Crimson Project - GeoTechnical Rock Lab

Present By: Xinyuan Chen, Linh Do, Hayden Lau

Agenda

1. Overview of the method
2. Modelling:
    a. Clusters Selection
3. Decision Tree and Rule Extraction (6 clusters)

1. Overview of the method:

In this model, we use agglomerative hierarchical clustering algorithm with Grower distance.

- Agglomerative hierarchical clustering is an unsupervised machine learning technique with a bottom-up approach
- Gower distance measures similarity between samples with mixed data types

1. Overview of the method:

How does agglomerative hierarchical clustering work?

Step 1: Each rock sample = its own cluster

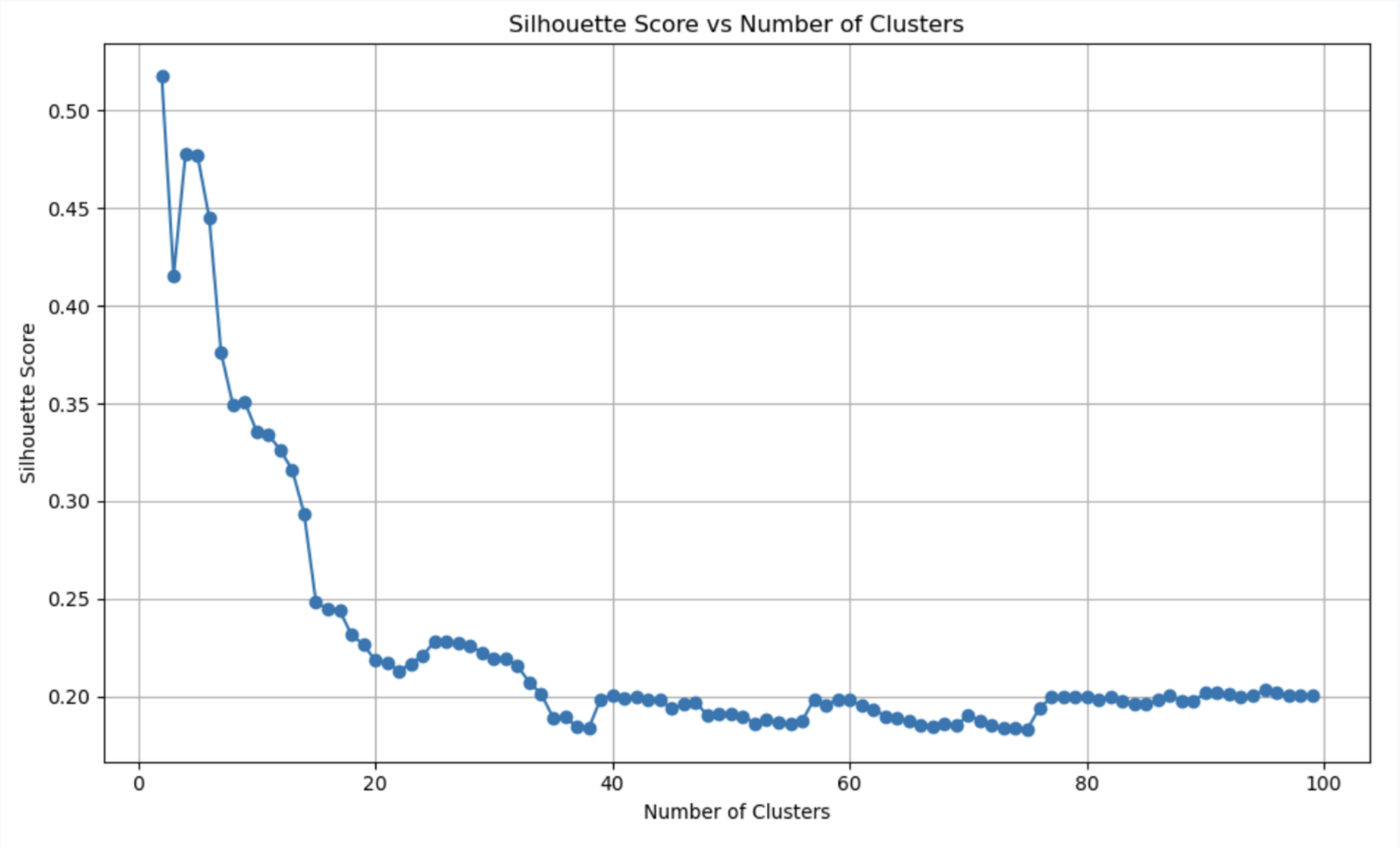
Step 2: Join the two most similar clusters based on Gower distance

