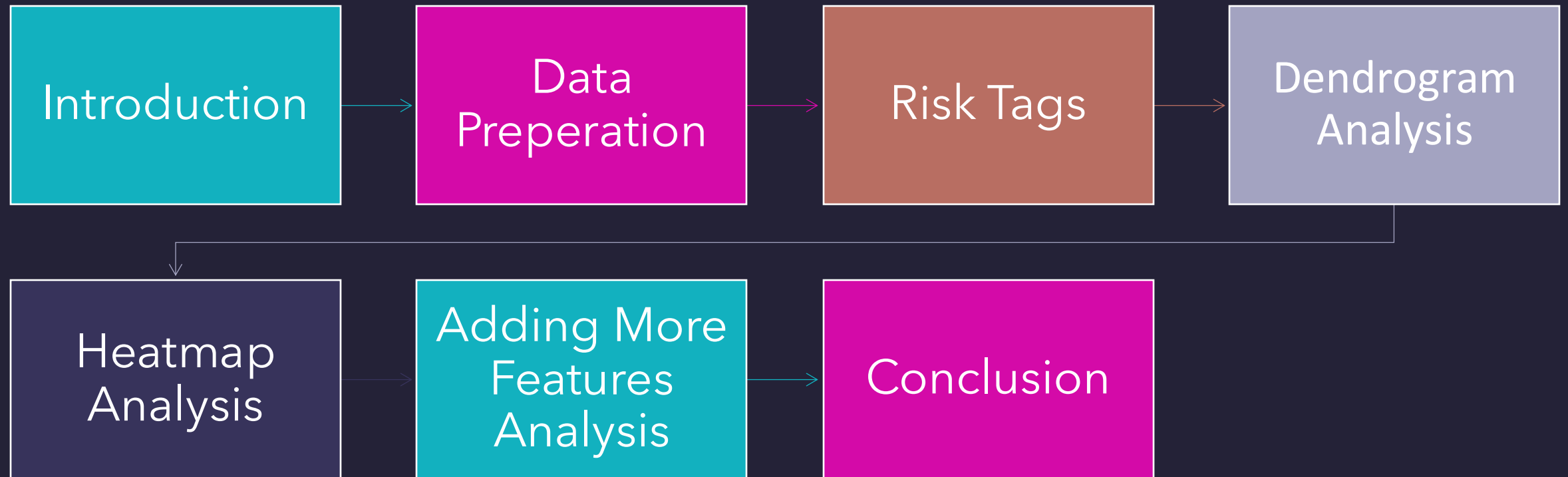


Webacy Externship Cluster Report Analysis Slides

By Ben (Minh) Pham

Introduction



Data Preparation | Feature Selection:

Identifying key features that may affect smart contract security, such as exploitation, buy_tax, and sell_tax. These two were picked because they stand as the major risk factors in blockchain security.

I then added the features ``is_closed_source`` and ``is_honeypot`` later in the analysis, which would allow identification of the contracts that likely hide vulnerabilities or inversely deliberately try to fool.

Data Preparation | Data Transformation:


Normalization: Since buy_tax and sell_tax are continuous features, they have to be normalized in order to bring them to the same scale. Otherwise, the other features with a large range may dominate the clustering process, which would be undesirable.

Missing Data Handling: The missing values in the dataset have been either substituted with the median value for that feature or, if not significant, removed.

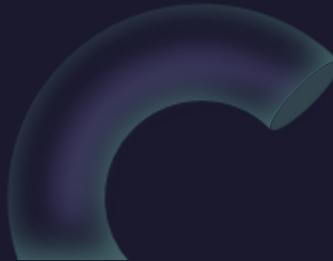

Data Preparation | Data Cleaning:

Outlier Removal: This is far from the data points that might bias the outcome of clustering; outliers are often considered contracts with extreme values. These might not represent the general population.

Data Validation: Making sure that there is a dataset that is consistent and correct on whether the features actually describe the nature of smart contracts.



