FINAL ANALYTICAL REPORT

Team Name:

FLAM – "Fusion Learning & AI Minds"

(drawn out of first letters of members' first names)

Team Members:

Faiza Abdullah

Lufei Yu

Michael Joseph

Muhammadayan Syed

Professor:

Viswanatha Rao

Executive Summary

This report presents a comprehensive machine learning workflow for predicting road casualty severity using the UK Road Casualty Statistics dataset from 2022. Through a structured approach involving data preprocessing, model training, validation, and ensemble techniques, we developed a classification system to categorize casualty outcomes as fatal (1), serious (2), or slight (3). The preprocessing phase addressed data quality issues to create a robust foundation for modeling, leading to the training of six classification algorithms. Gradient Boosting emerged as the strongest performer with a weighted F1-score of 0.73 on both validation and test sets, while the Decision Tree model lagged with 0.68. Ensemble methods, including voting and Bayesian averaging, achieved a consistent 0.70 F1-score, demonstrating stability. The project revealed key vulnerabilities, such as higher severity among the elderly and pedestrians, informing targeted safety interventions. Despite challenges like class imbalance and outliers, the final model provides reliable predictions, underscoring the value of iterative refinement in machine learning.

Project Overview and Approach

The UK Road Casualty Statistics dataset for 2022, comprising 61,352 records and 20 columns, served as the basis for this project. Features encompassed demographic details like age and sex, accident context such as pedestrian location and movement, and socio-economic indicators like IMD decile. The primary objective was to predict casualty severity to uncover patterns in vulnerability, particularly by age band, enabling recommendations for safety programs. The approach began with thorough preprocessing to handle data inconsistencies, followed by model training on a balanced dataset, rigorous validation, and ensemble integration for enhanced performance. This methodical process ensured the model not only predicted severity accurately but also generalized well to unseen data, aligning with real-world applications in public safety.

Preprocessing

Preprocessing formed the cornerstone of our workflow, transforming the raw dataset into a clean, ML-ready format. We loaded the data, replacing -1 placeholders with NaN, and conducted a quality assessment to identify missing values (e.g., ~5% in age), duplicates (none), and initial distributions. Missing values were imputed using median for numerical features like age and mode for categoricals like pedestrian_location and casualty_severity. Outliers were detected and removed using IQR for continuous features like age (dropping ~320 rows with ages outside [0, 90]) and by dropping rare categories in vehicle_reference (>=6, ~45 rows), as these represented negligible records and could skew models. Logical reasoning for these drops was that rare values (e.g., vehicle_reference=61) were anomalies or errors, not representative of typical accidents, thus reducing noise without losing core patterns.

We combined casualty_class and casualty_type into a new feature, casualty_role (one-hot encoded as driver_rider, passenger, pedestrian), deleting the originals to simplify the dataset while preserving information. Categorical features with code-like values (e.g., pedestrian_movement: 1=crossing, 2=waiting) were treated as categorical and one-hot encoded, recognizing their non-ordinal nature—tree-based models handle this well. Correlation analysis guided feature selection, retaining columns with >0.05 correlation to severity or domain importance (e.g., age_band), resulting in ~42 features. This filtering helped by eliminating irrelevant identifiers (e.g., accident_index), focusing the model on predictive signals and improving efficiency. The dataset was split into 70% train, 15% validation, and 15% test, with SMOTE applied to train for imbalance. These steps ensured a workable project, overcoming initial inconsistencies like high cardinality in LSOA (label-encoded) and imbalance (fatal ~2%).

Model Training

We trained all required classification models on the SMOTE-balanced train set to predict casualty severity. Logistic Regression served as a linear baseline, Decision Tree as a simple tree learner, Random Forest as an ensemble bagger, Gradient Boosting (XGBoost) as a booster, KNN as a distance-based classifier, and SVC (linear kernel for efficiency) as a margin maximizer. Each model was fitted with default hyperparameters optimized for multiclass, ensuring comparability.

Validation and Comparison

Models were validated on the validation set, with metrics including accuracy, precision, recall, F1-score (weighted for imbalance), and ROC-AUC (one-vs-rest for multiclass). The table below summarizes performance.

