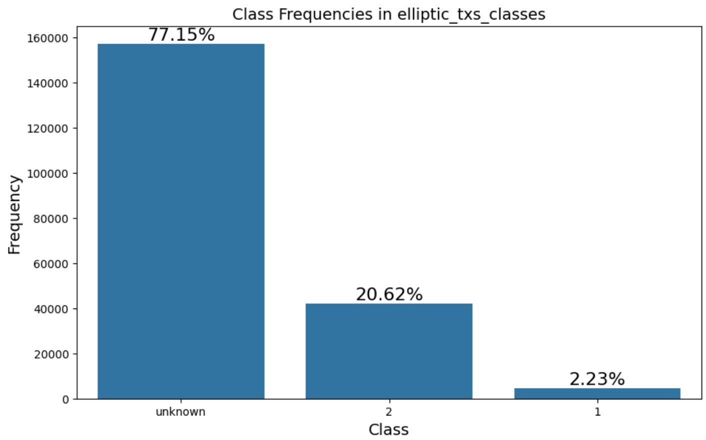
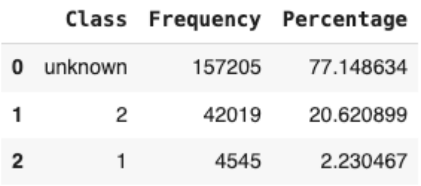
# Business Case:
# Evaluating AML Risk Through Data Science

Jhoan Flores Luna

## 1.- Exploratory Data Analysis and Preprocessing:

The data set is made of 204K nodes and 234K edges, where nodes are categorized as follow:

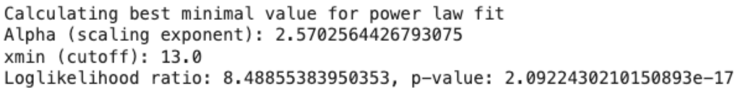




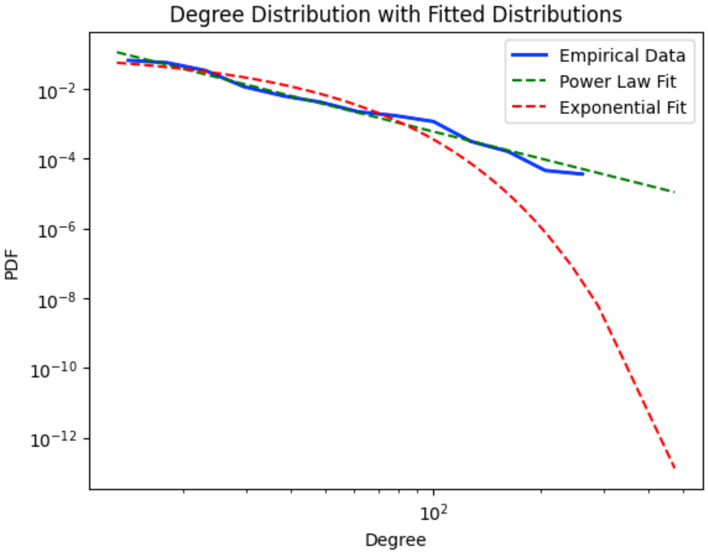
| | Class | Frequency | Percentage |
|---|---|---|---|
| 0 | unknown | 157205 | 77.148634 |
| 1 | 2 | 42019 | 20.620899 |
| 2 | 1 | 4545 | 2.230467 |

First, we notice it is an unbalanced target, only 2% of observations are labeled as illicit transactions.