Risk Tags We're Analyzing

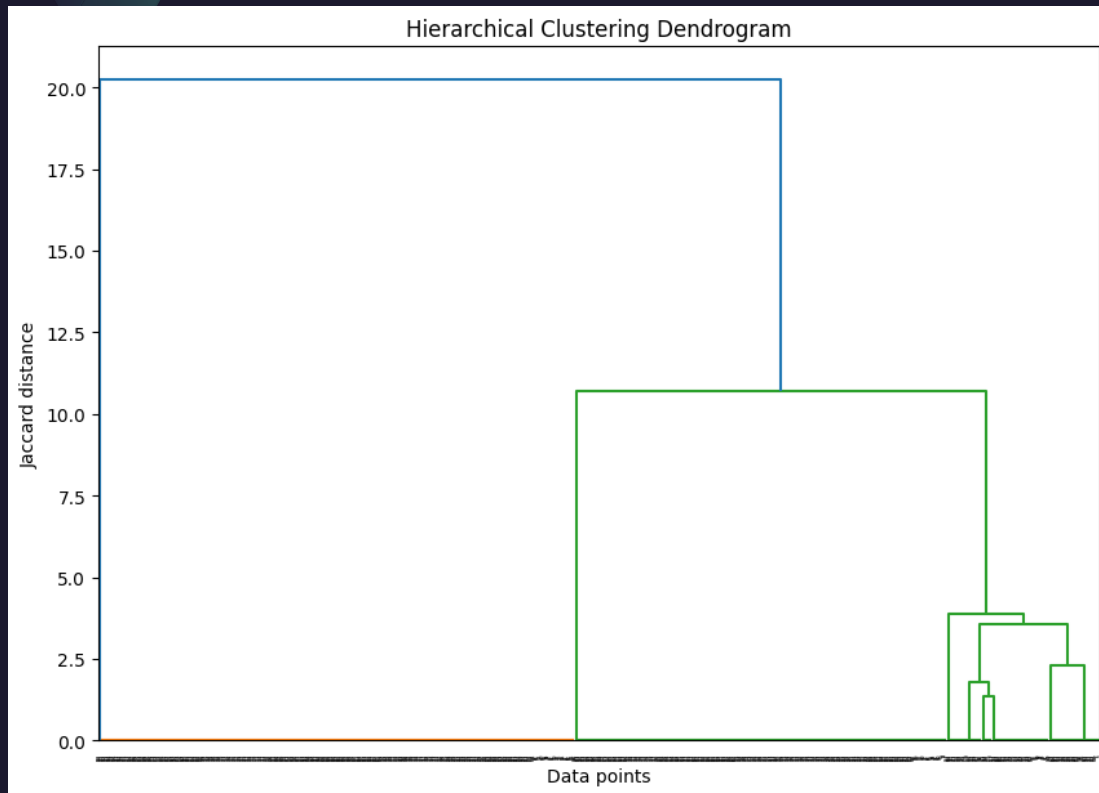


Exploitation: Utilizing smart contract weaknesses to steal money or interfere with operations.

Buy Tax: An expense associated with buying tokens inside of a smart contract; occasionally, this cost is set high without being fully disclosed, which can impact investment returns.

Sell Tax: This charge is applied when selling tokens, just like the buy tax. When token holders want to sell, high or concealed sell taxes may catch them off guard.

