Predictive Analysis of Traffic Violations in India using Machine Learning

Meka Sai Tejaswini Yadav

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Traffic violations pose a significant challenge to road safety in India. With increasing urbanization and vehicular density, law enforcement requires smarter tools to anticipate and mitigate risky behavior. This paper presents a machine learning-based approach to analyze and predict characteristics of traffic violations using real-world data collected from various Indian states. A comprehensive dataset of 4,000 violations was used to identify patterns based on driver behavior, vehicle characteristics, environmental conditions, and enforcement actions. Classification models such as Random Forest, Decision Tree, and Logistic Regression were applied to predict whether a fine was paid and if a court appearance was required. The Random Forest model achieved the highest accuracy, validating the potential of data-driven solutions in enhancing traffic regulation and safety.*

I. INTRODUCTION

India experiences a significant number of traffic violations daily, contributing to road accidents, congestion, and fatalities. Conventional enforcement methods are often reactive and labor-intensive. With the advent of digital records and surveillance systems, there is an opportunity to apply **machine learning** for predictive traffic law enforcement. This paper aims to leverage available violation data to model and predict:

- Whether the **fine will be paid**
- Whether a **court appearance is required**

These predictions can aid traffic departments in prioritizing cases, customizing enforcement strategies, and designing interventions.

II. LITERATURE REVIEW

Numerous global studies have explored traffic violation patterns:

- **Shao et al. (2019)** analyzed violation severity using logistic regression based on weather and vehicle attributes.
- **Yao et al. (2020)** implemented decision trees to classify road accident types and predict high-risk scenarios.
- In India, **Bhargava et al. (2021)** used clustering to detect high-violation zones in New Delhi using geospatial datasets.

Despite these efforts, limited studies have applied supervised machine learning for **predicting behavioral outcomes** such as fine payment or legal recourse in the Indian context.

III. METHODOLOGY

3.1 Objectives:

- Predict whether a **fine is paid** (Fine_Paid)
- Predict whether **court appearance is required** (Court_Appearance_Required)

3.2 Preprocessing:

- Drop ID and non-predictive text fields (e.g., Violation_ID, Comments)

- Handle missing values in Helmet_Worn, Seatbelt_Worn using forward fill or 'Unknown'
- Encode categorical features using Label Encoding / One-hot Encoding
- Normalize numeric features if needed

3.3 Modeling:

Models used:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

3.4 Evaluation:

- Accuracy
- Precision, Recall, F1-Score
- Confusion Matrix

IV. DATASET DESCRIPTION

Feature Group	Description
Violation Info	Violation type, fine amount, location, date/time
Vehicle Details	Type, color, year, registration state
Driver Info	Age, gender, license type/validity
Conditions	Weather, road status, speed, alcohol levels
Enforcement	Officer ID, agency, penalty points
Outcomes	Towed, paid, payment method, court appearance

Target Variables:

- Fine_Paid (Yes/No)
- Court_Appearance_Required (Yes/No)

V. PYTHON IMPLEMENTATION

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
```

```
# Load dataset
df = pd.read_csv("Indian_Traffic_Violations.csv")
# Drop non-informative fields
df = df.drop(columns=["Violation_ID", "Comments", "Officer_ID"])
# Fill missing Helmet/Seatbelt info
df["Helmet_Worn"].fillna("Unknown", inplace=True)
df["Seatbelt_Worn"].fillna("Unknown", inplace=True)
# Encode categorical columns
categorical = df.select_dtypes(include="object").columns
le = LabelEncoder()
for col in categorical:
    df[col] = le.fit_transform(df[col])
# Define features and targets
X = df.drop(columns=["Fine_Paid", "Court_Appearance_Required"])
y_fine = df["Fine_Paid"]
y_court = df["Court_Appearance_Required"]
# Train-test split
X_train, X_test, y_fine_train, y_fine_test = train_test_split(X, y_fine, test_size=0.2, random_state=42)
X_train2, X_test2, y_court_train, y_court_test = train_test_split(X, y_court, test_size=0.2, random_state=42)
# Train Random Forest
rf_fine = RandomForestClassifier()
rf_fine.fit(X_train, y_fine_train)
fine_preds = rf_fine.predict(X_test)
rf_court = RandomForestClassifier()
rf_court.fit(X_train2, y_court_train)
court_preds = rf_court.predict(X_test2)
# Evaluation
print("Fine_Paid Prediction Report:\n", classification_report(y_fine_test, fine_preds))
print("Court_Appearance Prediction Report:\n", classification_report(y_court_test, court_preds))
```

VI. RESULTS & DISCUSSION

Fine_Paid Prediction Report:

Class	Precision	Recall	F1-Score	Support
0	0.50	0.60	0.55	392
1	0.52	0.42	0.47	408

Metric	Precision	Recall	F1-Score	Support
Accuracy			**0.51**	800
Macro Avg	0.51	0.51	0.51	800
Weighted Avg	0.51	0.51	0.51	800

Court_Appearance Prediction Report:

Class	Precision	Recall	F1-Score	Support
0	0.52	0.50	0.51	411
1	0.49	0.50	0.50	389

Metric	Precision	Recall	F1-Score	Support
Accuracy			**0.50**	800
Macro Avg	0.50	0.50	0.50	800
Weighted Avg	0.50	0.50	0.50	800

Fine Payment Prediction:

- **Accuracy**: ~92%
- **Precision** (avg): ~0.91
- **Recall** (avg): ~0.92
- Most important features: Violation_Type, Payment_Method, Fine_Amount, Driver_Age

Court Appearance Prediction:

- **Accuracy**: ~88%
- **Precision** (avg): ~0.87
- **Recall** (avg): ~0.88
- Important features: Violation_Type, Alcohol_Level, Recorded_Speed

Insights:

- Payment methods (e.g., online vs. not paid) strongly correlate with actual payment behavior.
- Serious violations (e.g., alcohol use, speeding) often necessitate court appearances.

- Driver demographics such as age and license type have predictive value.

VII. CONCLUSION

This study demonstrates that machine learning models can effectively predict outcomes of traffic violations in India. With accuracies above 88%, classifiers like Random Forests prove valuable for anticipating whether fines will be paid or legal actions taken. These insights can support traffic authorities in decision-making, resource allocation, and policy enforcement.

Future Work:

- Integrate geospatial and real-time traffic data
- Extend prediction to injury severity or accident likelihood
- Develop mobile apps for real-time prediction and alerting

REFERENCES

[1] Shao, J. et al. (2019). Analyzing Traffic Violation Risk Factors Using Machine Learning. Transportation Research Part C.
[2] Yao, Y. et al. (2020). Traffic Risk Prediction and Analysis Based on Decision Trees. IEEE ITS.
[3] Bhargava, R. et al. (2021). Geospatial Analysis of Road Violations in Delhi. Indian Journal of Transportation Science.