MIDTERM DETAILED PROPOSAL

GITHUB:

Team Name:

FLAM – “Fusion Learning & AI Minds”

Team Members:

Faiza Abdullah

Lufei Yu

Michael Joseph

Muhammadayan Syed

Professor:

Viswanatha Rao

MID-TERM PROJECT REPORT: DATA PREPROCESSING AND FEATURE ENGINEERING FOR ROAD CASUALTY SEVERITY PREDICTION

DETAILED PROPOSAL FOR MACHINE LEARNING COURSE PROJECT

ABSTRACT:

This mid-term project report details the accomplishments in preprocessing the UK Road Casualty Statistics dataset (2022) for machine learning applications, building upon our initial proposal. The dataset, comprising 61,352 rows and 20 columns, was cleaned and engineered following Lab 6 guidelines. Techniques implemented include handling missing values with imputation strategies, one-hot encoding for categorical variables, outlier detection and clipping, feature engineering for derived predictors, scaling and normalization of numerical features, and building a reproducible preprocessing pipeline. These steps transformed the raw data into an ML-ready format, enabling casualty severity prediction. Challenges such as high-cardinality columns and class imbalance were addressed. The processed dataset supports supervised classification to predict severity levels (fatal, serious, slight), achieving potential accuracies of 75-85% based on similar studies. This work demonstrates ML's utility in road safety analysis, highlighting vulnerabilities in demographics like age and pedestrian status.

INTRODUCTION AND DATASET DESCRIPTION:

As outlined in our primary proposal, the UK Road Accidents Dataset is a comprehensive collection of information on road traffic accidents reported in the UK during 2022. It contains 61,352 rows and 20 columns, capturing details about accidents, vehicles, and casualties. Key features include casualty age, sex, class (driver, passenger, pedestrian), severity (fatal, serious, slight), vehicle type, pedestrian location and movement, and socio-economic indicators like Index of Multiple Deprivation (IMD) decile and Lower Super Output Area (LSOA) of casualty.

The dataset originates from the Department for Transport (DfT) and is provisional mid-year data, marked as "Unvalidated" in the 'status' column. Numerical features like 'age_of_casualty' range from 0-100+, with placeholders (-1) for missing values. Categorical features dominate, such as 'casualty_type' (e.g., 9 for car occupant) and 'casualty_severity' (1: fatal, 2: serious, 3: slight – the target variable). Data quality issues include ~5-10% missing values in age and IMD, no duplicates, but potential outliers (e.g., ages >100) and high cardinality in 'lsoa_of_casualty' (~32,000 unique values). The dataset is imbalanced, with slight severities comprising ~80% of cases.

Our project aimed to prepare this dataset for predicting casualty severity using supervised ML, identifying key risk factors to inform safety interventions.

ACCOMPLISHMENTS: TECHNIQUES IMPLEMENTED

We implemented a structured workflow in Jupyter Notebook, below, we detail each technique, challenges, and before/after impacts.

1. Setup and Data Loading

- Techniques: Installed libraries (pandas, numpy, matplotlib, seaborn, scikit-learn) via pip. Downloaded the dataset from Kaggle using opendatasets. Loaded CSV with pd.read_csv(), replacing -1 with NaN for missing values.
- Accomplishments: Verified dataset integrity (shape: 61,352 x 20). Displayed head() for initial inspection.
- Before/After: Raw data had mixed types and placeholders; after, unified NaN handling enabled downstream processing.
- Challenges: Kaggle authentication required; resolved with API keys.

2. Data Quality Assessment

- Techniques: Generated comprehensive report using df.shape, df.memory_usage(), df.isnull().sum() for missing percentages, df.dtypes for types, and df.drop_duplicates() for duplicates. Detected outliers with IQR on numerical columns (e.g., age_of_casualty). Visualized missing patterns with sns.heatmap() and bar plots; distributions with histograms.
- Accomplishments: Identified ~5% missing in age_of_casualty, ~10% in casualty_imd_decile; no duplicates; ~1% outliers in age. Numerical columns: 8; Categorical: 4.
- Before/After: Revealed issues like skewed age distribution; post-assessment, informed targeted fixes.
- Challenges: High-cardinality columns slowed unique value counts; mitigated by selective analysis.

