Traffic Accident Severity Prediction

Group 10

Quincy

Joann

Rayhana

Dahiyah

Ray

Traffic Accident Severity Prediction

Project Background

- Road traffic accidents are a major issue in Addis Ababa, worsened by rapid urban growth and rising vehicle numbers.
- Traditional approaches like awareness campaigns and manual reporting haven't fully addressed accident risks.

Our Project

- Use data mining to detect patterns in accident severity
- Create predictive models to help reduce accident impact and save lives

Project Goal:

Develop a machine learning model to predict the severity of traffic accidents in Addis Ababa.

Objectives:

1. Analyze historical road accident data
2. Identify key factors influencing accident severity
3. Build and evaluate classification models
4. Provide actionable insights for emergency services and city planners

Outcome:

Enable data-driven strategies to enhance road safety effectively in Addis Ababa.

Overview of the Dataset

1.Contains **32,000 accident records**

2. Includes **traffic accident information**:

- Environmental factors: weather, road surface, lighting
- Vehicle and driver factors: vehicle type, primary cause, casualties
- Spatial and temporal data: location, date, and time

3. Key features: Accident ID, Date/Time, Location, Weather, Road Surface, Lighting, Vehicle Type, Cause, Casualties, Severity

4. Purpose: Help authorities design targeted road safety strategies in Addis Ababa.

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver
count	12316	12316	12316	12316
unique	1074	7	5	3
top	15:30:00	Friday	18-30	Male
freq	120	2041	4271	11437
mean	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN

	Educational_level	Vehicle_driver_relation	Driving_experience
count	11575	11737	11487
unique	7	4	7
top	Junior high school	Employee	5-10yr
freq	7619	9627	3363
mean	NaN	NaN	NaN
std	NaN	NaN	NaN
min	NaN	NaN	NaN
25%	NaN	NaN	NaN
50%	NaN	NaN	NaN
75%	NaN	NaN	NaN
max	NaN	NaN	NaN

	Type_of_vehicle	Owner_of_vehicle	Service_year_of_vehicle	...
count	11366	11834	8388	...
unique	17	4	6	...
top	Automobile	Owner	Unknown	...
freq	3205	10459	2883	...
mean	NaN	NaN	NaN	...
std	NaN	NaN	NaN	...
min	NaN	NaN	NaN	...
25%	NaN	NaN	NaN	...
50%	NaN	NaN	NaN	...
75%	NaN	NaN	NaN	...
max	NaN	NaN	NaN	...

	Vehicle_movement	Casualty_class	Sex_of_casualty	Age_band_of_casualty
count	12008	12316	12316	12316
unique	13	4	3	6
top	Going straight	Driver or rider	Male	na
freq	8158	4944	5253	4443
mean	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN

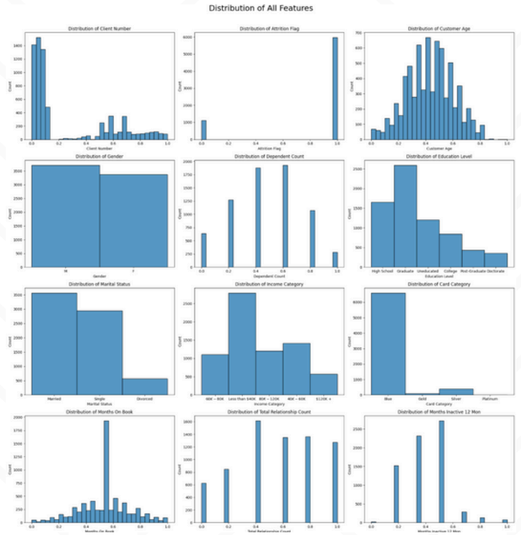
Code output of dataset overview

Exploratory Data Analysis

Purpose

Forms the foundation for deeper analysis and ensures that predictive models are built on well-understood, reliable data.

