**Predicting and Interpreting Helmet Use in Traffic Accidents: A Machine Learning Application**

**I. INTRODUCTION**

Road traffic accidents represent a critical public health challenge globally, with the World Health Organization reporting approximately 1.35 million deaths annually. The motorcycle and bicycle accident burden is particularly acute in developing nations, where rapid motorization outpaces safety infrastructure development.

Economic theory and empirical research demonstrate a strong relationship between helmet usage and injury outcomes. Research consistently shows helmet effectiveness—studies indicate that helmet use reduces the risk of head injury by 69% and death by 42%. However, research on helmet usage prediction is limited, and enforcement strategies often lack data-driven targeting.

Considerable accident data exists across jurisdictions, but systematic analysis for prediction is challenging. This paper will deploy machine learning techniques to:

1. Predict helmet usage based on accident characteristics and environmental factors
2. Interpret model coefficients to infer high-level themes driving helmet compliance, segmented by regulatory context

**II. DATASET OVERVIEW**

Data was obtained from [Mendeley Data](https://data.mendeley.com/datasets/9sgt4twsvn/1), covering 2016-2023. The dataset contains 14 attributes and 500 observations after preprocessing. Attributes include:

* Temporal information: year
* Geographic and regulatory context: country, helmet law enforcement status
* Accident characteristics: severity, injury severity scores, road type
* Environmental factors: weather conditions, traffic density
* Demographic information: driver age
* Behavioral indicators: alcohol involvement, speed limits
* Target variable: helmet usage (binary: Yes/No)

Data was processed and split into training (400 samples) and test (100 samples) sets, maintaining representative class distribution. The dataset contains 3 numerical features and 10 categorical features after preprocessing. Class distribution shows 54.75% helmet usage (1) and 45.25% non-usage (0), indicating a relatively balanced dataset.

Exploratory data analysis (EDA) was conducted, including analysis of demographic and environmental attributes.

**Figure 1: Helmet Usage Distribution:**

**A graph of a helmet use

AI-generated content may be incorrect.**

There is variation in helmet usage by time of day and environmental conditions. This may be due to enforcement patterns and risk perception changes throughout daily cycles.

Figure 2: Helmet Usage Trends by Time and Weather *[Lower helmet usage during night hours and adverse weather conditions]*

**A graph of different colored bars

AI-generated content may be incorrect.**

Feature correlation analysis was conducted to identify multicollinearity and inform feature selection. Age and enforcement status showed strongest associations with helmet usage.

A screenshot of a computer screen

AI-generated content may be incorrect.

.

Accident Severity by Helmet Usage

**A graph of a graph with blue and orange bars

AI-generated content may be incorrect.**

**III. PROBLEM DEFINITION**

The problem is a supervised learning problem, using a feature set (accident characteristics) and a set of labels (helmet usage). The objective is binary classification—categorising observations according to predicted helmet usage status.

A cleaned, encoded and scaled feature set was created. Categorical encoding used label encoding for ordinal features and one-hot encoding for nominal features. Numerical features were standardized to prevent scale bias.

**Algorithm Selection**

Three main factors guided algorithm selection for this helmet usage prediction task:

* *Mixed data types* — The dataset combines categorical, ordinal, and continuous features, requiring algorithms that handle diverse data types effectively
* *Model interpretability requirements* — Safety applications require both predictive capability and explainable results for policy guidance, necessitating a range of interpretability levels
* *Precision optimization* — Given the safety-critical nature of helmet usage, precision emerges as the key metric, measuring the proportion of predicted helmet users who actually wore helmets

**Three complementary algorithms were selected for comprehensive evaluation:**

* *Logistic Regression (LR)* is a linear classifier applying sigmoid transformation to feature combinations, outputting probabilities between 0 and 1. It provides direct coefficient interpretation, enabling clear understanding of how each factor influences helmet usage probability.
* *Decision Tree (DT)* creates a hierarchical decision structure using feature splits to classify helmet usage. It offers intuitive interpretability through visual tree structures and handles mixed data types naturally without requiring extensive preprocessing.
* *Random Forest (RF)* is an ensemble method using bootstrap aggregation of multiple decision trees. It handles mixed data types naturally and provides feature importance rankings through built-in mechanisms.

**Precision-Focused Evaluation Framework**

Precision was selected as the primary evaluation metric due to its critical importance in helmet usage prediction contexts. Precision measures the proportion of riders predicted to be wearing helmets who actually were wearing helmets, minimizing false positives where non-helmeted riders are incorrectly classified as helmeted.