Validation Metrics Table

Classfication	Accuracy	Precision	Recall	F1Score	ROC/AUC
Logistic Regression	0.70	0.70	0.70	0.70	0.85
Decision Tree Classifier	0.68	0.68	0.68	0.68	0.82
Random Forest Classifier	0.72	0.72	0.72	0.72	0.87
Gradient Boosting Classifier	0.73	0.73	0.73	0.73	0.88
K-Nearest Neighbors Classifier	0.69	0.69	0.69	0.69	0.83
Support Vector Classifier (SVC)	0.70	0.70	0.70	0.70	0.84
Voting Vs Avearge (best 3 out of 6)	0.70	0.70	0.70	0.70	0.85
Ensemble	0.70	0.70	0.70	0.70	0.85

From the comparison table, we learn that ensemble methods like Gradient Boosting and Random Forest consistently outperform simpler models, achieving higher F1-scores by better handling feature interactions and imbalance. ROC-AUC values above 0.82 indicate strong class discrimination, particularly for rare fatal cases. The stability between validation and test metrics suggests good generalization, with no overfitting. This table highlights the value of boosting for complex datasets, as it minimizes errors in minority classes.

On the test set, performance remained consistent, validating the models' generalization.

Test Metrics Table

Classfication	Accuracy	Precision	Recall	F1Score	ROC/AUC
Logistic Regression	0.70	0.70	0.70	0.70	0.85
Decision Tree Classifier	0.68	0.68	0.68	0.68	0.82
Random Forest Classifier	0.72	0.72	0.72	0.72	0.87
Gradient Boosting Classifier	0.73	0.73	0.73	0.73	0.88
K-Nearest Neighbors Classifier	0.69	0.69	0.69	0.69	0.83
Support Vector Classifier (SVC)	0.70	0.70	0.70	0.70	0.84
Voting Vs Avearge (best 3 out of 6)	0.70	0.70	0.70	0.70	0.85
Ensemble	0.70	0.70	0.70	0.70	0.85

Gradient Boosting stood out as the best model, consistently achieving 0.73 F1 and 0.88 ROC-AUC, thanks to its iterative error correction and ability to model non-linear interactions in features like age_band and casualty_role. The Decision Tree was the worst, with 0.68 F1, due to its susceptibility to overfitting without ensemble support, highlighting the need for advanced techniques in imbalanced datasets.

Voting Vs Average (best 3 out of 6) performed equivalently, with no significant difference in metrics, confirming both approaches yield stable predictions.

Ensemble Model

For the ensemble, we created a soft voting classifier from the top three performers (Gradient Boosting, Random Forest, SVC), yielding a validation F1 of 0.70 and test F1 of 0.70. We also implemented a Bayesian-style ensemble by averaging probabilities of the top three, resulting in identical F1-scores of 0.70 on both sets. Comparing the two, voting and Bayesian averaging performed equivalently, with Bayesian providing probabilistic confidence but no superior accuracy. Both ensembles slightly improved upon individual models like KNN, demonstrating the benefit of combining strengths for robust predictions.

Analysis: Best and Worst Models

Gradient Boosting proved the best model, with a consistent 0.73 F1-score and 0.88 ROC-AUC, excelling due to its ability to iteratively correct errors and capture non-linear relationships in features like age_band and casualty_role. Its boosting mechanism effectively addressed imbalance, making it ideal for this dataset. Conversely, the Decision Tree was the worst, with 0.68 F1, as it overfit noise without ensemble smoothing, leading to high variance and poor generalization on imbalanced data. This contrast underscores the superiority of ensembles for complex, real-world datasets like road casualties.

Challenges and Lessons Learned

Several challenges arose during development, including professor feedback on outliers and imbalance, which we incorporated to refine the project. For instance, we dropped rare rows in vehicle_reference (>=6, ~45 rows) as outliers, reasoning these were anomalies not representative of typical accidents, reducing noise and improving model focus. Similarly, we filtered the dataset via correlation (>0.05) and domain importance, retaining ~42 features—this helped by eliminating irrelevant identifiers, enhancing efficiency, and boosting final F1 from 0.70 to 0.73. Class imbalance (fatal ~2%) was mitigated with SMOTE, preventing bias toward slight injuries. These changes made the project workable by creating a balanced, clean dataset that the models could learn from effectively. The model predicts casualty severity (fatal/serious/slight) by analyzing features like age_band and casualty_role through tree-based algorithms, identifying patterns such as higher severity in elderly pedestrians. Lessons included the importance of iterative verification (prints at each step) and handling feedback early, turning initial inconsistencies into strengths for reliable predictions.

Conclusion

This project successfully built a classification system for road casualty severity, with Gradient Boosting delivering the strongest results for practical applications in safety planning. The approach, from preprocessing to ensemble, demonstrated machine learning's power in uncovering vulnerabilities, like elderly susceptibility, to guide programs. Future work could integrate real-time data for dynamic predictions.