Dendrogram Analysis | Blue Line

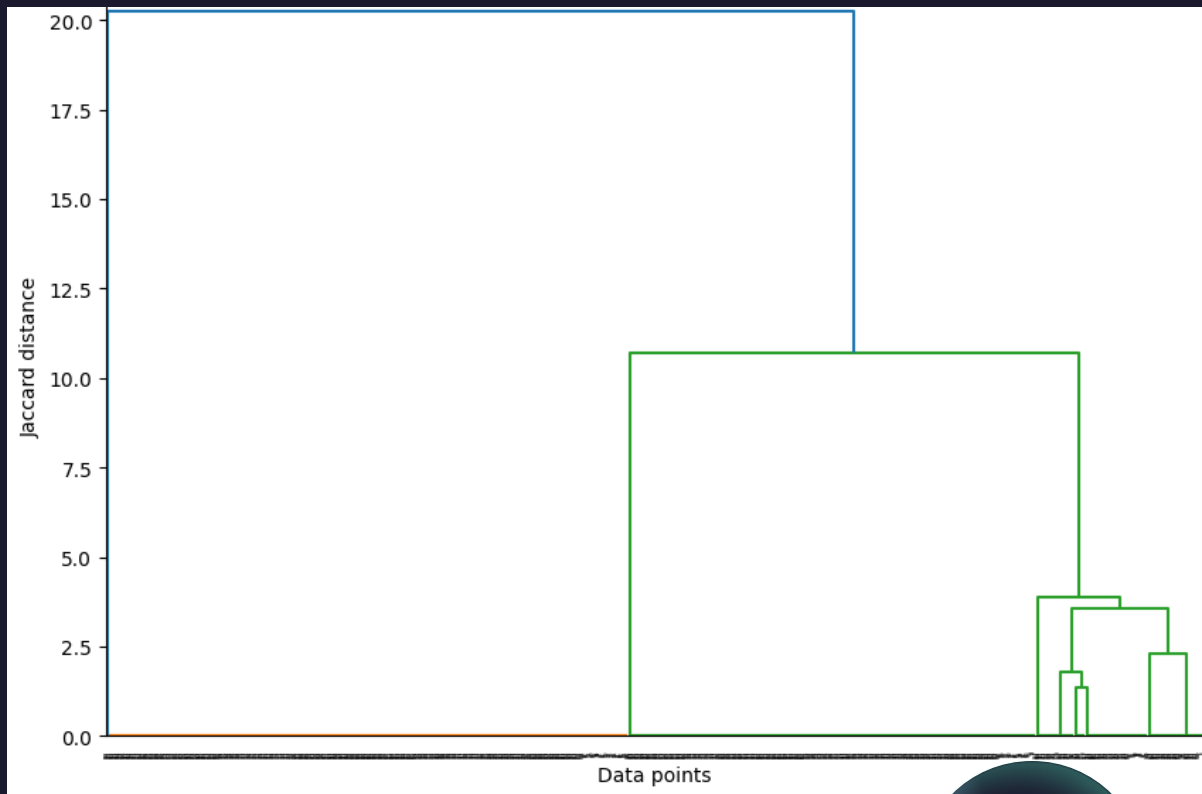


Height: The height of the blue vertical line is approximately a Jaccard distance of 20. One can imply that one cluster and others differ significantly from one another.

Implications: Shows a considerable difference between risk profiles, which can be utilized to isolate an outlier or new risk from the dataset.

There is a huge jump, from the green line to the blue.

