Extended Abstract

Title: *Machine Learning-Based Categorization of Central Roads in Athens Using Crash Risk Analysis*

Authors: Stelios Peithis, Paraskevi Koliou, George Yannis

Background and Objective:

Urban road traffic accidents (URTAs) represent a significant threat to public health and safety, particularly in densely populated cities like Athens. These incidents often occur at hazardous road segments where infrastructure, environmental conditions, and driver behavior intersect to elevate crash risk. While traditional road safety assessments rely heavily on historical crash data or manual observation, this study introduces a **machine learning-based approach** that integrates **telematics data**, **OpenStreetMap (OSM) data**, and **official crash records** to systematically **categorize urban road segments** in Athens as "Safe" or "Unsafe".

Methodology:

The methodology followed a structured three-phase approach:

1. **Problem Definition:** Focused on identifying high-risk urban road segments by analyzing areas with recorded crash incidents to generate actionable insights for traffic safety interventions.
2. **Data Integration and Preprocessing:** Combined telematics data (harsh braking, harsh acceleration, speeding), OpenStreetMap (OSM) data, and crash reports from Greek Authorities. A Python-based OSMnx framework matched textual crash location data with corresponding coordinates, creating a comprehensive graph of streets and municipalities in the Athens region. The study filtered telematics data from a total of **2614 trips** by **257 drivers** across **8 municipalities** to include only trips along crash-registered street segments.
3. **Modeling and Classification:** Using the processed data, an **XGBoost classification model** was developed. Road segments were labeled as “Safe” (≤2 crashes) or “Unsafe” (>2 crashes). To address class imbalance, the **Synthetic Minority Over-Sampling Technique (SMOTE)** was applied.

Feature Engineering and Analysis:

Driving behavior was quantified through the calculation of **ratios** for each key metric—speeding, harsh braking, and harsh acceleration—normalized against the number of trips per segment to mitigate the bias introduced by high-traffic roads. This normalization allowed the model to focus on **relative driving behavior intensity** rather than absolute counts.

Visual Insights:
Scatter plots illustrated the correlation between crash frequency and the normalized telematics behavior metrics, revealing distinct trends. For instance, higher speeding and harsh braking ratios correlated with segments labeled as unsafe, indicating their predictive potential for crash risk classification.

Model Performance and Evaluation:
The **XGBoost model** was evaluated using a confusion matrix and a classification report. It achieved:

- **Precision:** 0.88 (Safe), 0.77 (Unsafe)

- **Recall:** 0.83 (both classes)
- **F1-Score:** 0.86 (Safe), 0.80 (Unsafe)

The model showed strong generalization capabilities, effectively identifying high-risk segments with minimal false classifications. The analysis emphasized that the model was slightly more conservative in predicting unsafe segments—favoring higher precision for the "Safe" class.

Findings and Contributions:

- The study validated that **telematics data**—even without extensive physical infrastructure—can effectively contribute to urban road safety assessments.
- The **classification framework** provides an efficient, scalable, and low-cost alternative for identifying hazardous urban road segments.
- Policymakers and traffic authorities can use this model to **prioritize interventions**, **allocate safety resources**, and **deploy preventive measures** in high-risk areas.

Limitations and Future Work:

While the model's performance was robust, the study acknowledges the potential for improvement through the inclusion of **additional contextual variables** such as weather conditions, road surface quality, lighting, and pedestrian density. Further research could also **expand geographical coverage** to include other urban areas for external validation and model generalization.

Conclusion:

This research underscores the value of integrating **data science and urban planning** to address the growing challenge of urban traffic safety. By bridging crash records with telematics-derived behavioral data and applying machine learning, the study delivers a **practical tool for evidence-based decision-making**. The results not only contribute to safer road design and traffic management in Athens but also provide a replicable methodology for other urban centers seeking to adopt smart and proactive road safety strategies.