PREDICTING TRAFFIC ACCIDENT OUTCOMES: A LOGISTIC REGRESSION APPROACH FOR INJURY PREDICTION

GitHub Repository Link: https://github.com/akhil9390/Predicting-Traffic-Accidents-Outcomes

Table of Contents

1. Introduction

Overview of the Dataset

For this study, a dataset is used that contains information about traffic accidents, features about the circumstances and outcomes of each accident. The driver's age, driver experience, weather conditions, road type, accident severity and traffic density are some of the key features in the dataset. The dataset's primary target variable is an accident having injured or not (injury). The problem of predicting an accident outcome is a binary classification problem, as this is highly related to predicting different aspects of a traffic accident like understanding patterns, as well as predicting injury severity (Jamal *et al.*, 2021).

Aim

This report aims to develop a Logistic Regression model for predicting the possibility of an injury in a traffic accident with respect to different parameters like weather conditions, road type, driver characteristics, etc. The logistic regression algorithm is used for the purpose of estimating the probability of an accident causing injury so that decision-makers can concentrate on those factors that play a role in causing the severity of accidents.

Significance of the Study

It is crucial for the traffic safety management improvement. The model predicts injury outcomes and allows identifying high-risk factors so that injuries and fatalities can be reduced with appropriate interventions. The findings may help develop traffic safety policies, design improvements on the road, and driver education programs.

2. Methodology

2.1 Data Cleaning

Proper data cleaning was required to prevent any inaccurate and disorganized dataset.

- **Missing Value Handling:** The SimpleImputer 'most frequent' strategy is used to fill most frequent category (e.g., "Weather, Road_Type) and numerical features (e.g, Driver_Age) and left all missing value as np.nan.

- **Removing Duplicates:** The repetitive data was brought down to just one row per type.
- **Outlier Detection:** Potential outliers in features such as Driver_Age and Traffic_Density were used to detect outliers in features like these using box plots. There were no significant outliers to affect the analysis.

2.2 Exploratory Data Analysis (EDA)

EDA brought relationships and patterns in the dataset to light.

- **Univariate Analysis:** using bar plot to show particular variables such as Weather and Accident_Severity. The 'no injury' and 'injury' were extremely imbalanced in an 'Accident_Severity' variable.
- **Bivariate Plots:** The features and the target variable were related using correlation matrices and scatter plots. Strong correlations were found between features, such as Traffic_Density and the likelihood of accidents.
- **Missing Data:** After imputation there was no remaining significant missing data, the missing data was included in a heatmap and confirmed no missing data.
- **Feature Correlation:** The correlation matrix informed of the relationship of Driver_Age to Driver_Experience and so featured selection was chosen to avoid multicollinearity.

2.3 Data Preprocessing

This data was then transformed in order to be compatible with the logistic regression model.

- **Encoding Categorical Variables:** Weather and Road_Type were encoded using Label Encoding, making categorical variable numeric values for easier model compatibility.
- **Feature Scaling:** StandardScaler was used for standardizing numerical features like Driver_Age and Speed_Limit so that all have a mean 0 and a standard deviation of 1.
- **Feature Selection:** Recursive Feature Elimination (RFE) was used to pick the best 10 features and retain those like Driver_Age, Traffic_Density, and Weather which were the factors which best predicted accident outcomes.

2.4 Model Implementation

Accident Outcomes were predicted by the implementation of the logistic regression model.

- **Model Selection:** For this binary classification problem, Logistic regression was chosen as the model to predict injury occurrence (1) or no injury (0). It is perfectly suitable for a linear relationship and is used to represent probabilities (Song *et al.*, 2021).
- **Splitting the Data:** The dataset was split into 80% training set and 20% test set so that the model train on most of the data and evaluate some of them on the test set.
- **Model Training:** The logistic regression model was trained using the LogisticRegression class in scikit-learn (Saran and Nar, 2025). It used liblinear solver with optimal performance and class_weight 'balanced' handling of class imbalance (Kumar, 2022). Injury occurrence was the target variable and the features were scaled and later used for training the model.
- **Model Evaluation:** The accuracy, precision, recall, F1-score and confusion matrix are used to evaluate the performance of the model in predicting the severity of accidents.