Dendrogram Analysis | Green Cluster Grouping



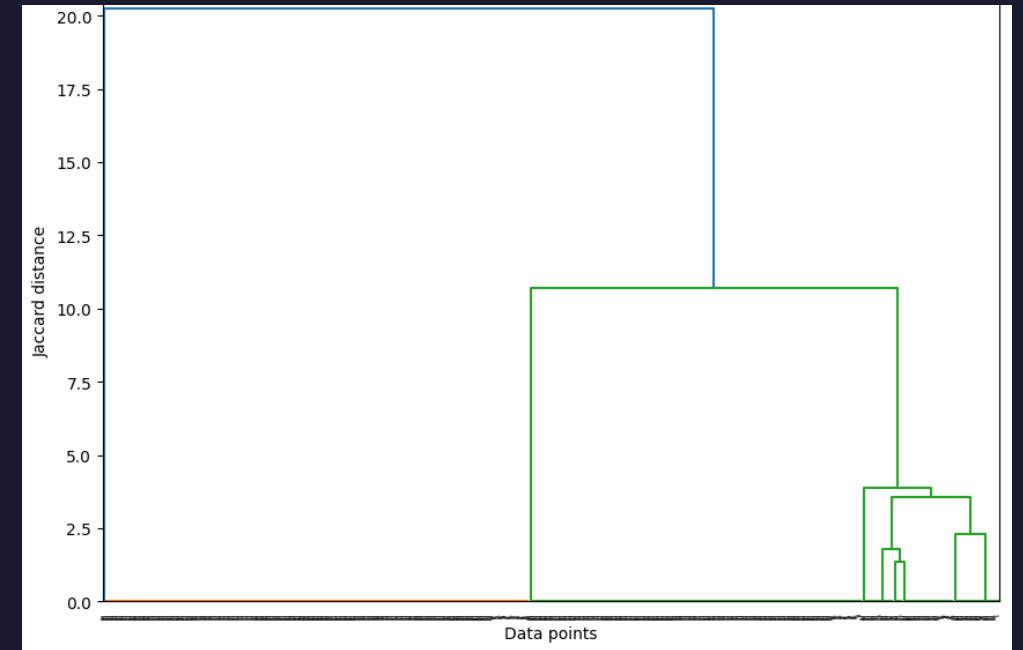
Sub-clusters: Several other clusters at a lower proximity are shown at the bottom as a number of green lines. These are signifiers of a set of smart contracts that are somewhat similar in terms of the risk tags that they both come with or the risks they both pose.

Close Proximity: The short distances indicate that these clusters are quite similar when it comes to risk.

Dendrogram Analysis | Optimal Number of Clusters

Cut-Off Point: The perfect cut-off is going to be around a Jaccard distance of about 10 before a major rise to the blue line at 20.

Number of Clusters: Cutting the dendrogram at this height of about 10 would most likely result in 2 to 3 clearly separated groups—one big and one or two small groups or outliers.



Dendrogram Analysis | Examining Outliers

Unique Cluster There is an outlier to the data at the top of this graph, around height 20. This would indicate a unique smart contract, unlike the others. An investigation here could expose an anomaly or a specific cluster of contracts with certain unique risk features.

Possible reasons: it may be due to its abnormally high or low tax, a unique characteristic of the exploitation, or a multiplex set of these two that give the marks of distinction to this contract.

Dendrogram Analysis | Why was Jaccard used?



Measured dissimilarity will be estimated by the Jaccard distance because it is binary data, with risk tags represented as present (1) or absent (0).



This metric measures dissimilarity, which estimates feature presence used in the analysis of binary data. It does that by determining the fraction of shared positive features about all of the features in either item.



This is particularly useful with datasets where risk inclusion outweighs risk exclusion.