Step 3: Keep merging until creating a big tree

Step 4: Slice the trees to create clusters

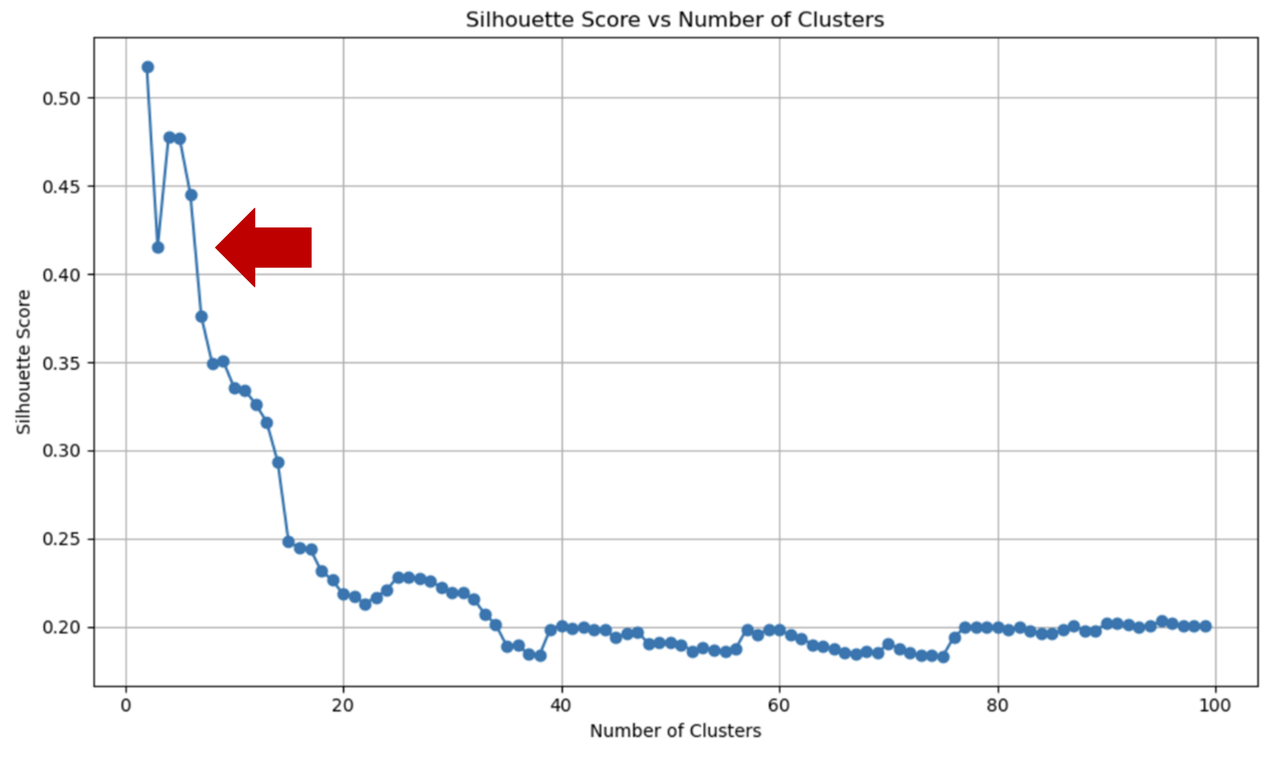
Cluster Evaluation Metrics

Silhouette Score: Measures how well samples fit within their clusters (higher is better)



