Project Title: Road Accident Analysis by Municipality

Team Members: Ariel Koren, Yosef Bencormono

Dataset Selection: The dataset selected for this project contains detailed information about road accidents categorized by municipal authorities. It includes data on the number of accidents reported in each municipality, sourced from publicly available government portals. This dataset allows for the identification of patterns, trends, and geographical disparities, providing a reliable foundation for analyzing road safety challenges and exploring factors influencing accident rates.

Motivation: Road accidents pose significant societal challenges, leading to fatalities, injuries, and economic costs. Municipal authorities play a vital role in addressing these challenges, yet there is often limited data-driven insight to guide interventions. By analyzing accident data across municipalities, this project aims to identify high-risk areas, uncover trends, and support effective policymaking. The novelty of this project lies in applying machine learning techniques in a comprehensive way to uncover actionable insights and improve upon existing methods for addressing road safety issues.

Method: To analyze the dataset, we will apply a complete end-to-end classification pipeline alongside unsupervised analysis techniques:

1. **End-to-End Classification Pipeline:**
   - **Data Preprocessing:** Handle missing values, normalize data, and encode categorical features.
   - **Feature Engineering:** Derive meaningful metrics such as accident rates per capita or road density.
   - **Model Selection and Training:**
     - Logistic Regression: Baseline model to classify municipalities by accident risk levels.
     - Random Forest: To identify key features affecting accident rates and capture non-linear relationships.
     - Gradient Boosting (e.g., XGBoost): To enhance classification accuracy and handle imbalances.
   - **Evaluation:** Use metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
   - **Interpretation:** Analyze feature importance and understand model predictions.
2. **Unsupervised Analysis:**
   - **Clustering:**
     - K-Means: To group municipalities with similar accident characteristics.
     - DBSCAN: For anomaly detection, highlighting municipalities with extreme accident rates.
   - **Dimensionality Reduction:** Apply PCA to reduce dimensionality and visualize patterns effectively.
   - **Validation:** Evaluate clustering results using silhouette scores and Davies-Bouldin Index.

Intended Experiments:

1. **Classification Models:** Train and evaluate at least five classification models, including logistic regression, random forest, gradient boosting, support vector machines, and neural networks, to classify municipalities by accident risk. Metrics specific to classification tasks, such as accuracy, precision, recall, F1-score, and ROC-AUC, will be used to assess performance. Additionally, we will ensure the evaluation does not include regression metrics, focusing solely on classification outcomes.
2. **Clustering Analysis:** Perform K-Means, DBSCAN, and an additional clustering method (e.g., Agglomerative Clustering) to uncover natural groupings and detect anomalies, enabling a robust comparison. Validate clusters using silhouette scores and Davies-Bouldin Index.
3. **Feature Engineering Impact:** Experiment with derived metrics, such as accident rates per capita and road density, to evaluate their impact on model performance.
4. **Cross-Validation and Error Analysis:** Use k-fold cross-validation to ensure robustness and conduct error analysis to identify patterns in misclassified cases.

By combining an end-to-end classification pipeline with unsupervised analysis, this project aims to provide actionable insights into road safety, enabling targeted interventions to reduce accident rates and improve public safety.