CLUSTER ANALYSIS OF SMART CONTRACT RISKS

TASK - 4

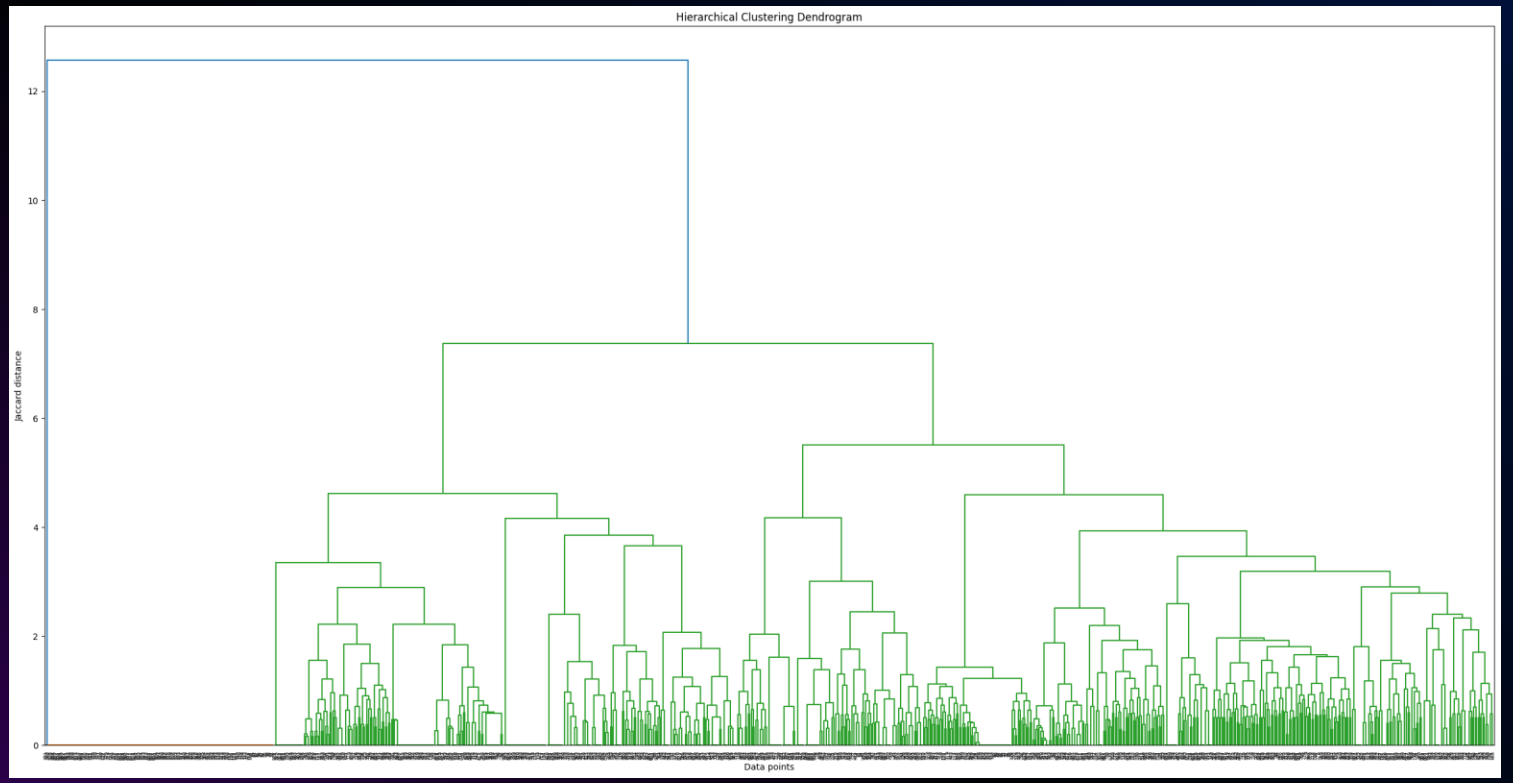
BY

VYSHNAVI

FEATURE SELECTION

- Analyzed 967 smart contracts with binary risk tags representing vulnerabilities.
- Redundant risk tags like 'sell_tax' were dropped using a phi correlation threshold of 0.7 to retain the most informative features.
- Hierarchical Clustering technique was used to find the clusters.
- Used Jaccard distance, well-suited for binary data, to measure contract dissimilarity based on risk profiles.

HIERARCHICAL CLUSTERING AND DENDROGRAM ANALYSIS



DENDROGRAM ANALYSIS

Large Blue Vertical Line: Extends up to a Jaccard distance of approximately 12, indicating a significant dissimilarity.