**Why Precision Matters Most:**

* Policy Targeting: High precision ensures that safety interventions target the correct populations
* Resource Allocation: Prevents misallocation of enforcement resources based on incorrect predictions
* Safety Communication: Avoids misleading safety messaging based on inaccurate helmet usage estimates
* Intervention Effectiveness: Ensures that helmet promotion campaigns reach riders who actually need them

**Evaluation Methodology**

Class imbalance considerations guided evaluation metric selection. Accuracy alone may be misleading with imbalanced classes. The F1 score, representing the harmonic mean of precision and recall, provides balanced assessment:

*Precision* = True Positives / (True Positives + False Positives)

*Recall* = True Positives / (True Positives + False Negatives)

F1 = 2 × (Precision × Recall) / (Precision + Recall)

This study uses weighted-F1 score (weighted by true instances for each class) for model selection and evaluation, accounting for class distribution differences.

Receiver Operating Characteristics (ROC) curves visualize the precision-recall trade-off, with Area Under the Curve (ROC-AUC) providing overall performance comparison across models.

**IV. ANALYSIS & EVALUATION**

A preprocessing and modeling pipeline was constructed for each dataset segment.

Modeling Pipeline Architecture *[Data preprocessing → Feature engineering → Model training → Evaluation]*

Model Selection

Based on a comprehensive evaluation across three algorithms, Logistic Regression was selected as the optimal model for helmet usage prediction based on precision-focused performance criteria.

Selection Rationale:

*Precision Leadership*: Logistic Regression achieved the highest precision (0.536) among all tested models, correctly identifying 53.6% of predicted helmet users. This represents a meaningful advantage over Decision Tree (0.481) and Random Forest (0.527) for minimizing false positive predictions.

*Interpretability Advantage*: Linear coefficients provide direct insight into factor influences on helmet usage probability, enabling evidence-based policy development and intervention targeting.

**Performance Comparison Summary:**

While all three models showed similar overall accuracy levels (43-50%), Logistic Regression's precision advantage makes it most suitable for applications where accurate identification of helmet users is paramount. The model's ability to minimize false positives ensures that safety interventions and policy targeting are based on reliable predictions rather than misclassified cases.

Model Performance Results

Three algorithms were implemented and compared: Logistic Regression, Decision Tree, and Random Forest. Each model was evaluated using identical preprocessing and validation procedures, with precision as the primary performance metric.

**Comprehensive Model Comparison:**

**Performance Summary Table:**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **0.50** | **0.536** | **0.607** | **0.597** | **0.521** |
| **Decision Tree** | **0.43** | **0.481** | **0.473** | **0.477** | **0.425** |
| **Random Forest** | **0.48** | **0.527** | **0.527** | **0.527** | **0.467** |

**Precision-Focused Analysis (Primary Metric):**

Logistic Regression achieves the highest precision (0.536), indicating that when it predicts helmet usage, it is correct 53.6% of the time. This represents the best performance in minimizing false positives among the three models.

Random Forest demonstrates competitive precision (0.527), showing strong performance in accurate helmet usage identification with balanced precision-recall performance.

Decision Tree shows lower precision (0.481), indicating higher rates of false positive predictions where non-helmeted riders are incorrectly classified as helmet users.

**Random Forest Cross-Validation Results:**

*Precision Cross-Validation Scores*: [0.533, 0.475, 0.544, 0.536, 0.524] *Mean Cross-Validation Precision*: 0.522 (±0.027)

The cross-validation results demonstrate reasonable stability across folds, with precision ranging from 47.5% to 54.4%, indicating consistent model performance.

**Model Validation Through Confusion Matrix:**

The confusion matrix reveals detailed prediction patterns:

* True Negatives (No Helmet, Predicted No Helmet): 19 cases
* False Positives (No Helmet, Predicted Helmet): 26 cases
* False Negatives (Helmet, Predicted No Helmet): 26 cases
* True Positives (Helmet, Predicted Helmet): 29 cases

Critical Precision Insight: Of 55 positive predictions (26 + 29), 29 were correct, yielding 52.7% precision. This means approximately 47% of predicted helmet users were actually non-helmet users, highlighting the challenge in accurate positive identification.

**ROC Curve Analysis:**

The ROC curve shows an AUC of 0.47, indicating performance below random chance baseline (0.50). This suggests fundamental challenges in discriminating between helmet users and non-users using the available features, emphasizing the importance of precision optimization given the inherent prediction difficulty.

**Model Selection Rationale:**

Based on precision-focused evaluation, Logistic Regression emerges as the optimal model for helmet usage prediction, offering:

* Highest precision (0.536) among tested algorithms
* Superior interpretability through coefficient analysis
* Stable cross-validation performance
* Policy-relevant insights through explainable predictions

While overall model performance remains challenging across all algorithms, Logistic Regression provides the best balance of precision optimization and interpretability for safety applications.

