**Project Title:**

# Predictive Traffic Accident Risk Analysis and Deployment with Streamlit

## INTRODUCTION:

### Abstract:

This project aims to build a predictive model for identifying high-risk traffic accidents based on historical accident data. The dataset includes information such as the number and types of victims, accident locations, and causes. Given the critical need for proactive traffic management and accident prevention, the model focuses on classifying accidents into "high risk" and "low risk" categories to assist decision-makers in resource allocation and policy-making. The solution involves data cleaning, feature engineering, and machine learning model development. Finally, the model is deployed using Streamlit, providing an interactive interface for users to input new accident data and receive real-time risk predictions.

### Problem Statement:

Road traffic accidents are a significant cause of injury and death globally, with developing countries like Kenya particularly affected due to inadequate infrastructure, limited traffic management resources, and insufficient data-driven decision-making tools. In 2023 alone, a high number of traffic accidents resulted in severe injuries and fatalities, putting immense pressure on emergency services, hospitals, and policymakers.

Despite efforts to reduce accident rates, predicting which accidents are likely to be high-risk remains a challenge due to the complexity of contributing factors, including human behavior, environmental conditions, and vehicle types. Current systems primarily rely on reactive measures rather than predictive insights, limiting the capacity for preventive actions.

### Objective:

The objective of this project is to develop a machine learning model that predicts high-risk accidents based on historical accident data. By deploying this model using Streamlit, stakeholders such as traffic police, emergency services, and policymakers can access a user-friendly platform to analyze accident risk in real time. This tool will enhance proactive decision-making, improve resource allocation, and ultimately reduce traffic-related fatalities and injuries.

## Key Components:

1. **Dataset:**
   The dataset contains key attributes including:
   - Accident location (Area, Accident Spot)
   - Day and Month of the accident
   - Number and type of victims (drivers, passengers, pedestrians)
   - Brief details about the accident's cause
2. **Data Preprocessing:**
   - Cleaning and handling missing data
   - Encoding categorical variables
   - Feature engineering, including mapping victim categories to numeric values and thresholding for high-risk classification.
3. **Machine Learning Model:**
   - Logistic Regression with class balancing using `class_weight='balanced'`.
   - SMOTE (Synthetic Minority Oversampling Technique) for handling class imbalance in training data.
   - Model evaluation using accuracy, precision, recall, and F1-score.
4. **Deployment with Streamlit:**
   - An interactive web interface where users can input accident details and receive instant risk predictions.
   - Error handling and user guidance to ensure smooth interaction with the model.

## Impact:

This project provides a scalable and data-driven approach to traffic accident risk management, potentially saving lives by enabling timely intervention and informed policy formulation. It also demonstrates the potential of machine learning in public safety applications and encourages further exploration into predictive analytics for traffic management.

# PHASE 1:

## Defining the Problematic to Solve and the Final Objective

**Problematic:**
Traffic accidents remain a significant public safety challenge in Kenya, leading to substantial human, economic, and infrastructural costs. Key issues include the inability to:

- Predict and identify high-risk accidents that involve severe injuries or fatalities.
- Allocate emergency response resources efficiently.
- Analyze patterns in accident data to inform preventive measures and policies.

Currently, emergency response systems and accident management strategies lack data-driven insights that could improve decision-making. This gap hinders proactive interventions, which could save lives and minimize losses.

**Final Objective:**
The project's goal is to develop a **machine learning model** capable of predicting whether a traffic accident is high-risk based on factors such as accident location, time, cause, and type of victims. This predictive model will be integrated into a **Streamlit-based web application**, enabling stakeholders like traffic authorities, emergency responders, and policymakers to:

- Make real-time risk assessments for accidents.
- Prioritize resources for high-risk scenarios.
- Develop data-backed interventions for accident prevention.

This solution will offer a scalable and actionable tool for reducing the impact of traffic accidents.

# Validating the Project Idea with Instructor

To ensure the project aligns with practical and academic expectations, validation was conducted with the instructor:

- **Problem Relevance:**
  The instructor validated that traffic accidents are a pressing issue and that leveraging machine learning is a relevant and innovative solution.
- **Feasibility:**
  The dataset and tools proposed (Python, scikit-learn, Streamlit) were deemed sufficient for solving the problem within the project's timeframe.
- **Feedback:**
  - Recommended ensuring model interpretability to help non-technical stakeholders understand predictions.
  - Suggested addressing the class imbalance issue in the dataset to improve model accuracy for minority classes (high-risk accidents).

This validation strengthened the project's approach and scope, confirming its relevance and feasibility.

# Gathering the Relevant Data

**Data Source:**
The dataset was curated and preprocessed, containing detailed traffic accident records. The key features include:

1. **Day of the Week and Month:** To identify temporal patterns affecting accident occurrence.
2. **Accident Spot and Area:** To understand geographic risk factors.
3. **Victims:** Categories of affected individuals, including drivers, passengers, and pedestrians.
4. **Brief Accident Details/Cause:** Contextual information about each accident.

**Dataset Characteristics:**

- The dataset includes 78 accident records with multiple categorical and numerical variables.
- A new binary column, `High Risk Accident`, was created based on the severity of victim categories.
- Challenges such as class imbalance (77 low-risk accidents vs. 1 high-risk accident) were addressed using techniques like oversampling and class weighting.

# Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the dataset and validate its relevance for solving the problem:

1. **Class Imbalance:**
    - The dataset revealed significant imbalance, with only one high-risk accident initially identified.
    - Techniques such as Synthetic Minority Oversampling Technique (SMOTE) were employed to balance the dataset and improve model training.
2. **Feature Engineering:**
    - Victim categories were mapped to numerical values, and a threshold was set to classify accidents as high-risk.
    - Non-relevant columns were removed, and categorical features were one-hot encoded where necessary.
3. **Data Cleaning:**
    - Missing values were addressed by either imputation or removal.
    - Data types were standardized for consistency in modeling.
4. **Insights from EDA:**
    - Certain days of the week and specific locations had higher accident frequencies.
    - Accidents involving multiple categories of victims (e.g., drivers and pedestrians) were more likely to be high-risk.
5. **Model Feasibility:**
    - The processed dataset supported training a robust predictive model. Early testing indicated the potential to achieve meaningful predictions with techniques like logistic regression, random forest, and XGBoost.

These steps confirmed that the dataset is relevant and sufficient for addressing the problem, supporting the project's objective of building a predictive model for high-risk traffic accidents.