**Project 4 proposal (Group 5)**

**Members = Patrick Yau & Sharanjit Singh**

**Project topic:** Can we predict the severity of road accidents in the UK based upon 2021 data based on environmental and situational features (like weather, road conditions, and time)?

To build machine learning model and visualize them using various interactive data visualization techniques**.**

**Aims:**

* Predicting **accident severity** based on weather, road conditions, and vehicle type.
* **Identifying high-risk areas** where accidents frequently occur.
* **Analyzing patterns in accidents** related to time of day, road type, and urban vs. rural locations.

**Goals and ML Approaches:**

**1. Predict Accident Severity**

* **ML Type:** Classification
* **Models:** Scikit-learn (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost)
* **Dataset:** UK Road Accidents (2021)
* **Input Features:**
  + Weather\_Conditions
  + Road\_Surface\_Conditions
  + Light\_Conditions
  + Time, Speed\_limit
  + Urban\_or\_Rural\_Area
  + Vehicle\_Type
* **Target Variable:** Accident\_Severity (Multi-class: e.g., Slight, Serious, Fatal)

**2. Identify High-Risk Accident Zones**

* ML Type: Clustering
* Models: K-Means or DBSCAN
* Input Features: Latitude, Longitude, and optionally Time or Severity
* Goal: DBSCAN forms clusters based on density, detecting naturally occurring high-risk areas.

**3. Analyze Patterns and Relationships**

* **Tableau:** Use Tableau to show various graphical comparisons

**Technologies to Implement:**

| **Task** | **Technology** |
| --- | --- |
| **Data Preprocessing** | | | Python + Pandas |
| **Model Training** | | | Scikit-learn |
| **Visualization accident trends** | | | Matplotlib / Plotly / Seaborn / hvPlot / Panel |
| **Mapping High-Risk Areas** | | | Leaflet.js (via Folium) |
| **Database Integration** | | | MongoDB / SQLAlchemy with SQLite |
| **Hosting / Dashboard** | | | Tableau / AWS S3 / EC2 or Streamlit if permitted |

**Key Visualizations to Include:**

* Heatmaps of accident locations
* Clustered accident zones on Leaflet maps
* ROC curves or confusion matrices for classification models
* Feature importance from tree-based models

**Machine Learning Workflow**

1. **Data Preprocessing (Pandas)**
   * Handle missing values
   * Feature engineering (e.g., time of day from Time, urban/rural from Urban\_or\_Rural\_Area)
   * Encode categorical data
   * Scale numerical features
2. **Modeling (Scikit-learn)**
   * Train/test split
   * Build model (e.g., Random Forest or Logistic Regression)
   * Evaluate performance with classification report, confusion matrix
3. **Visualization**
   * Use **Matplotlib** to visualize feature importances and model metrics
   * Use **Tableau** for interactive maps, dashboards (e.g., severity by region, time trends)
4. **Database Storage**
   * Store cleaned data or predictions in MongoDB/SQL for access via JavaScript (Leaflet/Plotly)

**Purpose:**

Understanding what factors contribute to severe accidents can help improve safety measures, optimize road infrastructure, and support policymaking to reduce casualties.

**References:**

* <https://www.kaggle.com/datasets/xavierberge/road-accident-dataset>?