INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

DA 204o: Data Science in Practice
Course Project Proposal

Road Safety: Accident Severity Prediction

Anil Chandra Damodara, IISc, anilchandrad@iisc.ac.in

Gomathi Sankar S, IISc, gomathis@iisc.ac.in

Neeraj Kumar, IISc, neeraj3@iisc.ac.in

Ritesh Mishra, IISc, riteshmishra@iisc.ac.in

Image source: Internet

Course Project

- Compulsory!
- Marks: 20%
- Team size: 3-4
- Duration: ~6 weeks (Oct 15th to Nov 30th)
- Initial proposal: Oct 11th
  - Team formation (choose among yourself)
  - Select project topic/domain and/or datasets
- Final project proposal: Oct 18th
  - Detailed information: Problem definition, dataset(s), proposed methodology, and implementation plan.
  - Submission of slides (use the following slides)
- Checkpoints
  - First: Completion of data preparation and EDA (5%)
  - Second: Completion of model development and validation (5%)
  - Final: Final report, project presentation and demonstration (5%)
  - Peer feedback: 5%

Problem Definition

- Background of the problem
  - Despite major improvements in vehicle technology and road infrastructure, road accidents remain a key public safety issue . Factors like weather, road type, lighting, and driver behaviour interact in complex ways to determine accident severity.
- Why is it important?
  - Enables proactive road safety planning and emergency preparedness.
  - Helps identify high-risk locations and conditions.
  - Supports data-driven policymaking for transport authorities.
- Objectives of the project
  - Predict the severity of road accidents (Slight / Serious / Fatal).
  - Identify factors most correlated with severe outcomes.
  - Provide actionable insights for prevention and response.
- How can Data Science solve the problem?
  - Analyze and preprocess large-scale accident and vehicle datasets.
  - Build predictive models to classify accident severity.
  - Interpret key contributing features using explainable AI methods.

Data Collection and Preparation

- Data source(s) (where it's from, how it was collected)
  - Data Source**:** Kaggle – UK Road Safety: Traffic Accidents and Vehicles Dataset
  - Data Origin: UK Department for Transport (Open Data Portal).
  - URL: https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles/data/

- Description of the data (features, size, format)
  - **Size:** ~200,000 accidents (2005–2017)
  - **Features:**Accident_Severity, Weather_Conditions, Road_Type, Light_Conditions, Speed_limit, Vehicle_Type, Urban_or_Rural_Area, Time, Date, Location_Easting_OSGR, Location_Northing_OSGR
  - **Target Variable:** Accident_Severity (Slight / Serious / Fatal)
  - **Format:** CSV file (~200K rows)

- Any preprocessing steps required
  - Handle missing/inconsistent values (e.g., unknown weather).
  - Encode categorical attributes (e.g., One-Hot or Label Encoding)
  - Merge accident and vehicle datasets using Accident_Index
  - Normalize numerical features (speed limits, time bins).
  - Address class imbalance using resampling (e.g., SMOTE)

Proposed Methodology

- Overview of methods or models you plan to use
  - **Exploratory Data Analysis:** Identify trends, correlations, and patterns between conditions and severity
  - **Decision Tree Classifier:** Baseline interpretable model for accident severity prediction.
  - **Random Forest Classifier:** Ensemble model to improve accuracy and reduce overfitting.
  - **XGBoost Classifier:** Gradient boosting model for optimal performance and feature ranking.
  - **Model Evaluation:** Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
- Justification for choosing these methods, if any
  - Interpretable structure — easy to trace accident risk paths.
  - Handles categorical and numerical variables efficiently.
  - Suitable for explainable safety insights.
- Tools/Technologies (e.g., Python, libraries)
  - pandas, numpy, scikit-learn, xgboost, matplotlib, seaborn
  - Python

Implementation Plan

- Project phases (data collection, model building, testing)
  - **Data Understanding & Collection** – Import and explore Kaggle dataset
  - **Data Cleaning & Preprocessing** – Handle missing values, encode categories, and merge.
  - **Exploratory Data Analysis (EDA)** – Explore data distributions and relationships.
  - **Model Building** – Train Decision Tree, Random Forest, and XGBoost models.
  - **Model Evaluation** – Compare metrics and identify best-performing model.

