Project Proposal

Advanced Topics in Machine Learning

Title: **"Analyzing Radiation Intensity in Communication Sites"**
Team Member: Eyal Kadosh/208005280

Dataset Selection

The dataset, sourced from the Israeli Government Data Portal (data.gov.il), contains over 8,000 rows and 20 features describing communication sites across Israel. Key attributes include:

- **Location Data**: City, geographic coordinates (X_ITM, Y_ITM).
- **Technical Parameters**: Maximum theoretical radiation intensity ($\mu W/cm^2$), health threshold (%), and transmission technologies (2G, 3G, 4G, 5G).
- **Site Information**: Site type (e.g., roof antenna, ground tower).
  This dataset is well-suited for both supervised classification and unsupervised clustering due to its rich variety of numerical and categorical features.

Motivation

Radiation levels from communication sites are a growing public concern;
however, their efficient monitoring is challenging due to geographic and technological diversity. The aim of this project is to:
1. Forecast the site categories (e.g., antenna type) based on radiation and geographic factors for informed infrastructure management.
2. Cluster communication sites to identify patterns or anomalies, such as high-radiation locations, that can guide public health regulations and operational improvements.

The project addresses real-world needs that balance technological growth with public safety concerns.

Methodology

1.Supervised Analysis (Classification Task):

Objective: Predict site categories using features such as radiation intensity, health thresholds, and technology type.

Decision Trees: For interpretable classification of categorical features.

Random Forests: To capture feature interactions and enhance prediction robustness.

Logistic Regression: As a baseline linear model for site type prediction.

Support Vector Machines (SVM): For complex decision boundaries, especially in higher-dimensional feature spaces.

2.Unsupervised Analysis (Clustering and Anomaly Detection):

Objective: Identify site clusters based on radiation intensity, geographic proximity, and health threshold adherence. Detect anomalies such as high-radiation sites or outliers in densely populated areas.

K-Means Clustering: To group sites into clusters based on similar radiation and geographic attributes, providing insights into technological and locational trends.

DBSCAN (Density-Based Spatial Clustering): For identifying anomalies such as isolated high-radiation sites or clusters of concern near sensitive areas.

PCA (Principal Component Analysis): To reduce dimensionality and visualize site groupings, aiding in the interpretation of high-dimensional clusters.

Planned Experiments

Supervised Pipeline:

- Experiment with various classifiers to determine the optimal model.
- Use grid search for hyperparameter tuning and compare results across metrics such as F1-score and ROC-AUC.

Clustering and Visualization:

- Apply K-Means and DBSCAN to create clusters, followed by PCA for 2D visualizations of the groupings.
- Visualize clusters geographically using scatterplots of radiation intensity overlaid on site coordinates.
- Investigate high-radiation anomalies by correlating them with site type, location, and health thresholds.

Interpretation and Analysis:

- Examine how well clustering results align with known site categories and geographic regions.
- Investigate potential biases or trends, such as urban vs. rural site differences or correlations between radiation intensity and transmission technologies.

Conclusion

This project will provide a comprehensive analysis of communication site radiation levels, uncover patterns through clustering, and develop predictive models for site categorization. The findings can serve as a foundation for future regulatory frameworks, public safety initiatives, and technological implementations in other regions.