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 land is a major environmental and financial concern. Analyzing this data can help identify high-risk areas and track remediation progress, supporting better decision-making and resource allocation. This project is meaningful as it addresses real-world challenges and aligns with our coursework.

Method

1. **Data Preparation:** Clean data, handle missing values, and convert categories to numeric formats for machine learning.
2. **Supervised Learning:** Classify sites into "High Risk" or "Low Risk", using models like Logistic Regression and Random Forest.
3. **Unsupervised Learning:** Group sites with similar patterns using K-Means and DBSCAN clustering.
4. **Evaluation Metrics:** Use Accuracy and Silhouette Score for model validation.

Planned Experiments

- **EDA:** Explore the data to understand contamination types and remediation trends.
- **Modeling:** Train classification models and group sites through clustering.
- **Visualization:** Present results with graphs and maps to highlight patterns.

Evaluation Plan

- **Testing:** Train models on 80% of data and test on 20%.
- **Validation:** Use Silhouette Score and manual review for clustering.
- **Final Output:** The results will be presented as graphs and charts, showing contamination patterns and highlighting high-risk areas.