Cluster Choose - 6 Clusters

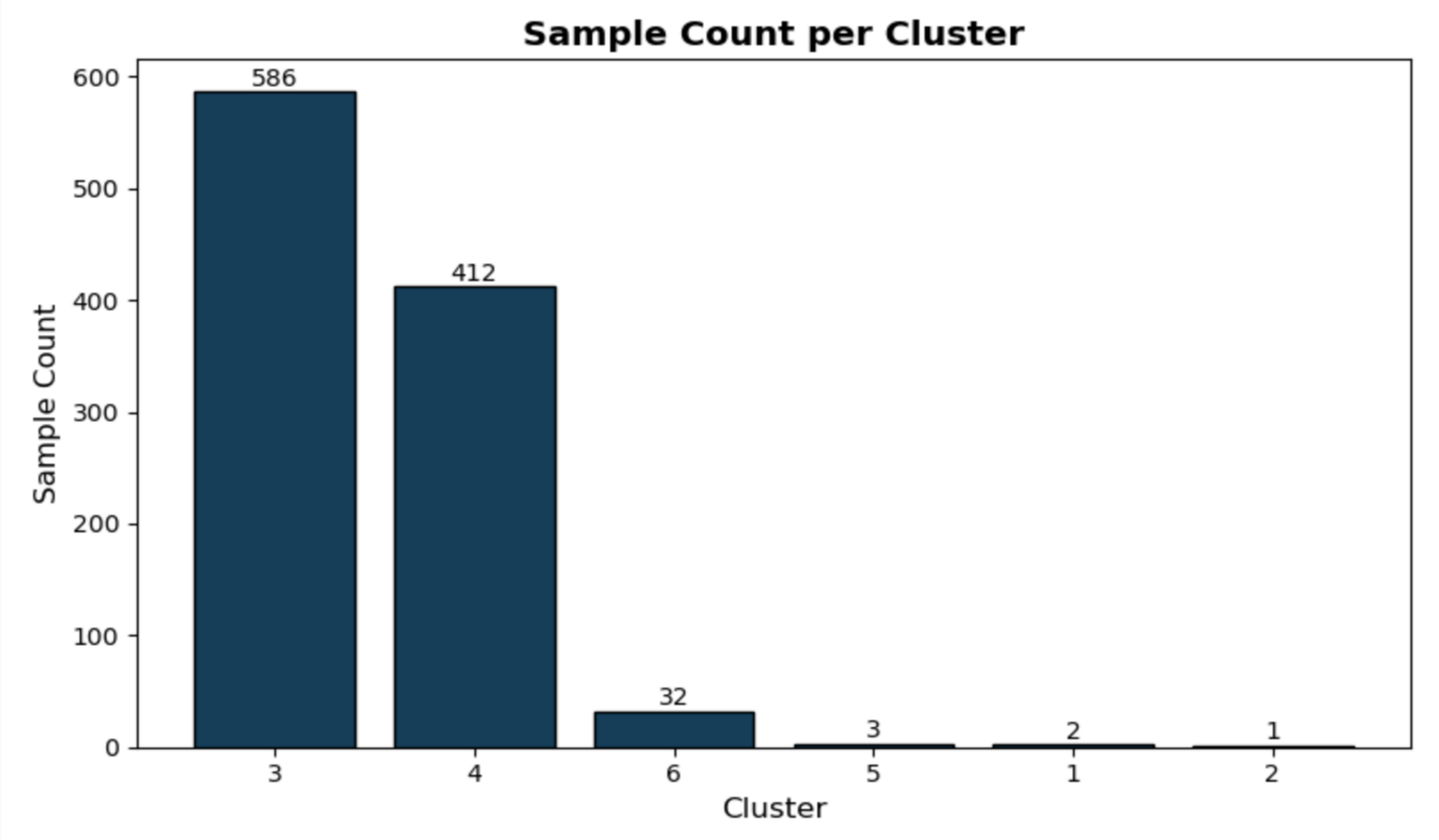
- Highest silhouette score across tested options
- Clear cluster separation for business interpretation
- Recommended as the final configuration