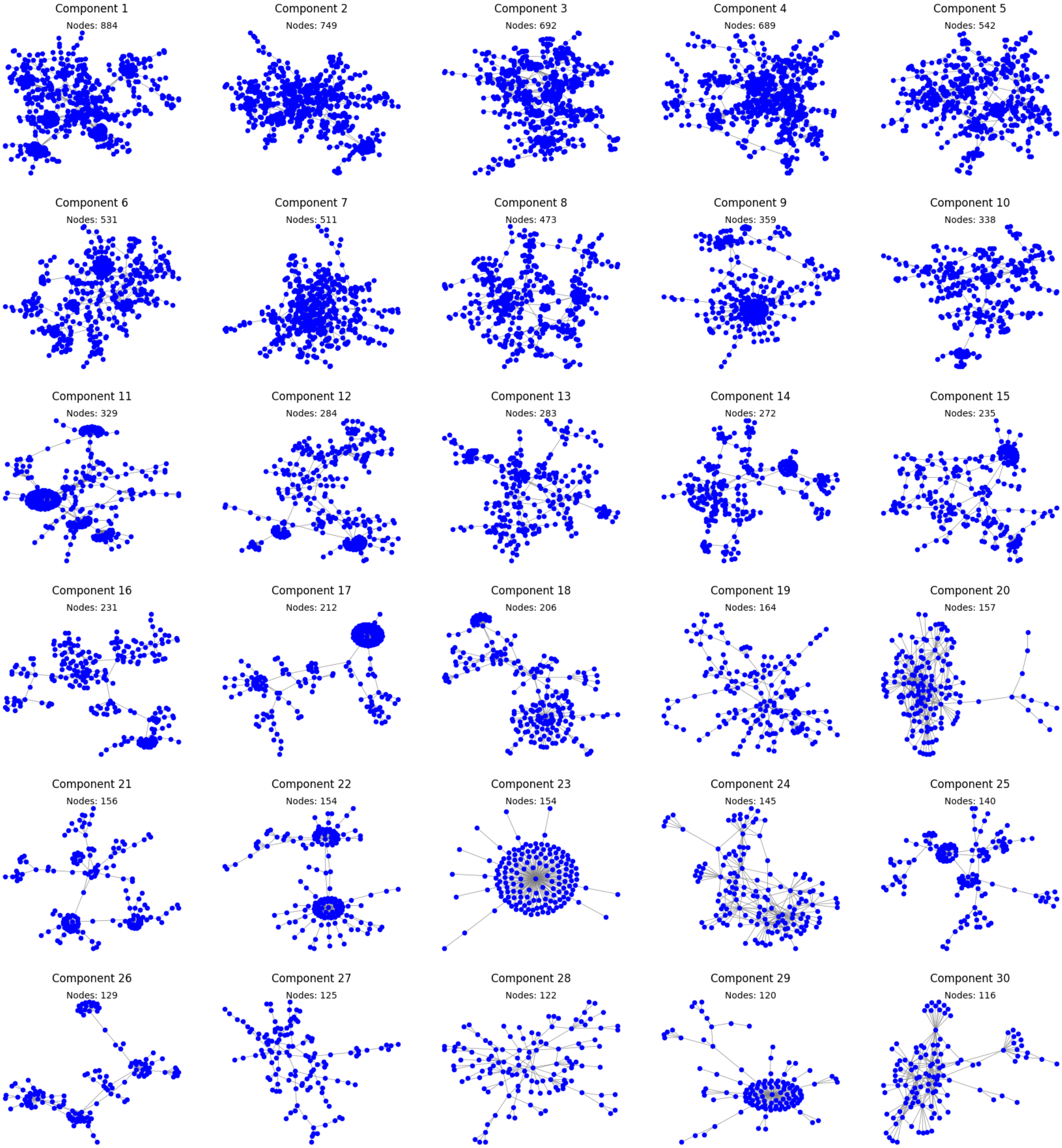
Regarding the network, we firstly highlight than the network is a free-scale network (power law distribution) for values >=13, that means there are few nodes which have the highest number of links.



```
Calculating best minimal value for power law fit
Alpha (scaling exponent): 2.5702564426793075
xmin (cutoff): 13.0
Loglikelihood ratio: 8.48855383950353, p-value: 2.0922430210150893e-17
```



Analyzing the components structure in the graph, we conclude there are multiple components similar to the “giant component” (with 884 nodes, the second one have 749 nodes and the third one with 692). Also, around 70% of the components has less than 3 nodes; this means we are facing a very **sparse network** (lot of small “islands”).



