Road Traffic Accident Severity and Risk Analysis

Group Members:

1. Mohammed Irfan Battegeri - mbattege@depaul.edu
2. Vishakha Maheshkumar Kamothi - vkamothi@depaul.edu
3. Sparshika Ajmaan Dinesh Kumar - sajmaand@depaul.edu
4. Sudipth Ajitesh Saravanan - ssarava4@depaul.edu
5. Praveen Kumar Jayakumar - pjayakum@depaul.edu

Course Name: DSC 478 Programming Machine Learning Applications

1. Problem Definition

The objective of this project is to analyze road traffic accident data to:

1. Predict **Accident Severity** on a multi-class scale (minor, moderate, severe) using Random Forest.
2. Predict the likelihood of **Multi-Vehicle Accidents** using an MLP.
3. Predict **Accident Severity** on a binary scale (high vs. low severity) using Logistic Regression.
4. Improve predictions using **Ensemble Modeling**, combining the outputs of different models.

2. Data Preparation and Preprocessing

Dataset

The dataset includes features like:

- **Demographics**: Age band of the driver, sex of the driver.
- **Environmental Conditions**: Weather conditions, road surface type, light conditions.
- **Accident Details**: Accident severity, number of vehicles involved.

Preprocessing Steps

- **Handling Missing Values**:
    - Categorical features: Mode imputation.
    - Numerical features: Median imputation.
- **Feature Encoding**: Label encoding for categorical variables.
- **Feature Engineering**:
    - Polynomial feature interactions for Logistic Regression.
    - Recursive Feature Elimination (RFE) for feature selection.
- **Class Imbalance Handling**:
    - Oversampling using SMOTE and ADASYN.
    - Class weighting for models sensitive to imbalance.
- **Scaling**: StandardScaler applied to numerical features for models requiring normalization.

3. Model Implementations

Model 1: Random Forest for Multi-Class Severity Prediction

Implemented by: Sparshika Ajmaan Dinesh Kumar
Objective: Predict accident severity across three levels: minor, moderate, severe.

Key Steps:

1. **Feature Selection**:
    - Selected features like driver age, vehicle type, and road conditions.
2. **Class Imbalance Handling**:
    - Applied SMOTE to oversample the minority classes.
    - Used class weighting to adjust for imbalances during model training.
3. **Model Training**:
    - Random Forest with 100 estimators and class weights derived from training data.
4. **Results**:

- **Accuracy**: 69.1%
- **Classification Report**:
  - **Severe (Class 2)**: High precision (85%) and recall (77%).
  - **Minor (Class 0)**: Poor recall (8%).
- **Observations**:
  - Struggled with minority classes, likely due to imbalances in the data.

Model 2: Driver Risk Profiling using K-Means Clustering

Implemented by: Vishakha Maheshkumar Kamothi
Objective: To identify driver risk categories based on demographic and behavioral factors using K-Means clustering, enabling actionable risk profiles.

Data Preparation:

- **Selected Features**:
  - Age_band_of_driver, Driving_experience, Type_of_vehicle, Accident_severity, Sex_of_driver, Educational_level, and Vehicle_driver_relation.
- **Feature Engineering**:
  - Categorical encoding using LabelEncoder.
  - Standardization with StandardScaler for equal feature contribution.
  - **Feature Refinement**:
    - **ANOVA F-test** to identify the most relevant features for clustering:
      - Selected Top 5 Features: Accident_severity, Driving_experience, Age_band_of_driver, Sex_of_driver, and Vehicle_driver_relation.
    - **Principal Component Analysis (PCA)**:
      - Applied PCA to analyze feature importance and dimensionality reduction.
      - Explained variance suggested two significant components, capturing ~34% of the total variance.

Clustering Methodology:

- **K-Means Clustering**:
    - Optimal clusters determined using:
        - **Elbow Method**: Suggested 6 clusters based on inertia.
        - **Silhouette Analysis**: Suggested 8 clusters with the highest silhouette score (~0.42).
    - Final clustering performed with **8 clusters** for better profiling.

Cluster Insights:

1. **High-Risk Clusters**:
    - **Cluster 0**: Young, inexperienced drivers with high accident severity.
    - **Cluster 6**: Older drivers with high accident severity.
2. **Low-Risk Clusters**:
    - **Cluster 4**: Older, cautious female drivers with low accident severity.
    - **Cluster 7**: Responsible middle-aged drivers with minimal accident severity.
3. **Moderate-Risk Clusters**:
    - **Cluster 1**: Slightly experienced young drivers.
    - **Cluster 2**: Experienced vehicle owners with severe accident involvement.

Key Findings:

- PCA revealed that a combination of Driving_experience and Accident_severity are the most impactful dimensions in driver profiling.
- Young and inexperienced drivers dominate high-risk clusters.
- Older and cautious drivers represent safer profiles.
- Behavioral and demographic factors significantly influence risk levels.