- Timeline and milestones
  - Week1 - Data understanding & setup
  - Week2 - Data cleaning & preprocessing
  - Week3 - Exploratory data analysis
  - Week4 - Model building
  - Week5 - Model evaluation & optimization

- Resources required
  - **Software:** pandas, numpy, scikit-learn, xgboost, matplotlib, seaborn
  - **Dataset:** Kaggle – UK Road Safety Dataset

Challenges and Risks

- Potential risks or challenges
    - Missing or inconsistent data entries (e.g., unknown weather/light).
    - Highly imbalanced target classes (few “Fatal” cases).
    - Complex feature interactions (weather × light × road type).
    - Risk of overfitting in tree-based models

- How you plan to mitigate them
    - Use imputation and standardization techniques.
    - Apply class-balancing methods (SMOTE, class weights)
    - Use pruning or cross-validation for robust model generalization
    - Evaluate interpretability via feature importance and SHAP values.

Expected Outcome

- What do you expect to achieve?
  - Cleaned, preprocessed dataset ready for ML modeling
  - Three trained classification models (Decision Tree, Random Forest, XGBoost).
  - Model performance comparison and explainability results.
  - Visual feature importance charts highlighting top severity predictors.

- How will you measure success?
  - Model Performance Metrics:
  - High model accuracy and F1-score for severe accidents.
  - Interpretability and transparency in feature influence.
  - Actionable insights for policymakers, traffic authorities, and emergency planners.

Role and Responsibilities

- Anil
  - Data Collection & Preprocessing
  - EDA & Feature Engineering
- Gomathi:
  - Data Collection & Preprocessing
  - EDA & Feature Engineering
- Neeraj:
  - Model Building & Optimization
  - Evaluation & Reporting
- Ritesh:
  - Model Building & Optimization
  - Evaluation & Reporting

Data Science Canvas

Project: Road Safety: Accident Severity Prediction

Team: Data Warriors

Problem Statement

Business Case & Value Added

Predicting accident severity helps improve road safety policies and emergency response readiness.

Data Landscape

Data Required:
Accident severity, weather, light, road type, speed limit, urban/rural flag, vehicle type, number of casualties.

All required attributes are available in the Kaggle dataset.
Future extensions could include real-time traffic or weather integration.

Model Selection

Classification models designed for interpretability and accuracy:
Decision Tree: baseline model (interpretable structure).
Random Forest: ensemble model for better generalization.
XGBoost: gradient boosting for optimized predictive performance.
.

Model Requirements

Accurate and interpretable multi-class classifier.
Handles mixed data types (categorical + numerical).
Robust to missing values and imbalanced data.
Evaluated using: Accuracy, Precision, Recall, F1-score, Confusion Matrix
.

Software & Libraries

Python, Jupyter/Colab, pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, Streamlit (for app).

Skills

Python programming, data preprocessing, visualization, classification modeling, model evaluation.

Execution & Evaluation

Model Evaluation

Metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix.
Validation: Stratified train-test split, k-fold cross-validation.
Compare Decision Tree vs Random Forest vs XGBoost.

Data Storytelling

Use visualizations and dashboards to:
Show key influencing factors (e.g., weather, time of day, road type).
Compare model results (feature importance, decision paths).
Communicate findings clearly to policymakers and public safety units
.

Data Collection & Preparation

Data Selection & Cleansing

Merge accident and vehicle records.
Clean missing or inconsistent entries (weather/light).
Standardize categorical codes.
Engineer contextual features (day/night, urban/rural).

Data Integration

Join Accident and Vehicle datasets via Accident_Index.
Handle one-to-many relationships (multiple vehicles per accident).
Cleaned dataset fed into ML pipeline for model training and evaluation.

Data Collection

Source: Kaggle dataset from UK Department for Transport.
Format: CSV files downloaded from official open data portal.
Content: Accident- and vehicle-level details with date, time, severity, and conditions..

Explorative Data Analysis

Study variable distributions and correlations.
Analyze severity trends by weather, lighting, and road type.
Visualize spatial and temporal accident patterns.
Summarize key insights to guide model design and explainability.