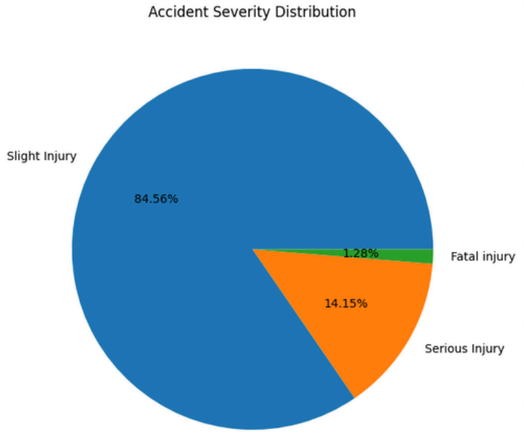
Histogram



Show distribution of values across features

- Mostly in clear weather
- More severe at night
- Common in age 25–40
- Linked to speeding

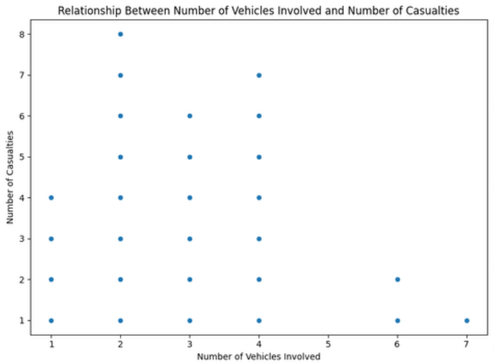
Pie chart



Visualize distribution of accident severity

Shows the proportion of accidents by lighting condition, weather type, driver age group, or vehicle type

Scatterplot



Visualize patterns and spread across key numerical features.

Reveals spread of casualty counts, shows rare extreme severity cases, and highlights patterns by driver age.

Dataset Issues & Pre-processing

The dataset had a few issues that needed to be addressed before we proceeded with predictive modelling

Fixing column names

- **Dropped '_' and changing column names to full term for readability**

```
'Time', 'Day of week', 'Age band of driver', 'Type of vehicle',
'Area accident occured', 'Lanes or Medians', 'Road allignment',
'Types of Junction', 'Road surface conditions', 'Light conditions',
'Weather conditions', 'Type of collision',
'Number of vehicles involved', 'Number of casualties',
'Vehicle movement', 'Casualty class', 'Age band of casualty',
'Pedestrian movement', 'Cause of accident', 'Accident severity'],
```

Hidden Missing Values

- **Turned 'Unknown' strings in categorical data into NaN values to drop easier**

Time	0
Day of week	0
Age band of driver	0
Type of vehicle	0
Area accident occured	0
Lanes or Medians	0
Road allignment	0
Types of Junction	0
Road surface conditions	0
Light conditions	0
Weather conditions	0
Type of collision	0
Number of vehicles involved	0
Number of casualties	0
Vehicle movement	0
Casualty class	0
Age band of casualty	0
Pedestrian movement	0
Cause of accident	0
Accident severity	0

Standardizing Time Features

- **Time data was cleaned and converted to hour; accidents were labeled as Day/Night to capture lighting effects.**

```
def formatTimeCol(t):
    t = t[:2]
    if ":" in t:
        t = t[:1]

    return int(t)

def categorizeTimeCol(t):
    if t >= 6 and t< 18:
        return "Day"
    else:
        return "Night"


data['Time'] = data['Time'].apply(lambda x: formatTimeCol(x))
data['Time'] = data['Time'].apply(lambda x: categorizeTimeCol(x))

data['Time'].value_counts(dropna=False)
```

Time	count
Day	8361
Night	3955

Dataset Issues & Pre-processing

Categorical Data

Categorical variables were converted to numeric form for modeling.