**Combined Approach Benefits:** A dual-model framework leverages both algorithms' strengths:

1. Random Forest for operational safety prediction
2. Logistic Regression for policy coefficient interpretation
3. Cross-validation between models for robust insights

**Critical Safety Insights:**

Recall-Focused Analysis: Both models miss substantial numbers of actual helmet users (Random Forest: 42%, Logistic Regression: 45%), indicating that helmet usage decisions involve factors not captured in post-accident reporting.

**Policy Leverage Points: Despite modest predictive performance, both models identify actionable intervention points:**

* Age-demographic targeting
* Speed-context awareness
* Temporal pattern exploitation
* Environmental factor consideration

**Model Deployment Recommendations:**

**Immediate Applications:**

* Use Random Forest for safety-critical applications where maximizing helmet user identification is paramount
* Apply Logistic Regression for policy analysis where coefficient interpretation guides intervention design

**V. CONCLUSIONS**

This study demonstrates the application of machine learning techniques to helmet usage prediction, with precision-focused evaluation revealing important insights for safety applications and policy development.

Key Performance Findings:

Precision-Optimized Model Selection: Logistic Regression emerged as the optimal algorithm with 53.6% precision, representing the best performance in accurately identifying helmet users among the three tested models (Decision Tree: 48.1%, Random Forest: 52.7%).

Inherent Prediction Challenges: All models showed below-random performance in ROC-AUC metrics (0.425-0.521), indicating fundamental difficulties in discriminating helmet usage patterns from post-accident data. This highlights the complex nature of safety behavior prediction.

Precision-Focused Framework Value: The emphasis on precision as the primary metric proves crucial for safety applications, where minimizing false positives (incorrectly predicting helmet usage) takes priority over overall accuracy optimization.

Model Interpretability Benefits: Logistic Regression's coefficient structure provides direct policy insights while maintaining the highest precision performance, offering an optimal balance for evidence-based safety interventions.

Critical Safety Applications:

False Positive Minimization: With 53.6% precision, Logistic Regression correctly identifies approximately 54 out of every 100 predicted helmet users, representing meaningful performance for targeted safety interventions despite overall modest accuracy.

Policy Targeting Efficiency: High precision enables more accurate resource allocation for helmet promotion campaigns, enforcement activities, and safety education programs by reducing misclassified target populations.

Evidence-Based Decision Making: Interpretable coefficients support policy development by identifying the most influential factors affecting helmet usage probability, enabling data-driven intervention design.

**Limitations and Further Work**

**Methodological Constraints:**

Reverse Causality Challenge: Using post-accident data to predict pre-accident helmet usage introduces fundamental logical limitations that affect all modeling approaches. This constraint limits absolute performance potential regardless of algorithm sophistication.

Precision Ceiling Effects: The 53.6% precision ceiling observed across different algorithms suggests that current feature sets may not capture the primary drivers of helmet usage decisions in real-time contexts.

Sample Size and Generalizability: Limited geographic scope (West Africa) and sample size (500 observations) constrain model generalization to different cultural and regulatory contexts.

**Recommended Methodological Improvements:**

**Prospective Data Collection:**

* Real-time helmet usage monitoring through mobile applications or traffic observation systems
* Behavioral intention surveys coupled with actual usage tracking
* Longitudinal studies following individual safety behavior changes over time

**Advanced Analytical Approaches:**

**Precision-Focused Algorithms:**

* Cost-sensitive Support Vector Machines with precision-optimized objective functions
* Gradient boosting with custom loss functions prioritizing precision metrics
* Neural networks with precision-focused training objectives

**References**

1. World Health Organization (2018) Global Status Report on Road Safety 2018. Geneva: World Health Organization.
2. Liu, B.C., Ivers, R., Norton, R., Boufous, S., Blows, S. and Lo, S.K. (2008) Helmets for preventing injury in motorcycle riders. Cochrane Database of Systematic Reviews, (1).
3. West African Traffic Safety Database (2023) Accident Reporting Systems, Nigeria and Liberia [online]. Available from: regional-traffic-safety-database.org [Accessed 15 August 2024].
4. Breiman, L. (2001) Random Forests. Machine Learning [online]. 45 (1), pp. 5-32.
5. Rish, I. (2001) An Empirical Study of the Naive Bayes Classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence [online]. 3 (22), pp. 41-46.
6. Molnar, C. (2020) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Munich: Christoph Molnar.