Dendrogram Analysis | Cluster Characteristics



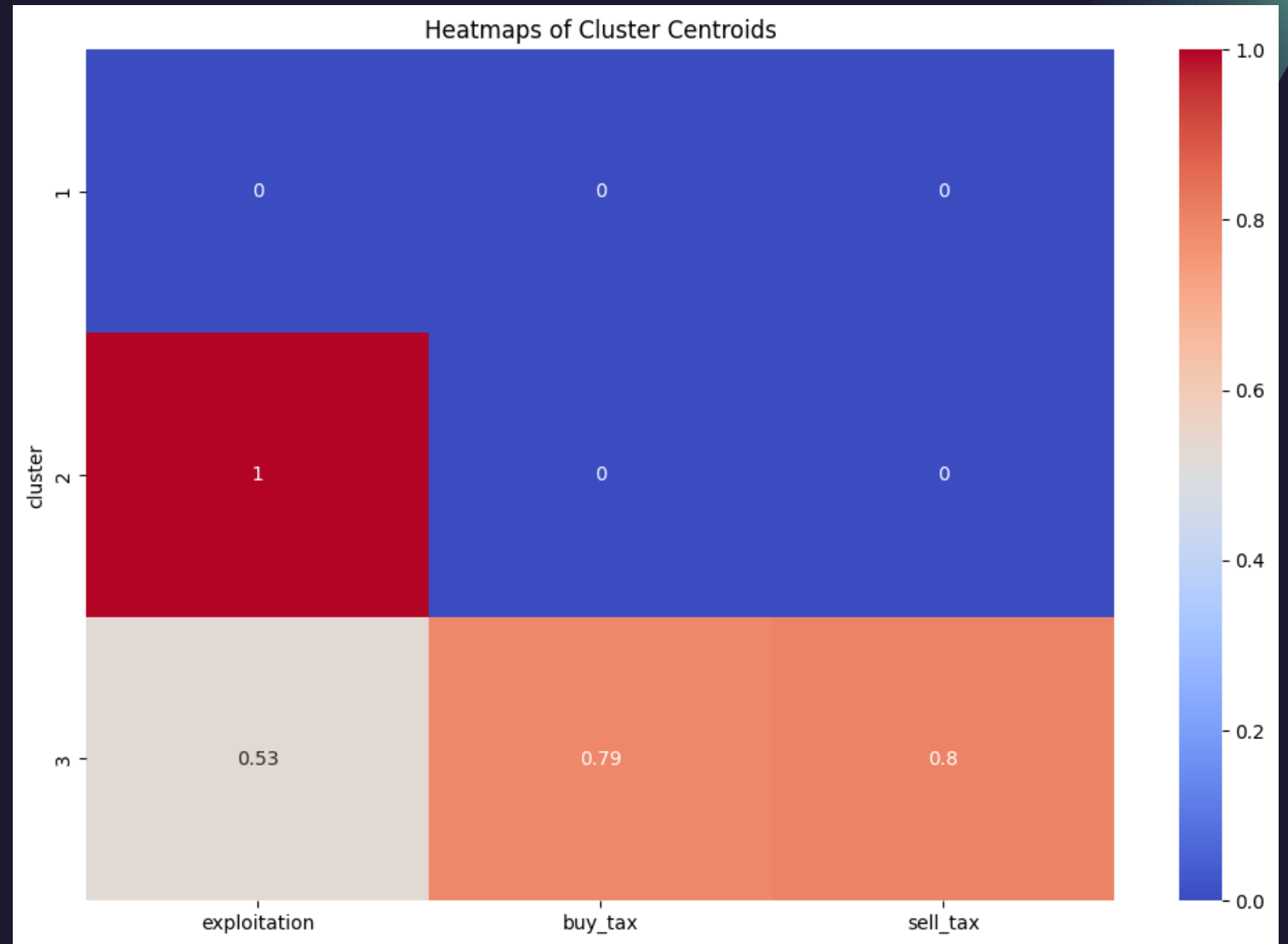
It would just be the mean of the mode values of Exploitation, Buy_Tax, Sell_Tax, etc. The green cluster works out the value that would be formed by the cluster of green lines in the larger cluster.



