORMAL** (85%): Regular legitimate account holders
- **GAMBLING_OPERATOR** (0.1%): Primary targets - gambling website operators
- **SHELL_COMPANY** (0.5%): Fake companies used for money laundering
- **COLLECTION_ACCOUNT** (1%): "Rekening pengepul" - fund aggregation accounts
- **MULE_ACCOUNT** (8%): Recruited accounts (students, low-income individuals)
- **MONEY_CONVERTER** (0.2%): Cryptocurrency/foreign exchange facilitators

- **RECRUITER** (0.3%): Individuals who recruit mule accounts

Dataset 3: Network Graph Structure (`garuda_network.gexf`)

Purpose: Graph structure showing relationships between accounts

Format: GEXF (Graph Exchange XML Format) - standard for network analysis

Primary Use: Graph neural network training, network analysis

Node Attributes

Attribute	Data Type	Description
`id`	string	Account identifier
`role`	string	Network role label
`bank`	string	Bank affiliation
`region`	string	Geographic location
`risk_level`	integer	Risk assessment (0-5)

Edge Attributes

Attribute	Data Type	Description	Example Values
`relationship`	string	Type of connection	"gambling_proceeds", "money_laundering"
`weight`	float	Connection strength (0-1)	0.7, 0.9, 0.3
`transaction_count`	integer	Number of transactions between accounts	5, 23, 156
`total_amount`	float	Total money transferred	25000000.0
`first_transaction`	date	Date of first connection	"2024-01-15"
`last_transaction`	date	Date of most recent connection	"2024-03-20"

Relationship Types

- **gambling_proceeds**: Money flow from gambling to operators
- **money_laundering**: Laundering transactions between accounts
- **recruitment**: Mule account recruitment connections
- **shell_company_flow**: Transactions through fake companies
- **collection_aggregation**: Fund collection patterns

Dataset 4: Temporal Features (`garuda_temporal_features.csv`)

Purpose: Time-series behavioral features for each account

Expected Size: ~10,000 rows (one per account)

Primary Use: Temporal pattern analysis, behavioral modeling

Variable Specifications

Variable	Data Type	Description	Example Values
`account_id`	integer	Account identifier	1, 2, 3
`account_number`	string	Bank account number	"0141234567890"
`peak_hour_activity_ratio`	float	Activity during gambling peak hours (19-23)	0.65
`weekend_activity_ratio`	float	Weekend vs weekday activity	0.45
`velocity_change_30d`	float	Transaction frequency change (last 30 days)	3.2
`amount_variance`	float	Variance in transaction amounts	15000000.0
`dormancy_periods`	integer	Number of inactive periods > 7 days	2
`burst_activity_events`	integer	High-activity periods (>10 transactions/day)	5
`evening_gambling_pattern`	boolean	Matches typical gambling time patterns	true
`monthly_cycle_score`	float	Correlation with salary cycle patterns	0.8
`cultural_event_correlation`	float	Activity during Indonesian cultural events	0.3
`seasonal_pattern_strength`	float	Strength of seasonal patterns	0.6

Dataset 5: Network Analytics (`garuda_network_metrics.csv`)

Purpose: Graph-theoretic metrics for each account

Expected Size: ~10,000 rows

Primary Use: Network-based feature engineering, centrality analysis

Variable Specifications

Variable	Data Type	Description	Example Values
account_id	integer	Account identifier	1, 2, 3
degree_centrality	float	Number of connections (normalized)	0.05, 0.89
betweenness_centrality	float	Bridging position in network	0.002, 0.156
closeness_centrality	float	Average distance to all other nodes	0.23, 0.78
eigenvector_centrality	float	Influence based on connections' importance	0.01, 0.45
pagerank_score	float	PageRank centrality score	0.0001, 0.0234
clustering_coefficient	float	How connected are neighbors	0.6, 0.9
community_id	integer	Detected community membership	1, 5, 12
community_size	integer	Size of community	15, 234, 567
triangle_count	integer	Number of triangles involving this node	0, 5, 23
k_core_number	integer	K-core decomposition level	2, 5, 8
local_clustering	float	Local network density	0.4, 0.8

Data Relationships and Integration

Primary Keys and Foreign Keys

```
garuda_accounts.account_id → garuda_transactions.from_account (via account_number)
garuda_accounts.account_id → garuda_transactions.to_account (via account_number)
garuda_accounts.account_id → garuda_temporal_features.account_id
garuda_accounts.account_id → garuda_network_metrics.account_id
```

Network Graph Integration

The GEXF network file contains the same account IDs as the CSV files, enabling seamless integration between tabular features and graph structure for hybrid machine learning models.

Data Quality and Validation

Missing Values

- **Designed with no missing values** - all fields populated during generation
- **Null handling**: External accounts (GAMBLER_, *MERCHANT_*) may appear in transactions but not in account profiles

Data Consistency

- **Account numbers**: Follow Indonesian bank account format (bank code + 10 digits)
- **Timestamps**: All in Indonesian timezone (UTC+7)
- **Amounts**: All in Indonesian Rupiah, positive values only

- **Risk scores**: Normalized between 0-1 for consistency

Validation Flags

Each dataset includes validation fields to ensure data integrity:

- **is_suspicious**: Ground truth labels for supervised learning
- **risk_level**: Ordinal risk classification (0=normal, 5=highest risk)
- **role**: Categorical network role for node classification tasks

Usage Recommendations

For Graph Neural Networks

1. Use `garuda_network.gexf` for graph structure
2. Use `garuda_accounts.csv` for node features
3. Use `garuda_transactions.csv` for edge attributes and temporal sequences

For Traditional ML Models

1. Use `garuda_accounts.csv` as primary feature table
2. Join with `garuda_temporal_features.csv` and `garuda_network_metrics.csv` for enriched features
3. Use `garuda_transactions.csv` for sequence modeling

For Federated Learning

Each bank's subset can be extracted by filtering on the `bank` field, maintaining realistic data distribution while enabling multi-institutional training scenarios.

File Formats and Compatibility

- **CSV files**: UTF-8 encoding, comma-separated, headers included
- **GEXF file**: XML format compatible with Gephi, NetworkX, PyTorch Geometric
- **Date formats**: ISO 8601 standard (YYYY-MM-DD HH:MM:SS)
- **Boolean values**: true/false (Python compatible)
- **Numeric precision**: Float64 for monetary amounts, Float32 for ratios/scores

This comprehensive dataset structure provides all necessary components for training sophisticated gambling money laundering detection systems while maintaining realistic data relationships and comprehensive ground truth labeling.