Trade-off – Cluster Count vs. Quality

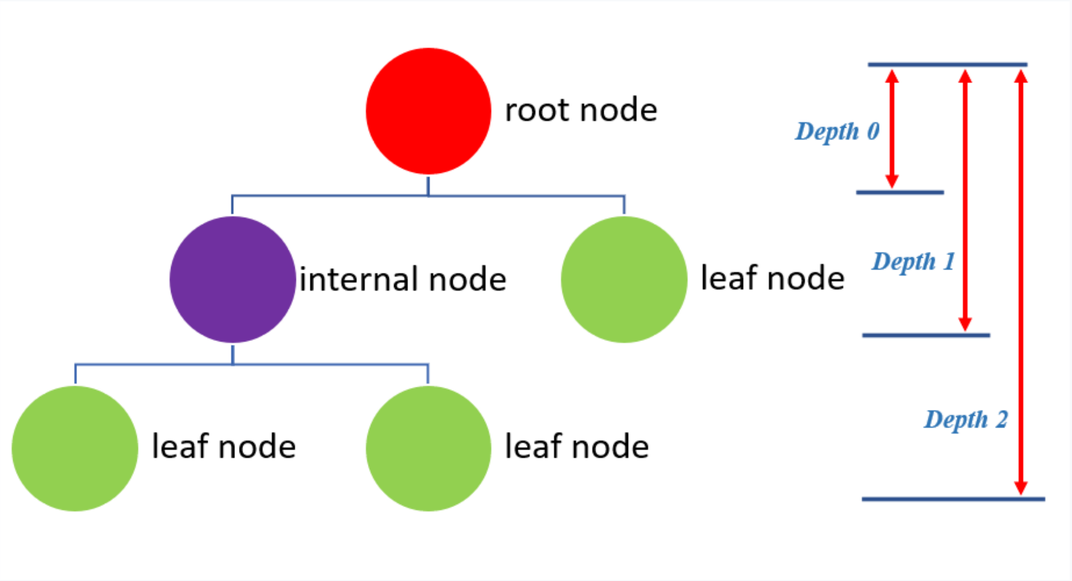
Cluster Count	Pros	Cons
3–6	Higher silhouette score (better fit)	Less detailed
8–12	More detailed groups	Lower silhouette (less reliable fit)
50–100	Extremely detailed	Very hard to interpret; noisy