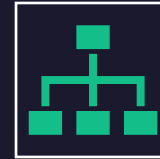
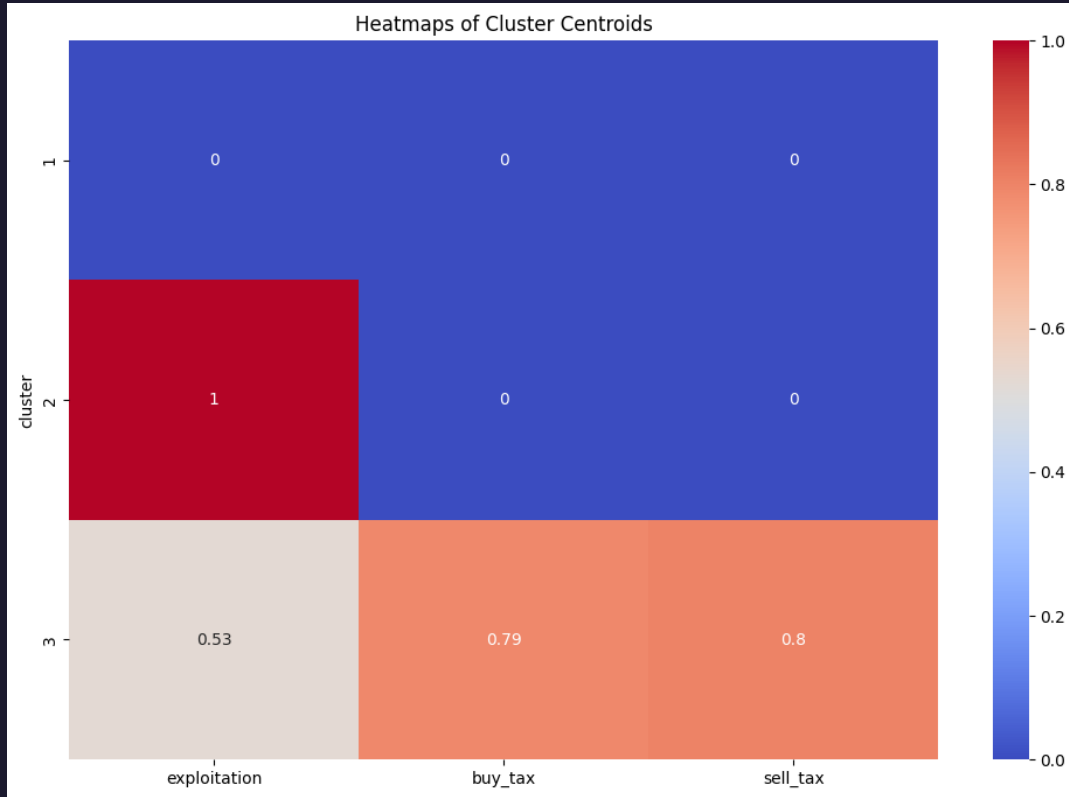
For example, if most of the contracts in that cluster have a similar buy/sell tax or similar vulnerabilities, it forms the profile of the cluster.

Cluster Heatmap | Observation

- Cluster 1: This cluster shows no indication of exploitation, buy_tax, or sell_tax risks.
- Cluster 2: This cluster exhibits a high risk of exploitation (value of 1) but does not show risk in buy_tax or sell_tax.
- Cluster 3: This cluster shows moderate risks across all three categories—exploitation (0.53), buy_tax (0.79), and sell_tax (0.8).



Cluster Heatmap Analysis | Risk Profiling



Cluster 1 is of the "low-risk" type. That is, the contracts in this cluster are expected to be very thin in terms of opportunities for adversaries.



Cluster 2 might represent a group of contracts that are highly vulnerable to exploitation. This could indicate that these contracts have a significant design flaw or are exposed to specific threats that others are not.



Cluster 3 seems to be a relatively balanced group, although with moderate risk. Contracts within such conditions would likely show more complex profiles of risk and thereby be open to numerous vulnerabilities.

Cluster Heatmap | Targeted Actions or Further Analysis

For Cluster 1, if these contracts include other, less prevalent dangers not included in the current data, or if they do, further research may be necessary to determine whether they really represent best practices.


For Cluster 2, it may represent a group of very exploitable contracts, which therefore could mean that these contracts either have some major design flaw or are a lot more vulnerable to certain specific threats than others are.

For Cluster 3, it probably contains a more balanced, relatively high-risk group of contracts with a slightly more complex risk profile or is more likely to have multiple vulnerabilities.

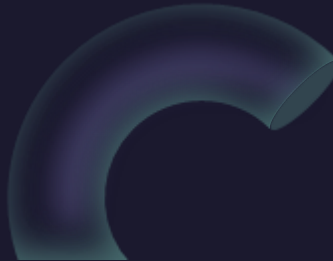

Cluster Heatmap | Policy or Monitoring Adjustments

Contracts in Cluster 2 should be prioritized for immediate security reviews and might warrant stricter monitoring. Those might merit closer scrutiny, perhaps with a focus on the development-time hardening of exploitation prevention strategies as the key.

The fact of the matter is that Cluster 3 contracts may require regular audits and focused improvements. That said, it might just be necessary for the improvement of the current risk detection tools.

