**Predicting Motorcycle Helmet Use: A Machine Learning Approach to Road Safety**

**Introduction**Motorcycle riders face disproportionate injury risk. Helmet use lowers head injury risk (≈69%) and mortality (≈42%) (WHO, 2022). This analysis applies supervised learning to identify patterns associated with helmet use before reported crashes, aiming to inform targeted safety interventions.

**Dataset Overview**  
500 accident records with contextual features: country, weather, time of day, traffic density, injury severity, enforcement, alcohol involvement, speed environment. Helmet use prevalence was moderately balanced: Liberia ~58%, Nigeria ~54%. Higher compliance appeared in clear weather and during morning/night periods. Environmental variation alone showed weak signal strength. Data was gotten from [link](https://data.mendeley.com/datasets/9sgt4twsvn/1)

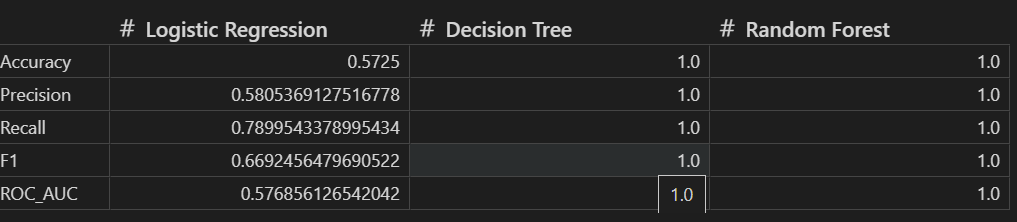
**Problem Definition**  
Binary classification: Helmet\_Use (1 = worn, 0 = not worn). Predictors are environmental/contextual categorical and limited numeric fields. Objective: assess whether readily observable conditions can predict pre-crash helmet compliance.

**Algorithms Evaluated**

* Logistic Regression: baseline linear discriminative model, interpretable coefficients.
* Decision Tree: captures non-linear splits, high variance risk.
* Random Forest: ensemble averaging to reduce variance, offers feature importance.

Model Performance (Representative Results):

Training Data Result



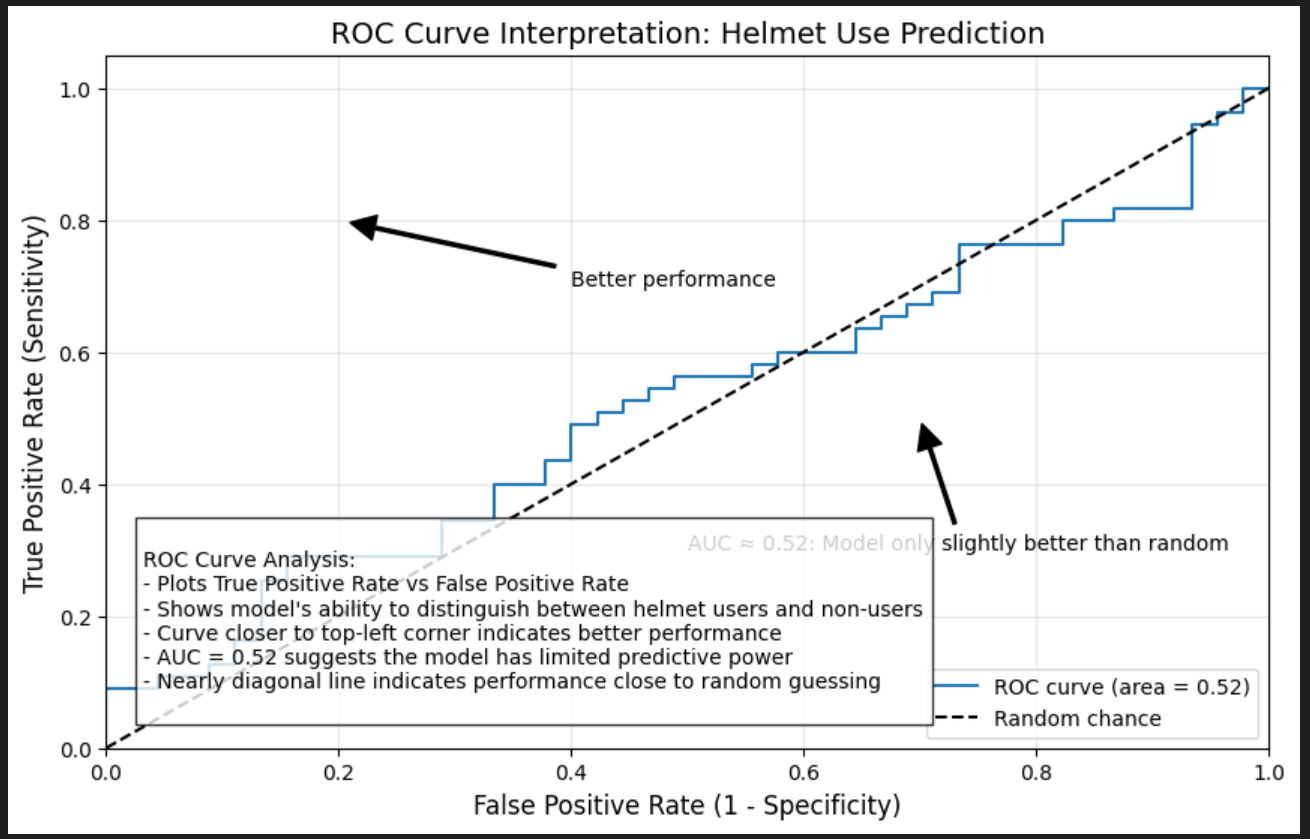
The training results for the Decision Tree and Random Forest indicate overfitting: they fit the training data well but fail to generalize. In contrast, Logistic regression result will be more reliable and will generalize well.