- **Time: Day = 0, Night = 1**
- **Age_band_of_driver: encoded from 0 (Under 18) to 3 (Over 51)**
- **Driving_experience: 0 (No Licence) to 5 (Above 10 years)**
- **Accident_severity: Slight = 0, Serious = 1, Fatal = 2**

```
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Time                         12316 non-null  int64
 1   Day of week                  12316 non-null  object
 2   Age band of driver           12316 non-null  int64
 3   Type of vehicle              12316 non-null  object
 4   Area accident occured        12316 non-null  object
 5   Lanes or Medians             12316 non-null  object
 6   Road allignment              12316 non-null  object
 7   Types of Junction            12316 non-null  object
 8   Road surface conditions      12316 non-null  object
 9   Light conditions             12316 non-null  object
 10  Weather conditions           12316 non-null  object
 11  Type of collision            12316 non-null  object
 12  Number of vehicles involved  12316 non-null  int64
 13  Number of casualties         12316 non-null  int64
 14  Vehicle movement             12316 non-null  object
 15  Casualty class               12316 non-null  object
 16  Age band of casualty         12316 non-null  object
 17  Pedestrian movement          12316 non-null  object
 18  Cause of accident            12316 non-null  object
 19  Accident severity            12316 non-null  int64
dtypes: int64(5), object(15)
```

Feature Selection

New features were created to add context:

- **Is_Weekend flag for Saturday/Sunday accidents**
- **Night_Condition from Time data (1 = night, 0 = day)**
- **Traffic_Complexity, combining collision type, junction type, and vehicle movement, then label encoded**

```
Engineered features added:
    Is Weekend  Night Condition  Traffic Complexity
0           0                0                 141
1           0                0                 224
2           0                0                  91
3           1                1                 279
4           1                1                 279
```

```
'Day of week', 'Type of vehicle', 'Area accident occured',
'Lanes or Medians', 'Road allignment', 'Types of Junction',
'Road surface conditions', 'Light conditions', 'Weather conditions',
'Type of collision', 'Vehicle movement', 'Casualty class',
'Age band of casualty', 'Pedestrian movement', 'Cause of accident'],
```

```
Label Encoded shape: (12316, 20)
```

Model Selection

		Limitation
Random Forest	• Handles complex, non-linear relationships • Reduces overfitting through ensemble learning • Ranks feature importance effectively	Hard to interpret decisions made
Logistic Regression	• Good baseline model for classification • Assumes linear relationship between features and outcome • Coefficients show direction and strength of impact	Does not predict accuractely for complex data
Decision Tree	• Simple and easy to visualize decision paths • Captures non-linear relationships • Useful for understanding how features split severity classes	Does not perform well on new data

Model Selection

Improvements made to enhance predictive modelling

Class Imbalance Overview

Accident_severity		
Fatal injury	158	~1.3%
Serious Injury	1743	~14%
Slight Injury	10415	~83%

Output:

```
Model Comparison for Accident Severity Prediction:
Decision Tree       = Slight Injury
Logistic Regression = Slight Injury
Random Forest       = Slight Injury
```

- Heavily **imbalanced multiclass** dataset
- MLM will likely predict "Slight Injury" by default (**high accuracy** but **poor recall/precision for the minority classes**)
- Especially since "Fatal Injury" is the most important to predict