Green-Cluster Groupings:

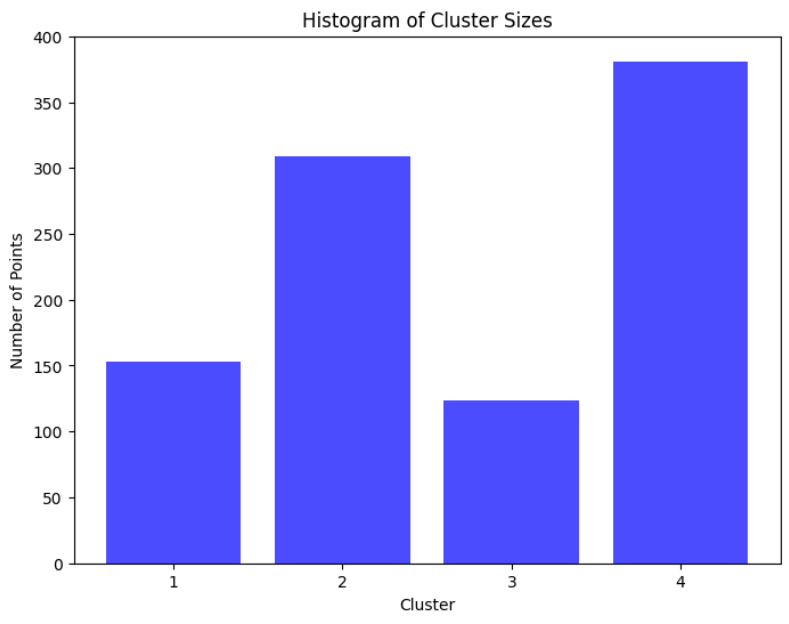
- Below the blue line, the green lines represent numerous smaller clusters that merge at shorter Jaccard distances, primarily in the range of 0 to 5.
- The shorter green lines indicate that these data points or sub-clusters are relatively similar, showing close relationships as they merge at low Jaccard distances.

Optimum number of Clusters:

Cut-Off Point: Based on the significant height of the blue line, an optimal cut-off might be around 5 Jaccard distance. This level balances meaningful distinctions without overly fragmenting the clusters.

Number of Clusters: Cutting at this point would likely result in four main clusters.