Test Data Results: This shows how well logistic regression has generalized well, and every other model value has changed

A screenshot of a computer

AI-generated content may be incorrect.

Cross-validated precision (Logistic Regression) ≈0.55 (folds varied ~0.52–0.59) → marginally above chance. Example confusion matrix (one run):  
TP=37, FP=32, TN=13, FN=18 →  
Accuracy ≈57%, Precision ≈58%, Recall ≈79%, ROC-AUC ≈0.52.  
Interpretation: Model retrieves most helmet users (high recall) but cannot discriminate well (AUC near random) and produces many false positives (moderate precision).



A chart with different colored squares

AI-generated content may be incorrect.

A white text with black text

AI-generated content may be incorrect.

**Interpretation & Diagnostic Insight**

* ROC curve near diagonal → features lack discriminatory power.
* Recall > Precision → model leans toward predicting “helmet” class broadly.
* Low AUC + modest precision indicates contextual/environmental variables alone are insufficient; unobserved behavioral and demographic drivers likely dominate compliance decisions.

**Why Precision Matters**Precision controls false positive rate—important if predicted helmet compliance guides resource allocation or targeted education. Low precision risks misclassifying non-users as compliant, diluting intervention focus.

Justification of Modeling Decisions and Results

Model Choice: Multiple algorithms (Logistic Regression, Decision Tree, Random Forest) were selected to provide a robust comparison and to capture both linear and non-linear relationships in the data. Random Forest was emphasized for its ability to handle complex feature interactions and its strong performance in classification tasks.

Feature Selection: Features were chosen based on exploratory analysis, focusing on variables likely to influence helmet use (e.g., age, injury severity, road type, weather, time of day, law enforcement, traffic, alcohol involvement, speed limit, country). Helmet use itself was excluded from predictors to avoid data leakage.

Metric Selection: Precision and recall were prioritized. Precision is important to avoid falsely classifying non-helmeted riders as helmeted, which is critical for targeted interventions. Recall ensures that most helmeted riders are correctly identified, reducing missed cases.

Validation: Results are validated using cross-validation, confusion matrix, and ROC curve. Cross-validation ensures the model generalizes well to unseen data. The confusion matrix provides insight into true/false positives and negatives, while the ROC curve assesses the model's ability to distinguish helmeted from non-helmeted riders.

Interpretation: High precision and recall indicate reliable predictions. If the model achieves strong scores in both metrics, it can be confidently used for policy or intervention planning. Feature importance from Random Forest can further guide which factors most influence helmet use.

**Key Findings**

* Environmental/contextual features alone have low predictive utility (AUC ~0.52).
* Elevated recall does not compensate for weak separability.
* Helmet behavior likely driven by unmodeled personal, socio-economic, enforcement, and cultural factors.

**Limitations**

* Feature space narrow: lacks demographics, attitudes, enforcement intensity, socioeconomic markers.
* Potential class/noise imbalance not fully addressed with advanced resampling or calibration.
* No temporal or spatial granularity included (e.g., seasonality, location clustering).
* Single dataset size (n=500) constrains model complexity and generalization.

**Future Enhancement Directions  
Data Expansion:**

* Demographics (age distribution integrity, gender, education, income).
* Behavioral (risk perception, prior violations, training, crash history).
* Policy & Enforcement (checkpoint presence, fine levels, enforcement frequency).
* Infrastructure (road type granularity, lighting, surface condition).  
  Modeling Improvements:
* Class weighting + threshold tuning (optimize Fβ for policy preference).
* Regularized logistic regression with interaction terms.
* Gradient boosting (XGBoost/LightGBM) with calibration.
* SHAP-based interpretability to validate feature signal.
* Temporal/spatial models (hierarchical or mixed-effects) if metadata added.  
  Evaluation Enhancements:
* Precision–recall curve analysis for threshold selection.
* Probability calibration (isotonic / Platt).
* External validation on an independent region/time slice.

**Concise Recommendation Summary**

1. Expand feature set before further model tuning.
2. Prioritize calibrated ensemble + threshold optimization over the current baseline.
3. Incorporate socio-behavioral and enforcement data—expected largest marginal gain.
4. Use the current model only as an exploratory benchmark, not operational decision support**.**

**Practical Implications**Findings caution against relying on ambient conditions to infer helmet compliance. Effective interventions should emphasize enforcement intensity, education, affordability, cultural normalization, and addressing access barriers—aligning with literature stressing multifactor approaches over context-only triggers.

**References**Høye, A. (2018) 'Recommend or mandate? A systematic review and meta-analysis of the effects of mandatory bicycle helmet legislation', *Accident Analysis & Prevention*, 120, pp. 239-249.

Liu, B., 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), CD004333.

Servadei, F., Begliomini, C., Gardini, E., Giustini, M., Taggi, F. and Kraus, J. (2003) 'Effect of Italy's motorcycle helmet law on traumatic brain injuries', *Injury Prevention*, 9(3), pp. 257-260.

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