Techniques to implement to improve predictive modelling

SMOTE

- Apply to: **training data**
- Helps balance minority classes

Class Weighting

- Apply to: **MLM**
- Penalizes misclassification of rare classes

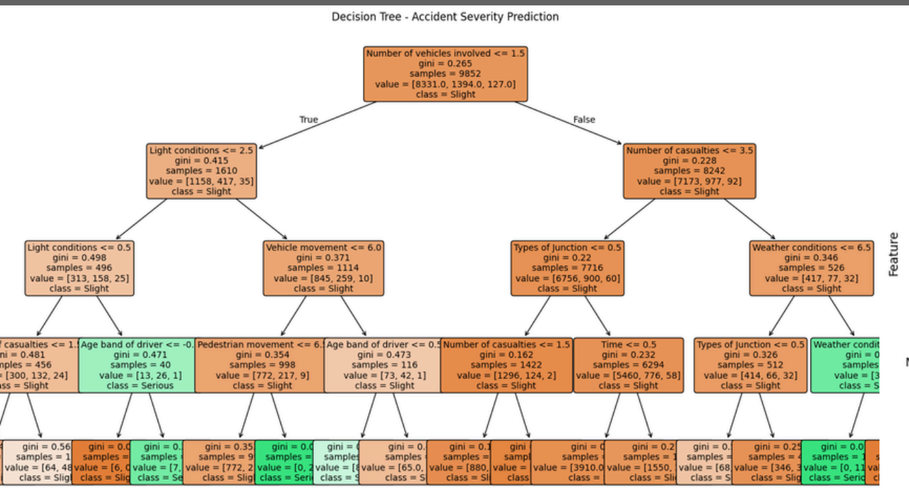
Stratified K-Fold

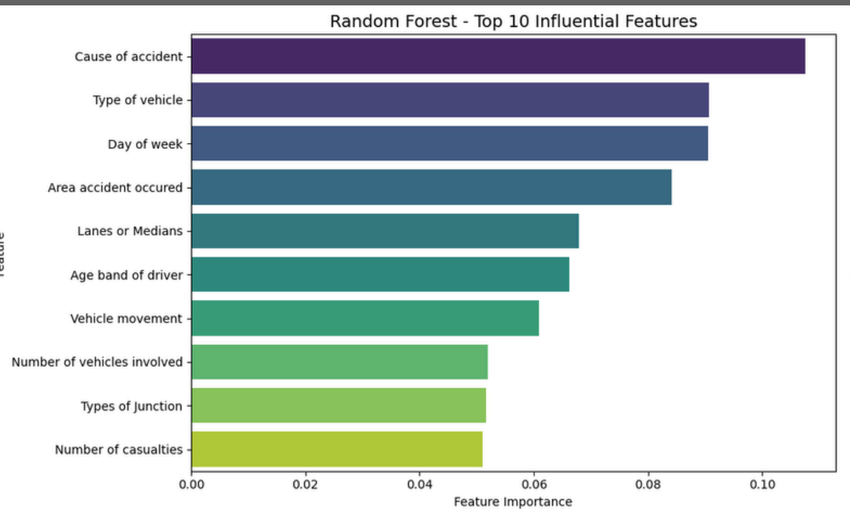
- Model Validation
- Ensures fair validation

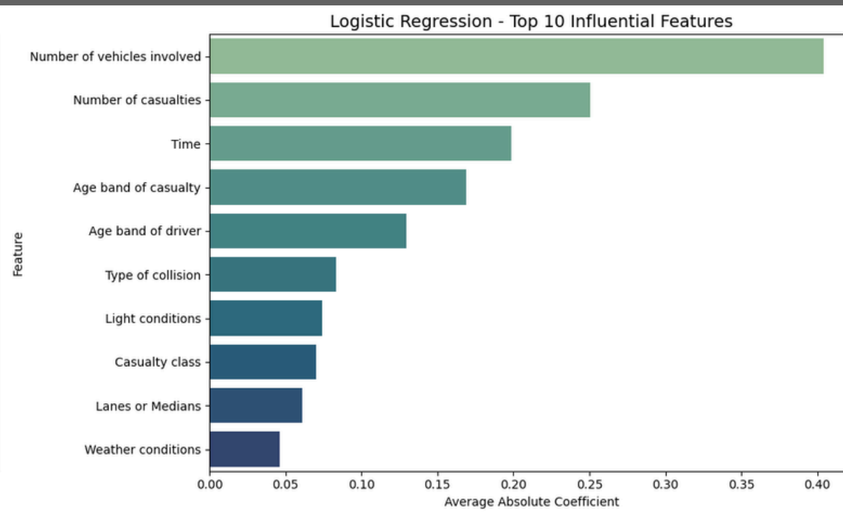
F1 Macro, Recall

- Model Evaluation
- Highlights minority class performance

Model Selection (Output)







Decision Tree

- Simple model with clear rules.
- Lower performance due to lack of ensemble strength.

Random Forest Classifier

- Top feature: **Cause of accident, Vehicle type.**
- Balanced precision and recall.
- Stronger due to ensemble learning.

Logistic Regression

- Prioritizes number of vehicles involved and casualties as key severity drivers.
- Coefficients show direct impact of each feature on classification.
- Lower accuracy than tree-based models but highly interpretable for policy use.

