US Accidents Analysis & Severity Prediction

By Sahil Karande

Objective

This project explores the **US Accident Dataset** to uncover key insights about traffic accident trends across different cities and times. It also involves preprocessing and balancing techniques to improve the performance of severity prediction models, addressing class imbalance issues.

Dataset Overview

The dataset contains **~500,000 accident records** from across the United States, with rich features including:

- Location (City, State, Zip)
- Time-based features (Start Time, End Time)
- Environmental factors (Temperature, Weather Condition, Visibility, etc.)
- Infrastructure indicators (Traffic Signal, Amenity, Crossing, etc.)
- Target: Severity (1 to 4)

Data Preprocessing

- **Missing Values**: High-NA columns like End_Lat, End_Lng, and weather metrics like Wind_Chill(F) and Precipitation(in) were removed.
- **Datetime Parsing**: Extracted Hour, Day, Month, and Weekday from Start_Time.
- **Feature Reduction**: Redundant or uninformative features like Turning_Loop, Airport_Code, etc., were dropped.
- **Label Encoding**: Applied to categorical variables for model compatibility.

Exploratory Data Analysis (EDA)

Severity Distribution

- **Class Imbalance**: Majority of accidents are of **Severity 2**, creating a biased model risk.
- **Oversampling** was suggested to fix this and improve model fairness.

Top Accident-Prone Cities

- **Houston** and **Miami** top the list with over **10,000** accidents.
- **Los Angeles, Dallas, Charlotte** also show high accident frequency.

- Tourist and tech hubs like **Austin and Orlando** follow closely.

Peak Accident Hours

- Morning peak: **7 AM & 8 AM**
- Evening peak: **4 PM – 5 PM**
- Lowest activity: **1 AM – 3 AM**

Day-of-Week Insights

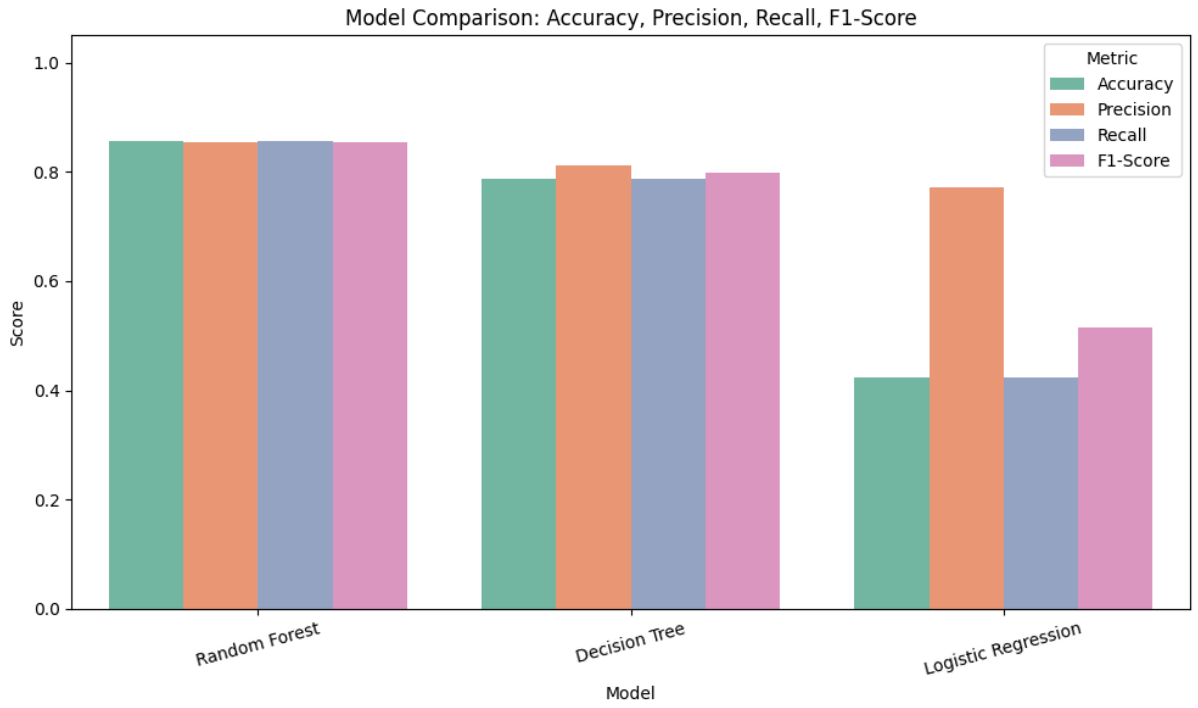
- **Friday** records the most accidents (~80K+).
- Accidents are concentrated on **weekdays**, with weekends having much fewer incidents.

Weather Conditions during Accidents

- Most accidents occur in **Fair** or **Clear** weather — not during bad weather!
- Light Rain, Fog, and Snow contribute less than expected.

Machine Learning Modeling Results

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.855887	0.855181	0.855887	0.855185
Decision Tree	0.788323	0.812495	0.788323	0.798874
Logistic Regression	0.422946	0.772277	0.422946	0.514188



Key Insights Summary

- Urban density and peak commute hours drive accident frequency.
- Surprisingly, **good weather** is when most accidents happen — indicating human error, congestion, or road design as root causes.
- **Temporal patterns** show the need for targeted interventions during weekday rush hours.

Conclusion

The US Accident dataset provides deep insights into the **who, when, and where** of traffic accidents. Through data cleaning, visualization, and strategic rebalancing, this project lays the foundation for building fair and effective accident severity prediction models.