PROJECT AND DATA MANAGEMENT PLAN *Aishwarya Shekar Babu (23093811)*

Project title	Machine Learning prediction of Road Accident Severity using Urban factors,vehicle data and Weather conditions
Research Question	**Research Question 1:** which machine learning approach is optimal in accurately predict the severity of UK road accidents using environmental (e.g. weather), temporal (e.g. time of day, season), spatial (e.g. urban/rural), and vehicle data which features contribute most to severe outcomes?
Project Objective	• Clean and preprocess the 2005–2017 STATS19 dataset (which includes weather, light, and road-surface fields) and Engineer features (Hour, Month, Day_of_Week) and encode categorical fields • Balance the training set to address severe-outcome imbalance (Slight, Serious, Fatal). • Train and compare four models—Logistic Regression (from scratch), custom MLP, Random Forest, and XGBoost—for multiclass severity prediction. • Evaluate models using F1Score, precision/recall, and confusion matrices. • Use feature-importance and SHAP to identify which environmental and spatial factors (e.g., weather, rural roads, high speed limits) most drive "Serious"/"Fatal" outcomes. • Build a weather-only XGBoost classifier to assign Low/Moderate/High risk levels to each weather category. • Provide actionable insights for targeted road-safety interventions based on 2005–2017 trends.

Summary of Project topic and background

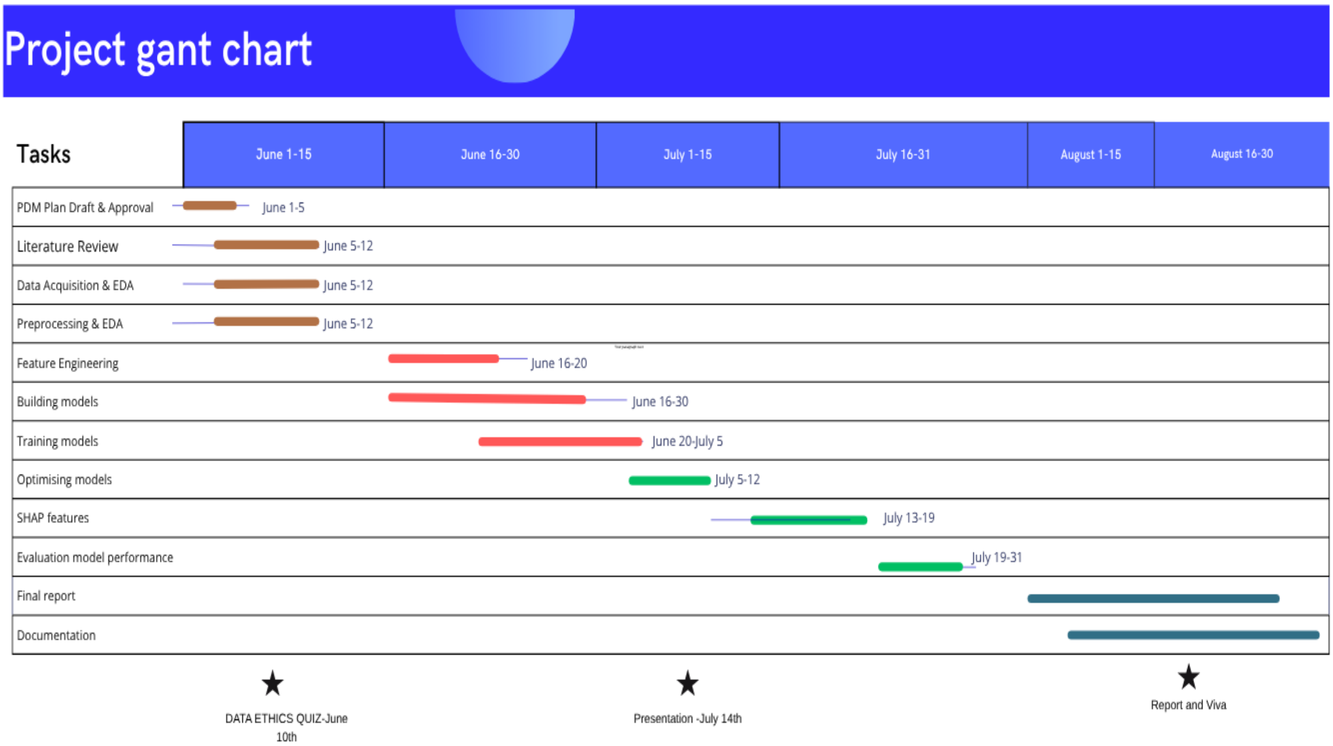
Road traffic accidents in the UK result in over 100,000 reported collisions annually, costing more in healthcare, emergency response, and economic losses . Although the STATS19 dataset records detailed factors—weather, light conditions, road surface, speed limits, and precise latitude/longitude—only a small fraction of collisions are classified as Serious or Fatal, making reliable severity prediction challenging. This project will compare four machine-learning approaches—multiclass Logistic Regression (from scratch), a custom two-layer MLP, Random Forest, and XGBoost—to forecast if a collision will be Slight, Serious, or Fatal. After balancing, models will be evaluated using accuracy, macro-F1, and confusion matrices to identify the optimal algorithm and, via feature-importance and SHAP analysis, reveal which environmental and spatial features (e.g. heavy rain, rural roads, high speed limits) most strongly drive severe outcomes. Additionally, a weather-only XGBoost model will estimate risk levels (Low/Moderate/High) per weather category (RQ2), and a analysis will pinpoint geographic hotspots for serious/fatal collisions (RQ3), informing targeted interventions.

