INTRODUCTION

Predicting Road Accident Severity Using Machine Learning

In this project, we aim to predict the severity of road accident casualties using machine learning models. The dataset includes various factors such as accident location, time, weather conditions, road type, and vehicle involvement. Our goal is to build an effective model that can help identify severe casualties, enabling better resource allocation and improved road safety measures.

The steps involved in this project are:

1. **Importing the necessary libraries and dataset.**
2. **Understanding and preprocessing the data.**
3. **Analyzing class imbalance and applying appropriate techniques.**
4. **Building and evaluating different machine learning models, including ensemble methods.**
5. **Identifying the best model and analyzing feature importance for interpretability.**

Through this study, we aim to enhance accident risk assessment and contribute to data-driven road safety improvements.

CONCLUSION

In this project, we developed a machine learning model to predict the severity of road accident casualties based on various features such as accident location, time, road conditions, and vehicle involvement. Given the **highly**

imbalanced dataset, where severe casualties (Class 1) were significantly underrepresented compared to minor ones (Class 0), we applied **class balancing techniques, threshold tuning, and ensemble methods** to improve model performance.

After testing multiple models, including **Decision Trees, Random Forest, SVM, and XGBoost**, the **best-performing model was XGBoost** with the following optimized parameters:

- **colsample_bytree:** 0.8
- **learning_rate:** 0.1
- **max_depth:** 3
- **n_estimators:** 100
- **scale_pos_weight:** 6
- **subsample:** 0.8

By tuning the classification threshold to **0.4910**, we achieved the best trade-off between precision and recall.

Final Model Performance (XGBoost with Threshold Tuning):

- **Accuracy:** 73.78%
- **ROC AUC Score:** 0.7791
- **Recall for Severe Casualties (Class 1):** 70%
- **Precision for Class 1:** 27%
- **F1 Score for Class 1:** 0.39

Challenges and Solutions

1. **Class Imbalance Impact**

a. Initial models struggled to detect Class 1 due to the severe imbalance (**Class 0: 88%, Class 1: 12%**).

b. Without balancing techniques, recall for Class 1 was very low (<10%), meaning most severe casualties were misclassified as minor.

c. **Solution:** We used **class weighting (scale_pos_weight=6)** and **threshold tuning** to significantly improve recall (70%).

2. **Precision vs. Recall Trade-off**

   a. The model achieved high recall for Class 1 (70%) but at the cost of lower precision (27%).

   b. This means the model **successfully identifies most severe casualties but also misclassifies some minor ones as severe.**

   c. Given that recall is more critical than precision in this context (missing a severe casualty is worse than overestimating risk), this trade-off is acceptable.

Key Takeaways

✅ **XGBoost with class balancing and threshold tuning achieved the best recall for Class 1 (70%) while maintaining a stable overall accuracy (73.78%).**
✅ **Class imbalance was a major challenge, but using scale_pos_weight helped significantly.**
✅ **Future improvements** could include testing **anomaly detection methods** or incorporating **more granular accident-related features** to refine predictions.

This study highlights that **carefully tuned machine learning models can effectively handle imbalanced datasets and improve the detection of critical but rare events, such as severe accident casualties.**