HISTOGRAM OF CLUSTER SIZES

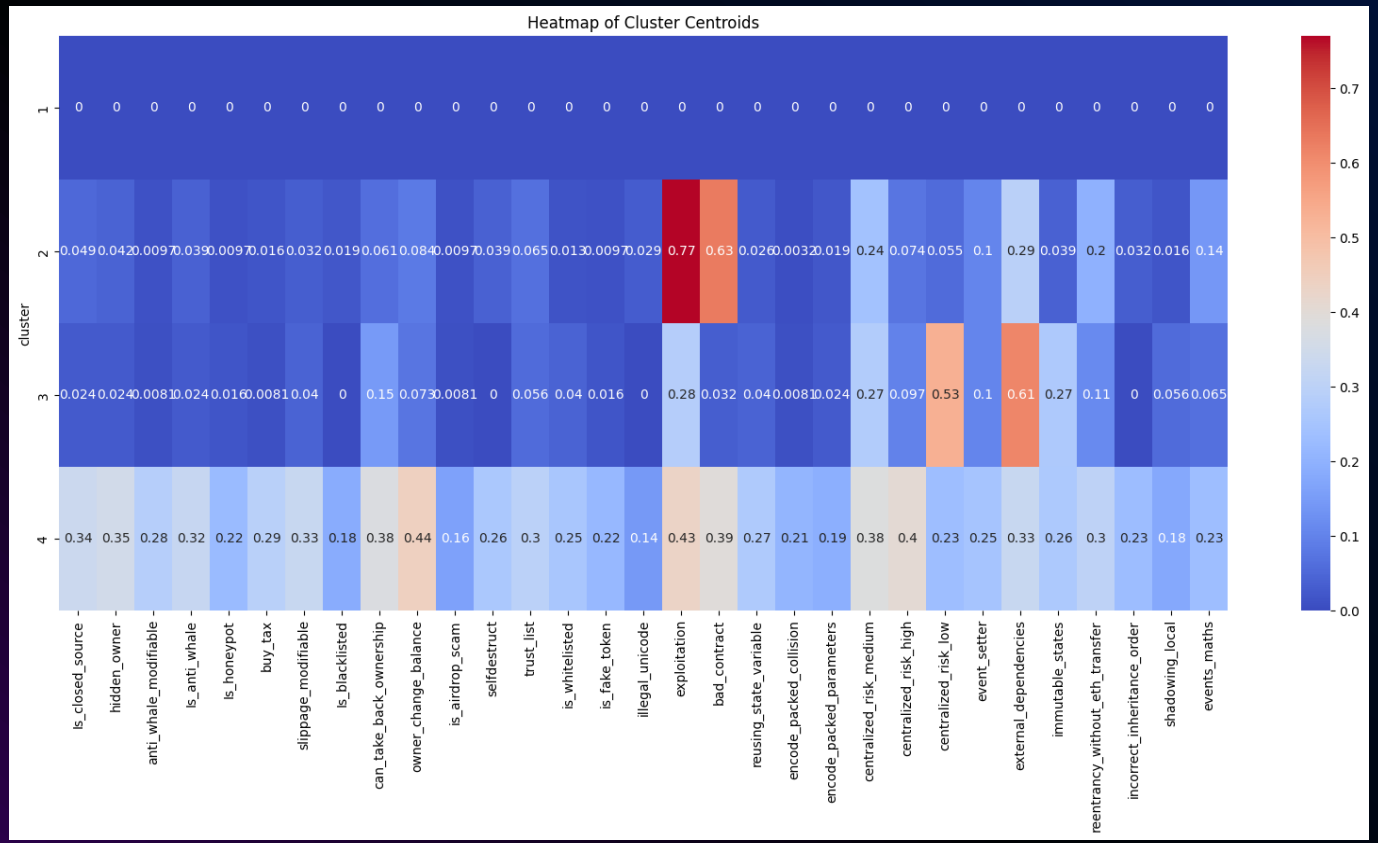


The histogram of cluster sizes indicates a skewed distribution:

- Cluster 4: The largest group, containing nearly 400 data points.
- Cluster 2: The second largest, with about 300 data points.
- Clusters 1 and 3: Smaller, containing around 150 data points each.

The size distribution suggests that while most data points share similar risk profiles, there are a few unique groups with distinct risk characteristics.

HEATMAP ANALYSIS OF CLUSTER CENTROIDS



HEATMAP ANALYSIS OF CLUSTER CENTROIDS

Cluster 1: All features have an average of 0.

- This group could be considered the "safest" or "lowest-risk" cluster among the smart contracts, as it lacks the specific vulnerabilities tracked in this analysis.

Cluster 2: Notable high-risk concentrations like "exploitation" (0.77) and "bad_contract" (0.63) being strong risks and "centralized_risk_medium" (0.24) and "external-dependencies" (0.29) being the moderate ones.

- This cluster represents contracts with deceptive characteristics and exploitation risks and this combination suggests potentially malicious contracts designed to deceive users.

HEATMAP ANALYSIS OF CLUSTER CENTROIDS

Cluster 3: High values in external_dependencies (0.61) and centralized_risk_low (0.63), low to moderate values in exploitation risks, immutable states, and other external vulnerabilities.

- This cluster represents legitimate but architecturally vulnerable contracts where risks stem from system design rather than malicious intent.

Cluster 4: Moderate but consistent risk values across multiple categories like 'owner_change_balance' (0.44), 'exploitation' (0.43), 'centralized_risk_high' (0.40) and consistent baseline of risk (0.2-0.35) across most other categories.

- This cluster contracts with diverse, distributed risks rather than acute singular vulnerabilities.

DRAWING INSIGHTS

Risk Profiling: Malicious intent (Cluster 2) and architectural vulnerabilities (Cluster 3) emerge as distinct patterns rather than mixing. Most concerning contracts show either high deceptive characteristics (Cluster 2) or high external dependencies (Cluster 3).

Targeted Actions or Further Analysis:

- For Cluster 2 (High Deceptive): Implement enhanced verification of contract code, especially checking for deceptive Unicode characters and exploitation vectors.
- For Cluster 3 (External Dependencies): Conduct thorough audits of external contract dependencies and implement fallback mechanisms.
- For Cluster 4 (Distributed Risks): Develop comprehensive monitoring systems that can track multiple moderate risks simultaneously

DRAWING INSIGHTS

Policy or Monitoring Adjustments:

1. Implement multi-tier monitoring system based on cluster characteristics:
   - High-frequency monitoring for contracts showing Cluster 2 patterns (potential scams).
   - Dependency health checks for Cluster 3 contracts.
   - Broad-spectrum monitoring for Cluster 4 contracts
2. Develop risk scoring system that weights both individual high risks and accumulated moderate risks.
3. Create early warning systems specifically tailored to each cluster's dominant risk patterns.

THANK YOU