References

Shirwaikar, R.D., Simha, A.H., Narvekar, A. & Pradeep, T.G., 2024. Predicting Accident Severity using Machine Learning. *Journal of Electrical Systems*, 20(6s), pp.2733–2746. Available at: Predicting Accident Severity using Machine Learning

Predicting and Analysing Road Traffic Injury Severity Using Boosting-Based Ensemble Learning Models Zhao, X., Zhou, H., Wang, J. & Ma, Y., 2022. Predicting and Analyzing Road Traffic Injury Severity Using Boosting-Based Ensemble Learning Models. *International Journal of Environmental Research and Public Health*, 19(18), p.11296. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8910532/

Crash Severity Analysis and Risk Factors Identification Based on an Interpretable Machine Learning Approach Wang, R., Liu, Q. & Shi, Y., 2023. Crash Severity Analysis and Risk Factors Identification Based on an Interpretable Machine Learning Approach. *Accident Analysis & Prevention*, 198, p.106908. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9732348/



Task	**Description**	**Start Date**	**End Date**
PDM Plan Draft & Approval	Develop the PDM Plan document; meet with supervisor and finalize for submission.	June 1 2025	June 5 2025
Literature Review	Conduct a detailed review of ML methods for crash-severity prediction and data ethics.	June 5 2025	June 12 2025
Data Acquisition & EDA	Download dataset inspect columns, handle missing values, and create initial plots.	June 5 2025	June 12 2025
Preprocessing & EDA	Clean raw data: drop unused columns, parse Date/Time, impute missing numeric, and plot distributions.	June 5 2025	June 12 2025
Feature Engineering	Derive features and cap outliers.	June 16 2025	June 20 2025
Building Models	Implement baseline model skeletons (LogReg, MLP, RF, XGBoost) on raw feature set.	June 16 2025	June 30 2025
Training Models	Train each classifier (LogReg, MLP, RF, XGBoost) on balanced data; log loss & accuracy.	June 20 2025	July 5 2025
Optimising Models	Tune hyperparameters (grid search/learning rates) for MLP, RF, and XGBoost.	July 5 2025	July 12 2025
SHAP Feature Analysis	Compute feature-importance and SHAP values for best model; plot top predictors.	July 13 2025	July 19 2025

Task	Description	Start Date	End Date
Evaluation (Model Performance)	Evaluate final models on test set: compute accuracy, macro-F1, confusion matrices.	July 19 2025	July 31 2025
Final Report	Write up Introduction, Methods, Results, Discussion, and Conclusion; format and proofread.	August 1 2025	August 30 2025
Documentation	Create README, finalize figures, back up all code/data to GitHub/OneDrive.	August 1 2025	August 30 2025

Data Management Plan

Overview of the Dataset:
UK STATS19 “Accidents.csv” and “Vehicles.csv” (2005–2017), collected by UK police and published by the Department for Transport under OGL v3.0. Accidents.csv (~1.98 M rows) records collision details (severity, date, time, location, weather, light, road conditions, speed limit, vehicle/casualty counts). Vehicles.csv (~3.2 M rows, ~20 columns, ~8 GB) includes vehicle type, driver age band, gender, and fuel type, merged via Accident Index.

Data Collection:
Downloaded from Kaggle: https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles

Original source: The data come from the Open Data website of the UK government, where they have been published by the Department of Transport.

Metadata:
Raw CSVs (UTF-8) total ~12 GB. A README file that entails full documentation on dataset, preprocessing steps, and usage instructions.

Document Control:
GitHub repo: https://github.com/aishwarya-shekar-babu/Road-Accident-Prediction

Weekly commits; branches: main

ReadMe File:
Includes project overview, data sources, setup instructions usage), directory structure, and licensing.

Security & Storage:
Code and small data (<100 MB) on GitHub (public); raw/processed files on OneDrive/Google Drive (weekly backups) with supervisor access.

Ethical Requirements:
All STATS19 data are fully anonymized—no personal names, addresses, or unique identifiers—and latitude/longitude are coarse, so no GDPR “special category” data are processed. As a secondary analysis of public data, the project complies with UH ethical policies without requiring new human-subject approvals. The Department for Transport publishes these datasets under the Open Government Licence v3.0, explicitly permitting academic use with proper attribution. Finally, the original STATS19 collection by UK police followed standardized, lawful protocols for recording accident, vehicle, and weather details, ensuring the data were gathered ethically.