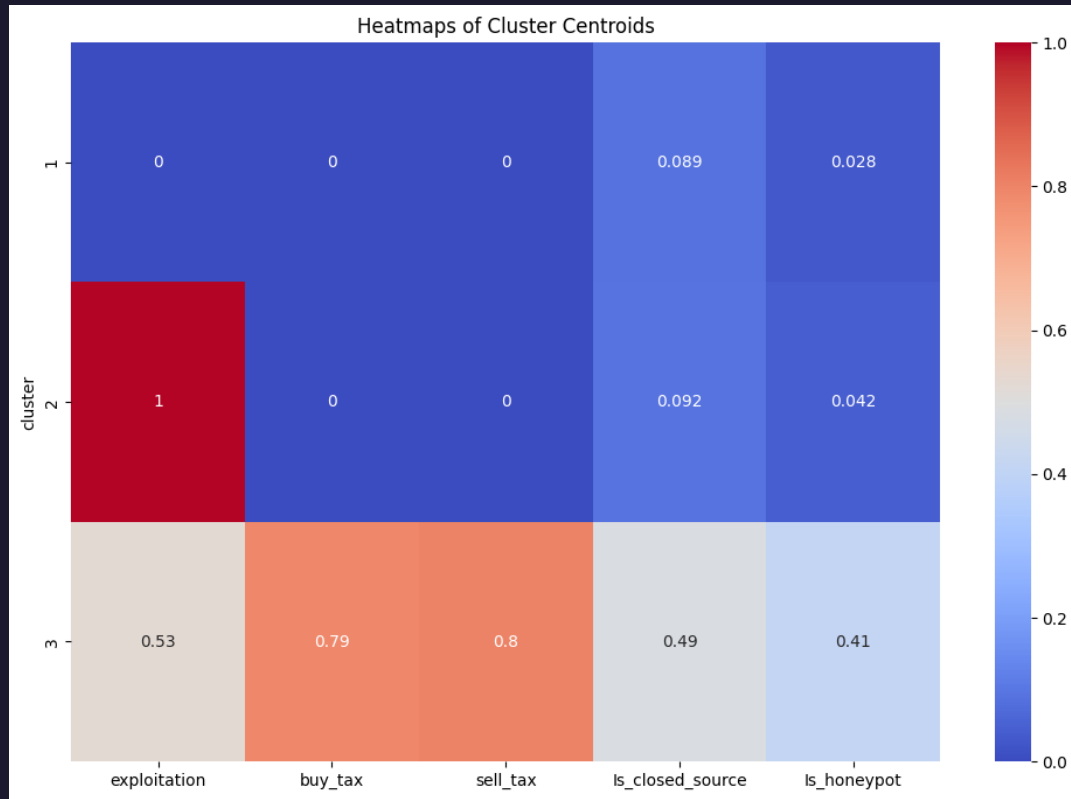


Adding More | Risk Tags We're Analyzing



`Is_closed_source`: "Closed source" refers to a smart contract whose code is not accessible to the general public. This opaqueness may conceal defects or malevolent operations.

`Is_Honey_Pot`: This is a dishonest smart contract that seems to entice investors but, in reality, keeps them from taking their money back after they have invested.



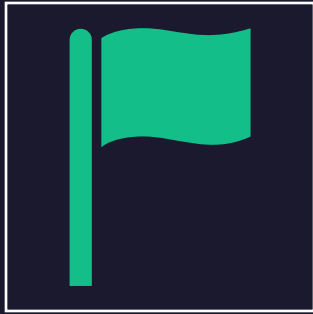
Adding More Features | Observation

Cluster 1: This cluster still shows no significant risks in exploitation, buy_tax, or sell_tax. However, there is a slight presence in is_closed_source (0.089) and a minimal risk in is_honeypot (0.028).

