**# Auto Crash Fatality Prediction - Final Report**

**1. Introduction**

Traffic accidents remain a critical concern, with fatalities being a major public safety issue. The goal of this project is to develop a **machine learning model** that predicts the likelihood of fatalities in road accidents based on key factors such as vehicle count, pedestrian involvement, and road conditions.

**2. Dataset Overview**

The dataset used in this project contains data on **fatal traffic accidents in Arizona from 2012 to 2016**. It includes various factors such as location, time, weather conditions, and accident details.

**Key Features in the Dataset:**

* **VE\_FORMS**: Number of vehicles involved.
* **PERSONS**: Number of people involved.
* **ROAD\_FNC**: Type of road where the accident occurred.
* **MAN\_COLL**: Type of collision.
* **A\_PED, A\_PED\_F**: Pedestrian involvement.
* **A\_DROWSY**: Indicator of driver drowsiness.
* **INDIAN\_RES**: Whether the accident occurred on tribal land.

**3. Data Preprocessing & Exploration**

**Handling Missing Values:**

Some columns such as FUNC\_SYS, ROAD\_FNC, RD\_OWNER, and cf1\_lit had missing values. These were handled using appropriate imputation techniques.

**Feature Selection:**

A correlation heatmap was used to identify the **top relevant features** influencing fatalities. The most important features were selected based on correlation with the target variable FATALS.

**Feature Importance Analysis:**

* The number of **persons involved (PERSONS)** had the **highest impact** on fatalities.
* **Road function (ROAD\_FNC)** and **type of collision (MAN\_COLL)** were also significant predictors.
* Other relevant features included **VE\_FORMS, VE\_TOTAL, SP\_JUR, and A\_DROWSY**.

**4. Model Selection & Training**

We trained and compared two different machine learning models:

**Model 1: Logistic Regression**

* **Accuracy:** **94.35%**
* Pros: Works well for simple classification problems, interpretable.
* Cons: May not capture complex patterns in the data.

**Model 2: Random Forest (Tuned)**

* **Accuracy:** **82.80%**
* Pros: Captures complex relationships, handles missing data well.
* Cons: Slightly lower accuracy due to class imbalance.

**Model Comparison:**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | **94.35%** |
| Random Forest | **82.80%** |

**5. Conclusion & Key Insights**

* The **Logistic Regression model** achieved the best accuracy at **94.35%**, making it the recommended model for prediction.
* The **most influential factor** in predicting fatalities was the **number of persons involved in the accident**.
* **Road function and type of collision** also played significant roles in determining accident severity.
* The **Random Forest model, though slightly less accurate, provided insights into feature importance**, which could be useful for accident prevention strategies.

**6. Next Steps & Recommendations**

* **Balance the dataset** using **SMOTE (Synthetic Minority Oversampling Technique)** to improve the Random Forest model.
* **Experiment with Gradient Boosting models** (e.g., XGBoost, LightGBM) to enhance accuracy.
* **Deploy the model** as an API to predict accident severity based on input features.
* **Share insights with policymakers** to improve road safety measures.