**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.

Identifying predictive factors can guide targeted interventions and enforcement strategies, but measuring intervention effectiveness requires robust predictive frameworks. There are two main regulatory contexts. Strict enforcement jurisdictions (such as urban areas with active traffic police) generally achieve higher compliance rates through visible enforcement and penalty systems. Limited enforcement contexts (such as rural areas with minimal police presence) show lower compliance despite helmet laws, often relying on education and community-based interventions

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

**2. DATASET OVERVIEW**

Data were obtained from the traffic accident [Mendeley](https://data.mendeley.com/datasets/9sgt4twsvn/1), originally collected from accident reporting systems in West Africa, covering the period from 2016 to 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.

The feature set includes both numerical variables (age, year, injury severity scores) and categorical encodings (country, weather conditions, enforcement status, etc.).

Exploratory data analysis (EDA) was conducted, including analysis of demographic and environmental attributes. . Class distribution shows 54.75% helmet usage (1) and 45.25% non-usage (0), indicating a relatively balanced dataset

A graph of a helmet use

AI-generated content may be incorrect.

A group of green and orange bars

AI-generated content may be incorrect.

**3. PROBLEM DEFINITION**

This is a supervised binary classification task: given accident-level features describing the event, environment, and rider characteristics, the model predicts whether the rider was wearing a helmet (Helmet\_Use\_Binary). The goal is to produce reliable, generalizable predictions of helmet usage that can inform operational decisions and targeted safety interventions.

Binary classification performance — Healthcare and safety prediction tasks require algorithms proven effective for binary outcomes with class imbalance considerations (He and Garcia, 2009).

Three complementary algorithms were selected for comprehensive evaluation:

Random Forest Tree-based feature importance allows identification of key predictive factors without assumptions about linear relationships.

Decision Tree (DT) creates hierarchical decision structures through recursive binary splitting, as formalized by Quinlan (1986). It offers intuitive interpretability through visual tree structures and handles mixed data types without preprocessing requirements (Rokach and Maimon, 2005).

Logistic Regression (LR) applies the logistic function to linear combinations of features, providing probabilistic binary classification (Hosmer et al., 2013).Key limitations include linearity assumptions and sensitivity to outliers, addressed through appropriate feature engineering and regularization techniques (James et al., 2013).

**4. Analysis and Evaluation**

This section details the evaluation framework, compares the performance of the selected models, and provides a rationale for choosing the best approach for both policy interpretation and predictive accuracy.

Evaluation Methodology and Metrics

The evaluation focuses on the trade-off between identifying helmet users correctly and avoiding misclassifications.

* Precision: Measures the accuracy of positive predictions. In this context, it answers: "Of all the riders the model predicted were wearing a helmet, what proportion actually were?" High
* precision is critical for efficiently targeting interventions, ensuring resources are not wasted on riders who are already compliant.

**Machine Learning Modeling Pipeline**

**1.Feature Engineering & Selection:**

- Create a binary target variable, Helmet\_Use\_Binary, from the cleaned helmet usage column.

- Define the feature set (X) by selecting relevant columns and the target variable (y).

**2. Preprocessing & Transformation:**

**-Data Splitting:** Divide the data into training and testing sets using a stratified split to ensure the class proportions of helmet usage are maintained in both sets.

**-Encoding:** Apply one-hot encoding to nominal categorical features to convert them into a numerical format suitable for machine learning models.

**-Scaling:** Standardize numerical features using StandardScaler to bring them to a common scale, preventing features with larger ranges from dominating the model.

3. **Pipeline Creation:**

- Use ColumnTransformer and Pipeline from scikit-learn to chain all preprocessing steps (imputation, encoding, scaling) together. This ensures consistency and prevents data leakage from the test set into the training process.

**4.Model Training & Comparison:**

- Train three different classification algorithms:

**- Logistic Regression:** For a simple, interpretable baseline.

**- Decision Tree:** To capture non-linear patterns and generate simple rules.

**- Random Forest:** An ensemble model for improved accuracy and robustness.

Fit each model on the preprocessed training data.-

**5.Model Evaluation & Selection:**

**Performance Metrics:** Evaluate each model on the unseen test set using a comprehensive set of metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

**Confusion Matrix:** Generate a confusion matrix for each model to visualize its performance in distinguishing between true and false predictions for each class.

**Cross-Validation:** Perform k-fold cross-validation on the most promising model to verify its stability and ensure its performance is not due to a random chance split of the data.

**Final Selection:** Choose the final model(s) based on the project's primary goal, balancing the trade-offs between interpretability (Logistic Regression) and predictive power (Random Forest).

* Recall (Sensitivity): Measures the model's ability to find all actual positive instances. It answers: "Of all the riders who truly were wearing a helmet, whaproportion did the model correcidentify?" High recall is important if the goal is to study the outcomes of the entire helmet-wearing population
* F1-Score: The harmonic mean of precision and recall. It provides a single, balanced measure of a model's performance, making it an excellent primary metric for comparing models when both false positives and false negatives carry costs.
* ROC-AUC (Area Under the Receiver Operating Characteristic Curve): This metric evaluates the model's overall ability to discriminate between helmet-wearers and non-wearers across all possible classification thresholds. An AUC of 0.5 indicates performance equivalent to random chance, while an AUC of 1.0 signifies perfect classification.