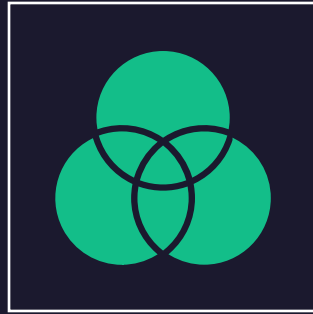
Cluster 2: This cluster continues to show a high risk of exploitation (1) but no risks in the buy_tax, sell_tax, is_closed_source, or is_honeypot features.

Cluster 3 shows moderate risks across all categories: Exploitation: 0.53, Buy_tax: 0.79, Sell_tax: 0.8, is_closed_source: 0.49, is_honeypot: 0.41

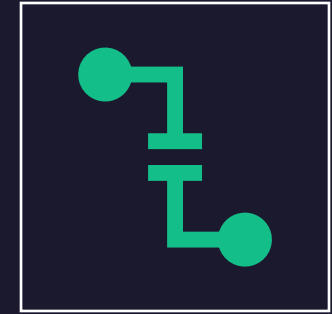
Adding More Features | Risk Profiling



Cluster 1 would comprise the low-risk contracts, even though there will be minor red flags in most of them, either closed source or potentially honeypots.



Cluster 2 obviously shows a very exposed contract. Clearly, these are only open contracts where there is nothing that could be taken advantage of, with no element of honeypotted or hidden source.



Cluster 3, therefore, seems to have the most sophisticated risk profile in the sense that it encompasses lots of features with medium levels of associated risk. Such contracts are likely to have a mixed bag of problems—susceptible to tax manipulation and closed-source honeypotting.



Research why some contracts in this cluster are closed-source or flagged—though minimally—as honeypots in Cluster 1. This could include verifying that the use cases of said contracts are legitimate and that the rationale for being closed source does indeed add value or introduce additional risk.



Understand the reasons as to why, in particular, these contracts are so easy to exploit in Cluster 2. One would need to know whether all of these contracts share some common flaw or if some specific attacks are targeted against them.



For instance, Cluster 3 consists of contracts that specify the interaction between the risks. Some might want a balance between being closed source, trying not to be taken advantage of by others, and staying open to taxation-related risks.

Adding More Features | Targeted Actions or Further Analysis

Adding More Features | Policy or Monitoring Adjustments

Even though it is likely low-risk, periodic reviews of the contract will always be beneficial to Cluster 1, emphasizing transparency and accessibility to source code.

Security auditing of contracts in Cluster 2 should be as close as possible to all mechanisms against their exploitation.

All contracts within Cluster 3 need frequent, detailed audits in order to understand the interplay of multifaceted risks and then be reduced through focused effort.

Conclusion | Clustering in Blockchain

Pattern and Anomaly Detection: Clustering exposed different risk profiles; hence, outlier identification presented the need for targeted security measures.

Risk Profiling: Contracts were categorized into risk profiles (high, medium, and low); audits and interventions were thus prioritized.

Deep dive analysis: New features, like closed source status and honeypot behavior, give a more complete view of security vulnerabilities.



Conclusion | Future Application of Blockchain

Enhanced Threat Detection: Real-time grouping to identify new attack vectors ahead of time.

Adaptive Security: Dynamically adjusts protocols for security against changing risk profiles over heterogeneous cluster groups.

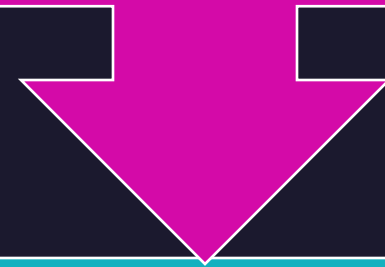
Automated Auditing: Automated auditing of smart contracts involves cleaning the low-risk clusters and flagging the high-risk clusters.



Conclusion | Leveraging Clustering for Proactive Security



Clustering, therefore, offers a proactive way of ensuring blockchain security; it can put insight into policymaking and further the process of auditing.



Further, artificial intelligence and machine learning can integrate for increased security in the blockchain in the years to come.



Thank you!
By Ben(Minh) Pham