GEO880 Project Concept

Identifying Modes of Transportation from Personal GPS Data in Switzerland

# Research Question

Can we accurately identify the mode of transportation (walking, jogging, biking, or skitouring) based on GPS data features such as speed, elevation change, and stop frequency?

# Research Plan

**Objectives**

1. Develop a method to classify different modes of transportation using GPS-derived metrics.
2. Investigate the accuracy of the classification models in differentiating between very similar modes (e.g., biking vs. e-biking).
3. Evaluate the influence of environmental factors like terrain on mode detection accuracy.

**Data Collection and Processing**

1. **GPS Data**: Utilize your collected GPS data that includes latitude, longitude, elevation, timestamps, and potentially derived speed.
2. **Data Enrichment**: Calculate additional features necessary for the classification, such as:
   * **Speed**: Calculate instantaneous or average speed between GPS points.
   * **Elevation Gain/Loss**: Compute changes in elevation between consecutive points.
   * **Stop Frequency**: Identify stops based on minimal movement over a certain period.

**Analysis Approach**

1. **Feature Engineering**
   * Derive statistical features from raw GPS data that may indicate different transportation modes (e.g., max speed, average speed, elevation gain, variability in speed).
2. **Exploratory Data Analysis (EDA)**
   * Visualize feature distributions by known transportation modes (if some data is labeled) to identify patterns and potential overlaps.
   * Use plots such as histograms, box plots, and scatter plots to explore the relationships between features and modes.
3. **Model Development**
   * Split the data into training and testing sets.
   * Select appropriate machine learning models for classification tasks, such as logistic regression, decision trees, random forests, or support vector machines.
   * Train models using the training set and validate using cross-validation techniques.
4. **Model Evaluation**
   * Evaluate the performance of each model on the test set using metrics like accuracy, precision, recall, and F1-score.
   * Analyze the confusion matrix to understand misclassifications between modes.
5. **Refinement and Optimization**
   * Perform hyperparameter tuning to optimize model performance.
   * If necessary, revisit feature engineering to include or exclude features based on model performance and importance metrics.

**Visualization**

* Create confusion matrices for model results to visualize accuracy across different modes.
* Plot ROC curves for models to evaluate and compare performance visually.
* Use feature importance graphs to highlight which features most significantly impact mode classification.

**Evaluation and Discussion**

* Discuss the challenges of distinguishing between similar modes (e.g., biking vs. e-biking) and strategies to overcome these issues.
* Reflect on the limitations of GPS data, such as signal loss in dense areas or under heavy tree cover, and how it affects model accuracy.
* Consider potential real-world applications of this model, such as in fitness apps or urban planning for better infrastructure decisions.

**Conclusion**

This project not only enhances understanding of how to process and analyze movement data but also offers practical insights into personal transportation habits. Through rigorous data analysis using R and tidyverse, you’ll develop robust data science skills applicable to a range of real-world problems.