Component 1
Nodes: 884

Component 2
Nodes: 749

Component 3
Nodes: 692

Component 4
Nodes: 689

Component 5
Nodes: 542

Component 6
Nodes: 531

Component 7
Nodes: 511

Component 8
Nodes: 473

Component 9
Nodes: 359

Component 10
Nodes: 338

Component 11
Nodes: 329

Component 12
Nodes: 284

Component 13
Nodes: 283

Component 14
Nodes: 272

Component 15
Nodes: 235

Component 16
Nodes: 231

Component 17
Nodes: 212

Component 18
Nodes: 206

Component 19
Nodes: 164

Component 20
Nodes: 157

Component 21
Nodes: 156

Component 22
Nodes: 154

Component 23
Nodes: 154

Component 24
Nodes: 145

Component 25
Nodes: 140

Component 26
Nodes: 129

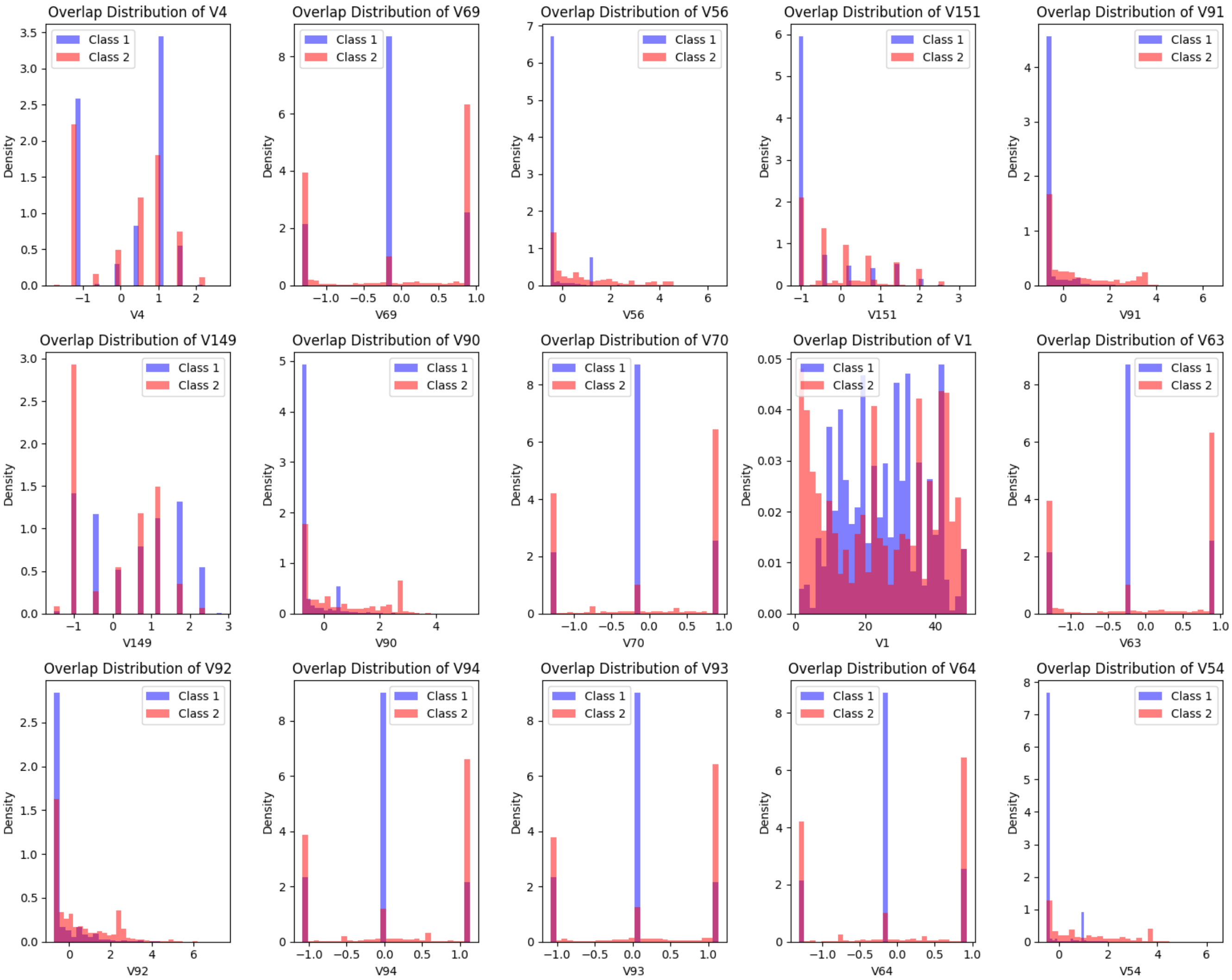
Component 27
Nodes: 125

Component 28
Nodes: 122

Component 29
Nodes: 120

Component 30
Nodes: 116

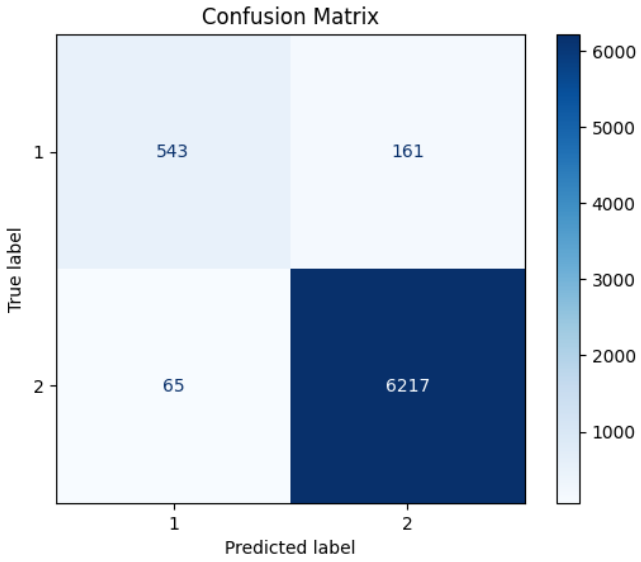
Based on a profiling analysis, we can notice there are some features with different behavior in the data set that can be used to build our classification model. Unfortunately, detailed description of the variables is not provided to give a better business-oriented analysis.

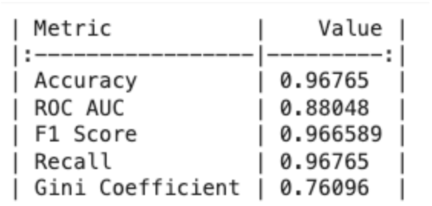


## 2.- GNN Model

Considering a classification problem (class 1 and 2). We removed the “unknown” labels to build a binary classification model.

