Risk Assessment and Clustering of Contaminated Lands in Israel

Team Members: Shiraz Hemo, Daniel Yesharim

Dataset Selection

- **Dataset Name:** "Contaminated Lands in Israel".
- **Description:** The dataset provides information on contaminated land sites in Israel, serving professionals in risk assessment, consulting, and other fields. It includes details on site activity, administrative status (investigation and remediation), update dates, and the presence of soil gases and groundwater.

Motivation

- **Problem and Significance:** Contaminated lands pose significant environmental and economic challenges. Identifying high-risk sites and analyzing remediation progress are essential for efficient resource allocation and policymaking. This project supports decision-makers by providing insights into remediation progress, clustering similar contamination sites, and identifying patterns for future risk management.

Method

1. **Data Preprocessing:** Handle missing data, convert categories to numeric formats, and normalize values.
2. **Supervised Learning:** Classify contamination risk levels using models like Logistic Regression and Random Forest.
3. **Unsupervised Learning:** Group similar sites using K-Means and DBSCAN clustering.
4. **Evaluation Metrics:** Use Accuracy and Silhouette Score for model validation.

Planned Experiments

- **EDA:** Analyze site distributions and remediation trends.
- **Classification:** Train and evaluate models to predict risk.
- **Clustering:** Detect contamination patterns and evaluate clusters.
- **Visualization:** Present findings using charts and maps.

Evaluation Plan

- **Model Validation**: Train models on 80% of the data and test on 20%.
- **Clustering Validation:** Assess results using clustering metrics like Silhouette Scores.
- **Final Output:** Produce clear and interpretable visualizations to summarize findings.