Cluster Result



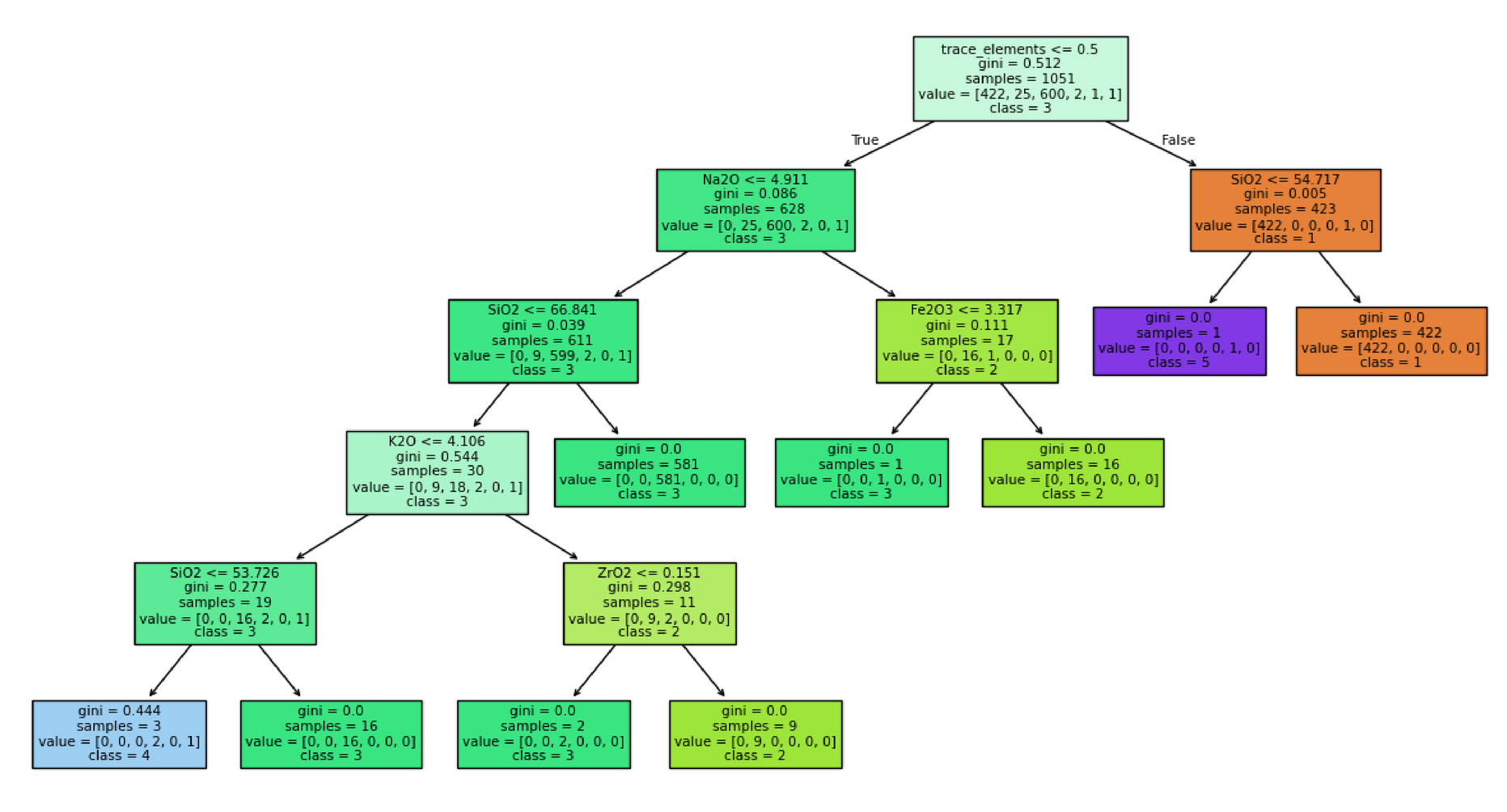
Decision Tree

- Finds the underlying rules of the clustering of the data
- Optimal tree depth will be decided by the algorithm
- Clusters with very few data points (i.e. < 5) may not appear in the decision tree



Decision Tree

- Optimal tree depth: 5



Rule Extraction

- Clusters with more data points will likely to have more rules
- OR relationship between each line of rules in each cluster

```
Cluster 3
- trace_elements = 0 AND Na2O <= 4.911 AND SiO2 > 66.841 (proba: 100.0%)
- trace_elements = 0 AND Na2O <= 4.911 AND SiO2 <= 66.841 AND SiO2 > 53.726 AND K2O <= 4.106 (proba: 100.0%)
- trace_elements = 0 AND Na2O <= 4.911 AND SiO2 <= 66.841 AND K2O > 4.106 AND ZrO2 <= 0.151 (proba: 100.0%)
- trace_elements = 0 AND Na2O > 4.911 AND Fe2O3 <= 3.317 (proba: 100.0%)

Cluster 1
- trace_elements = 1 AND SiO2 > 54.717 (proba: 100.0%)

Cluster 2
- trace_elements = 0 AND Na2O > 4.911 AND Fe2O3 > 3.317 (proba: 100.0%)
- trace_elements = 0 AND Na2O <= 4.911 AND SiO2 <= 66.841 AND K2O > 4.106 AND ZrO2 > 0.151 (proba: 100.0%)

Cluster 4
- trace_elements = 0 AND Na2O <= 4.911 AND SiO2 <= 66.841 AND SiO2 <= 53.726 AND K2O <= 4.106 (proba: 66.67%)

Cluster 5
- trace_elements = 1 AND SiO2 <= 54.717 (proba: 100.0%)
```

Classify new data point

- Find the cluster of a new unseen data point based on the rule extracted
- Format of data point entry: [SiO2, TiO2, ZrO2, Al2O3, Fe2O3, MnO, Na2O, K2O, trace_elements]
- Example: [60.123, 0.455, 0.120, 14.321, 6.789, 0.134, 5.322, 3.750, 0]
- Outcome:

```
Predicted cluster: 2
Probability distribution across clusters:
  Cluster 1: 0.0%
  Cluster 2: 100.0%
  Cluster 3: 0.0%
  Cluster 4: 0.0%
  Cluster 5: 0.0%
  Cluster 6: 0.0%
```

- Corresponding rule:

```
Cluster 2
- trace_elements = 0 AND Na2O > 4.911 AND Fe2O3 > 3.317 (proba: 100.0%)
- trace_elements = 0 AND Na2O <= 4.911 AND SiO2 <= 66.841 AND K2O > 4.106 AND ZrO2 > 0.151 (proba: 100.0%)
```