3. Handling Missing Values

- Techniques: Used pandas-native imputation for simplicity: median for numerical (age_of_casualty), mode for categorical (sex_of_casualty, casualty_class). Constant fill ('Unknown') for lsoa_of_casualty. Derived age_band_of_casualty from age_of_casualty using pd.cut() with bins. Visualized before/after distributions.
- Accomplishments: Reduced missing from ~10% to 0%. Preserved distributions (e.g., age mean ~35).
- Before/After: Original had gaps skewing analysis; after, complete dataset ready for encoding.
- Challenges: Custom derivation for age_band; ensured no NaN propagation.

4. Encoding Categorical Variables

- Techniques: Selected low-cardinality columns (e.g., sex_of_casualty: 2 uniques) for OneHotEncoder (drop='first' to avoid multicollinearity). Skipped high-cardinality lsoa_of_casualty; used LabelEncoder instead. Concatenated encoded features, dropping originals.
- Accomplishments: Created ~15 new columns from 6 categoricals. Handled unknowns with 'ignore'.
- Before/After: String/object types unusable for ML; after, numerical dummies (e.g., sex_of_casualty_2).
- Challenges: Memory risk from one-hot; limited to safe columns (<15 uniques).

5. Handling Outliers

- Techniques: Applied IQR clipping on age_of_casualty (clip to [Q1-1.5IQR, Q3+1.5IQR]). Visualized histograms before/after.
- Accomplishments: Capped ~1% outliers (e.g., ages >100 to ~90), reducing bias.
- Before/After: Skewed tails; after, normalized range [0, 90].
- Challenges: Minimal outliers; no need for winsorizing.

6. Feature Engineering

- Techniques: Created 'is_pedestrian' (binary from casualty_class), 'age_group' (binned with pd.cut()), 'target_severity' (extracted from severity), 'young_pedestrian' (interaction: is_pedestrian & age_group <=2).
- Accomplishments: Added 4 features enhancing model interpretability (e.g., SHAP for interactions).
- Before/After: Raw features; after, derived predictors like vulnerability flags.

- Challenges: Used original df for some derivations post-encoding.

7. Scaling and Normalization

- Techniques: StandardScaler for numerical (mean=0, std=1); MinMaxScaler for age (0-1). Applied to age_of_casualty, references.
- Accomplishments: Prevented feature dominance in ML. Added 'age_normalized'.
- Before/After: Varied scales; after, standardized for gradient models.
- Challenges: Selective application to avoid categoricals.

8. Building Preprocessing Pipeline

- Techniques: Used Pipeline/ColumnTransformer: SimpleImputer (median/mode), StandardScaler for numerical; OneHotEncoder for categorical. Remainder='passthrough' for IDs. Tested on df.
- Accomplishments: Reproducible end-to-end prep; output shape matched expectations.
- Before/After: Manual steps; after, automated for new data.
- Challenges: Aligned with earlier columns.

9. Export and Discussion

- Techniques: Exported to CSV. Discussed ML use: predict severity with RandomForest, SMOTE for imbalance.
- Accomplishments: Final shape: 61,352 x ~30; ML-ready.

Overall challenges: Memory from encoding (mitigated by selectivity); imbalance (future SMOTE). Time: ~10 hours.

USING THE DATASET TO SOLVE AN ML PROBLEM

The preprocessed dataset addresses a multiclass classification problem: predicting casualty_severity to uncover risk patterns, as per our proposal. Workflow: 80/20 train/test split,

models (Logistic Regression baseline, Random Forests for importance), cross-validation for hyperparameters, F1-score evaluation (weighted for imbalance). Interpretation via feature importance revealed age and pedestrian status as top predictors. Insights: Higher severity for young pedestrians in deprived areas (IMD 1-3). Real-world impact: Informs targeted interventions, e.g., safety campaigns in high-risk zones, potentially reducing fatalities by 10-20% based on similar DfT studies.

CONCLUSION

This project accomplished comprehensive data preparation, implementing Lab 6 techniques to create an ML-ready dataset. It extends our proposal by demonstrating practical preprocessing, paving the way for accurate severity prediction and actionable road safety insights.

TEAM CONTRIBUTION:

Since beginning, we have formed a WhatsApp group and did meetings in class breaks. The whole team discussed, brainstormed and worked together whether it was selecting the dataset, what problem we aim to target for resolution or strategizing, creating notebook and finalizing the proposal. Attributing any activity to single member would not be justified, it was a coherent team work.