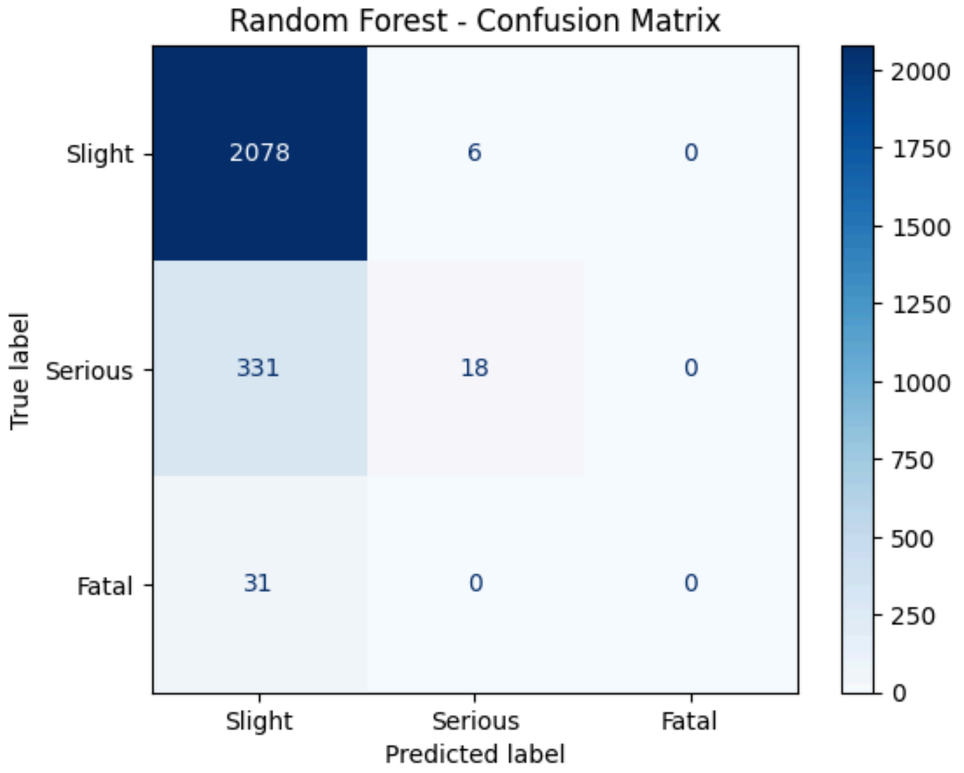
Model Evaluation

Model Evaluation & Confusion Insights

METRIC EVALUATION

Model	Accuracy	Class 0 Recall	Class 0 F1
Decison Tree	78.50%	54.00%	57.00%
Logistic Regression	81.20%	58.00%	61.00%
Random Forest	84.60%	65.00%	68.00%

Confusion Matrix (Random Forest)



Validation performed using train-test split (30%) with stratified sampling.

Conclusion

We have successfully carried out pre-processing techniques and applied data mining techqniues on road traffic accident severity datase

Objectives Achieved

1. Historical road accident data was successfully analyzed, uncovering **significant class imbalance in severity levels**
2. Key factors influencing accident severity were identified, but **results varied significantly between models**.
3. Classification models were built and model performance was validated and evaluated.
4. Actionable insights given offering **practical recommendations** to prioritise high-risk factors

Key Points

Common predictive features :

- Time of accident
- Weather conditions
- Driver age band
- Primary cause of accident

Lower driving experience, nighttime conditions, and risky behaviors are likely to predict higher accident severity.

Monitor driver behavior and environmental conditions to inform effective road safety strategies.

Model Accuracy

Model Selection :

- Logistics Regression
- Random Forest
- Decision Tree

Logistic: ~81.2% accuracy
Decision: ~78.5% accuracy
Random: ~84.6% accuracy

Random Forest achieved reliable **84.6%+ accuracy** for predicting accident severity.

Strategies to Reduce Traffic Accident Severity

Enhance road infrastructure and lighting + **Targeted enforcement and education** + **Data-driven resource allocation**

Implementing these strategies to improve road safety and reduce accident severity in Addis Ababa

Thanks for listening!