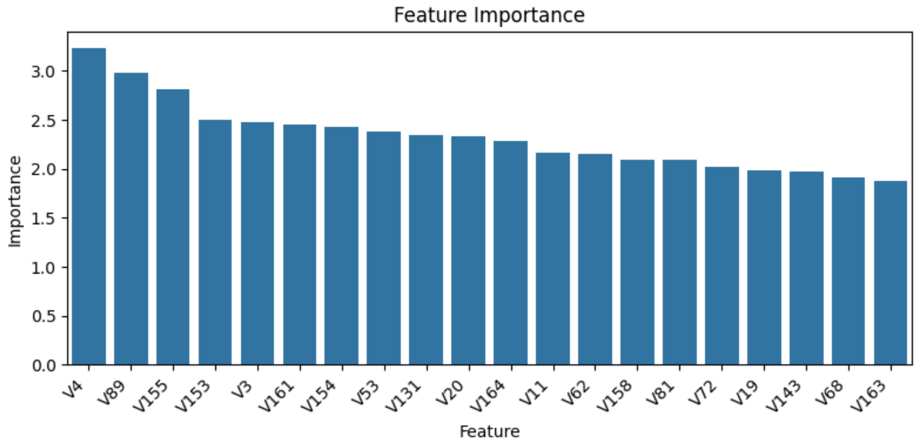
For our model we use a traditional GNN model, getting the following results:





```
| Metric           |    Value |
|:-----------------|---------:|
| Accuracy         | 0.96765  |
| ROC AUC          | 0.88048  |
| F1 Score         | 0.966589 |
| Recall           | 0.96765  |
| Gini Coefficient | 0.76096  |
```

With this ranking for feauture importance:



## 3.- Conclusion and forward improvements:

1. **High Sparsity Increases Modeling Challenges:**
   The network is extremely sparse, meaning most nodes have few connections. This makes it difficult to detect meaningful patterns, especially when modeling influence, communities, or flow dynamics.

2. **Lack of Variable Descriptions Limits Interpretability**
   Without clear descriptions or metadata for each variable, it is hard to assign business meaning to nodes or edges. This reduces the practical value of the insights extracted from the model.

3. **Scalability Issues Due to Network Size**

The size of the network poses computational challenges. Processing and analyzing the graph—especially for centrality metrics, community detection, or large matrix operations—requires significant memory and processing power. Standard machines struggle without efficient graph partitioning or parallel processing.