**5.Model Performance Comparison: Training vs. Test Results**

Three models were evaluated: Logistic Regression, Decision Tree, and Random Forest. A key part of the evaluation was comparing their performance on the training data versus the unseen test data to check for overfitting.

* Training Performance:
  + The Decision Tree and Random Forest models achieved near-perfect scores on the training data, a strong indicator of overfitting. They memorized the training data instead of learning generalizable patterns.
  + The Logistic Regression model had more modest scores on the training data, suggesting it was learning more general patterns.

A screenshot of a computer

AI-generated content may be incorrect.

* Test Performance: The models' true performance was measured on the test set:

A screenshot of a computer

AI-generated content may be incorrect.

**6. Conclusion**

The Logistic Regression model was selected as the best model for this task. It was the only model that showed stable performance between the training and test sets, and it achieved the highest Precision and ROC-AUC scores on the test data, making it the most reliable and interpretable choice.

Detailed Validation of the Logistic Regression Model

The selected Logistic Regression model underwent further validation to confirm its stability and understand its predictive behavior

Cross-Validation:

5-fold cross-validation was performed, yielding precision scores of [0.533, 0.475, 0.544, 0.536, 0.524].

The mean precision was 0.522, with a low standard deviation, confirming the model's stable performance across different subsets of the data.

Confusion Matrix Analysis:

A chart with different colored squares

AI-generated content may be incorrect.

CONFUSION MATRIX BREAKDOWN

* **Confusion Matrix Interpretation:**
  + The model correctly identified **29 helmet-wearers (True Positives)** and **19 non-wearers (True Negatives)**.
  + However, it made **26 False Positive** errors (incorrectly labeling a non-wearer as a wearer) and **26 False Negative** errors (failing to identify a helmet-wearer).
  + This highlights the primary challenge: the model struggles to reliably distinguish between the two classes, leading to a significant number of errors.

II. KEY PERFORMANCE METRICS ANALYSIS

Overall Accuracy: 50%

* Interpretation: The model correctly predicts helmet usage in exactly half of all cases
* Baseline Comparison: Performance at chance level, indicating fundamental prediction challenges
* Implication: Model provides limited discriminative capability between helmet users and non-users

Sensitivity/Recall: 67% (Critical Safety Metric)

Negative Predictive Value (NPV): 42%

**Conclusions & Final Recommendations**

This study successfully applied machine learning techniques to the complex challenge of predicting helmet usage. The analysis demonstrates that while the problem is inherently difficult with the available data, a carefully selected and validated model can still provide significant value for policy and safety applications.

Key Findings

Logistic Regression as the Optimal Model: While three algorithms were tested, Logistic Regression emerged as the most suitable model. It was the only algorithm that provided stable and generalizable performance, achieving the highest precision (53.6%) and ROC-AUC (0.521) on the unseen test data. In contrast, the Decision Tree and Random Forest models suffered from severe overfitting, making them unreliable for this task.

The Challenge of Prediction is a Core Insight: A critical finding is the limited predictive power of the available features. The low ROC-AUC scores across all models indicate that factors in the dataset are not strong discriminators of helmet-wearing behavior. This itself is a crucial insight, suggesting that future efforts must focus on acquiring more informative data to significantly improve predictive accuracy.

Stability and Reliability Over Raw Power: The Logistic Regression model's performance was validated through 5-fold cross-validation, which showed consistent precision scores (mean of 52.2% ±2.7%). This stability makes it a trustworthy baseline model, even if its overall performance is modest.

The Value of a Precision-Focused Framework: The decision to prioritize precision proved essential. In a real-world application, the primary cost is misallocating resources by targeting riders who are already wearing helmets (false positives). By optimizing precision, the Logistic Regression model, despite its limitations, becomes a practical tool for improving the efficiency of safety interventions.

**Actionable Implications for Safety Programs**

For Policy and Interpretation: The Logistic Regression model is immediately useful. Its coefficients can be analyzed to understand which factors are most strongly correlated with helmet use, providing an evidence-based foundation for designing new policies or educational campaigns.

For Targeted Interventions: With a precision of 53.6%, the model provides a better-than-random tool for identifying potential groups of non-compliant riders. While not perfect, it allows for more focused resource allocation than a random approach, ensuring that safety campaigns are more likely to reach their intended audience.

**7. Final Recommendation**

The Logistic Regression model should be adopted for its dual benefits of interpretability and stable, precision-focused performance. It is recommended to use this model to derive policy insights and guide strategic safety planning. Future work should be directed at enhancing the feature set to build a more powerfully predictive model, with the current analysis serving as a robust and informative baseline.

**References**

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied logistic regression*. 3rd ed. Hoboken: John Wiley & Sons.

Kotsiantis, S.B. (2007) Supervised machine learning: a review of classification techniques. *Informatica*, 31 (3), pp. 249-268.

Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, 1 (1), pp. 81-106.Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1 (5), pp. 206-215.

Zheng, A. and Casari, A. (2018) *Feature engineering for machine learning: principles and techniques for data scientists*. Sebastopol: O'Reilly Media.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An introduction to statistical learning: with applications in R*. New York: Springer.

Zheng, A. and Casari, A. (2018) *Feature engineering for machine learning: principles and techniques for data scientists*. Sebastopol: O'Reilly Media.

World Health Organization (2022) *Helmets: A Road Safety Manual for Decision-Makers and Practitioners*. Geneva: WHO Press

**Total word count: 1730**