Model 3: Logistic Regression for Binary Severity Classification

Implemented by: Sudipth Ajitesh Saravanan
Objective: Predict whether an accident is high or low severity.

Key Steps:

1. **Feature Engineering**:
    - Added polynomial interaction terms to capture relationships between road and weather conditions.
    - Selected features like road surface type, lane types, and weather conditions.
2. **Feature Selection**:
    - Used Recursive Feature Elimination (RFE) to retain the top five features.
3. **Class Imbalance Handling**:
    - Applied SMOTE to oversample the minority class.
    - Incorporated balanced class weights into the logistic regression model.
4. **Hyperparameter Tuning**:
    - Conducted RandomizedSearchCV to tune parameters like regularization strength and solver.
5. **Results**:
    - **Accuracy**: 51%
    - **ROC AUC**: 51%
    - **Observations**:
        - Logistic Regression benefited from feature interactions but struggled with non-linear relationships.

Model 4: MLP for Multi-Vehicle Accident Prediction

Implemented by: Mohammed Irfan Battegeri
Objective: Predict whether an accident involves multiple vehicles.

Key Steps:

1. **Feature Selection**:
    - Correlation analysis and Chi-Square tests to identify significant features.
    - Features included weather conditions, light conditions, and number of casualties.

2. **Class Imbalance Handling**:
    - Applied ADASYN oversampling to balance the dataset.
3. **Fine-Tuning**:
    - Experimented with hyperparameter tuning, additional dropout layers, and increased MLP depth.
    - Added early stopping to prevent overfitting.
    - Adjusted classification thresholds (optimal: 0.44) to improve recall for minority classes.
4. **Results**:
    - **Accuracy**: 74.3%
    - **Precision-Recall AUC**: 91.2%
    - **Observations**:

ADASYN significantly improved recall for the minority class (single-vehicle accidents).

4. Ensemble Modeling

Implemented by: Praveen Kumar Jayakumar

Approach 1: Voting Classifier

1. Combined **Random Forest** and **Logistic Regression** predictions.
2. Evaluated both soft and hard voting strategies:
    - **Soft Voting Accuracy**: 83.5%
    - **Hard Voting Accuracy**: 83.1%
3. Observations:
    - Soft voting improved overall accuracy but slightly reduced minority class recall.

Approach 2: Stacking Classifier

1. Combined Random Forest and Logistic Regression using Logistic Regression as the meta-classifier.
2. **Results**:
    - **Accuracy**: 83.8%
    - Observations:

        - Marginal improvement over voting but poor minority class recall.

Approach 3: Weighted Voting with SMOTE

1. Applied SMOTE to balance the training dataset.
2. Assigned weights to Random Forest and Logistic Regression models (e.g., 1:3).
3. **Results**:
    - **Accuracy**: 76.8%
    - **Observations**:
        - Improved minority class recall but slightly reduced overall accuracy.
        -

5. Final Results Summary

Model	Methodology/Techniques	Key Results
Random Forest	Feature selection, SMOTE, class weighting	Accuracy: 69.07%, macro F1: 37%; Insights: Importance of "Time," "Type of Vehicle."
Logistic Regression	Polynomial features, RFE, SMOTE, hyperparameter tuning	Accuracy: 51%, ROC AUC: 51%; Improved precision for binary accident severity classification.
MLP	ADASYN, focal loss, early stopping, dropout layers, classification threshold adjustment	Accuracy: 74.35%, PR AUC: 0.91; Boosted recall for multi-vehicle accidents; Improved minority class F1.
K-Means Clustering	PCA for dimensionality reduction, ANOVA F-test for feature selection, silhouette scores for optimal cluster count	Optimal Clusters: 8; Profiles: Age, experience, and accident severity highlighted key group differences.
Ensemble Modeling	Voting (hard/soft), stacking, weighted voting with SMOTE	Soft Voting: Accuracy: 83.56%, improved F1 for Class 2 (dominant class); Stacking: Accuracy: 83.80%.

6. Lessons Learned and Future Work

- **Class Imbalance**: Minority class recall remained a challenge across all models. Cost-sensitive learning or dynamic sampling could help address this issue.
- **Feature Engineering**: Incorporating additional domain-specific features (e.g., accident location) could improve predictions.
- **Alternative Models**: Advanced ensemble techniques like boosting (XGBoost) or deep learning architectures could better capture complex relationships.
- **Ensemble Refinements**: Experimenting with neural networks as meta-models for stacking might yield better results in future iterations.
- **Feature Dependence:** Clustering results heavily relied on selected features. Incorporating behavioral factors like driving habits could enhance insights.
- **Explained Variance:** PCA explained ~34% variance, suggesting other components might contribute. Future work could explore alternative clustering algorithms for deeper interpretations.