3. Results and Discussion

3.1 Data Cleaning

First few rows of cleaned data:

	Weather	Road_Type	Time_of_Day	Traffic_Density	Speed_Limit	\
0	2	0	2	-0.014651	0.943602	
1	0	3	3	-0.014651	1.583087	
2	2	1	1	-0.014651	-0.335369	
3	0	0	0	1.284374	-0.335369	
4	2	1	2	-0.014651	3.981158	

	Number_of_Vehicles	Driver_Alcohol	Accident_Severity	Road_Condition	\
0	0.874829	0.0	1	3	
1	-0.142088	0.0	2	3	
2	0.366370	0.0	1	1	
3	-0.142088	0.0	1	2	
4	3.925579	0.0	1	0	

	Vehicle_Type	Driver_Age	Driver_Experience	Road_Light_Condition	Accident
0	1	0.499691	0.581544	0	0.0
1	3	0.364337	0.245649	0	0.0
2	1	0.702722	0.850260	0	0.0
3	0	-0.650819	-0.560498	1	0.0
4	1	1.244139	1.051797	0	1.0

Figure 1: First Few Rows of Cleaned Data

The first impression of the dataset will be reflected by the following figure, which shows the first five rows that started being cleaned on the database, including features like Weather, Road_Type, Time_of_Day, Traffic_Density, Speed_Limit, etc. After the data cleaning process, it will give an overview of the dataset by the sorting and transformation of categorical and numerical features.

3.2 Exploratory Data Analysis (EDA)

Summary statistics of the cleaned dataset:

	Weather	Road_Type	Time_of_Day	Traffic_Density	Speed_Limit
count	798.000000	798.000000	798.000000	7.980000e+02	7.980000e+02
mean	1.171679	1.072682	1.107769	3.116416e-17	1.892109e-16
std	1.241018	0.949004	1.055981	1.000627e+00	1.000627e+00
min	0.000000	0.000000	0.000000	-1.313676e+00	-1.294597e+00
25%	0.000000	0.000000	0.000000	-1.313676e+00	-6.551120e-01
50%	1.000000	1.000000	1.000000	-1.465066e-02	-3.353693e-01
75%	2.000000	1.000000	2.000000	1.284374e+00	3.041162e-01
max	4.000000	3.000000	3.000000	1.284374e+00	4.556695e+00

	Number_of_Vehicles	Accident_Severity	Road_Condition	Vehicle_Type
count	7.980000e+02	798.000000	798.000000	798.000000
mean	1.113006e-17	1.199248	0.931078	1.314536
std	1.000627e+00	0.583460	1.157166	0.719029
min	-1.159005e+00	0.000000	0.000000	0.000000
25%	-6.505464e-01	1.000000	0.000000	1.000000
50%	-1.420880e-01	1.000000	0.000000	1.000000
75%	3.663704e-01	2.000000	2.000000	1.000000
max	5.450954e+00	2.000000	3.000000	3.000000

	Driver_Age	Driver_Experience	Road_Light_Condition
count	7.980000e+02	7.980000e+02	798.000000
mean	-1.513688e-16	8.904044e-17	0.563910
std	1.000627e+00	1.000627e+00	0.655295
min	-1.733653e+00	-2.038436e+00	0.000000
25%	-8.538507e-01	-8.292142e-01	0.000000
50%	9.362845e-02	1.112915e-01	0.000000
75%	8.380764e-01	8.502603e-01	1.000000
max	1.717878e+00	1.992303e+00	2.000000

Figure 2: Summary Statistics

Above are the summary statistics of the dataset, count, mean, standard deviation, minimum, and maximum variable for each feature. These statistics provide insights into the characteristics of the distribution and the center tendency of numerical features, that prove to be useful in finding out patterns or anomalies in the data.

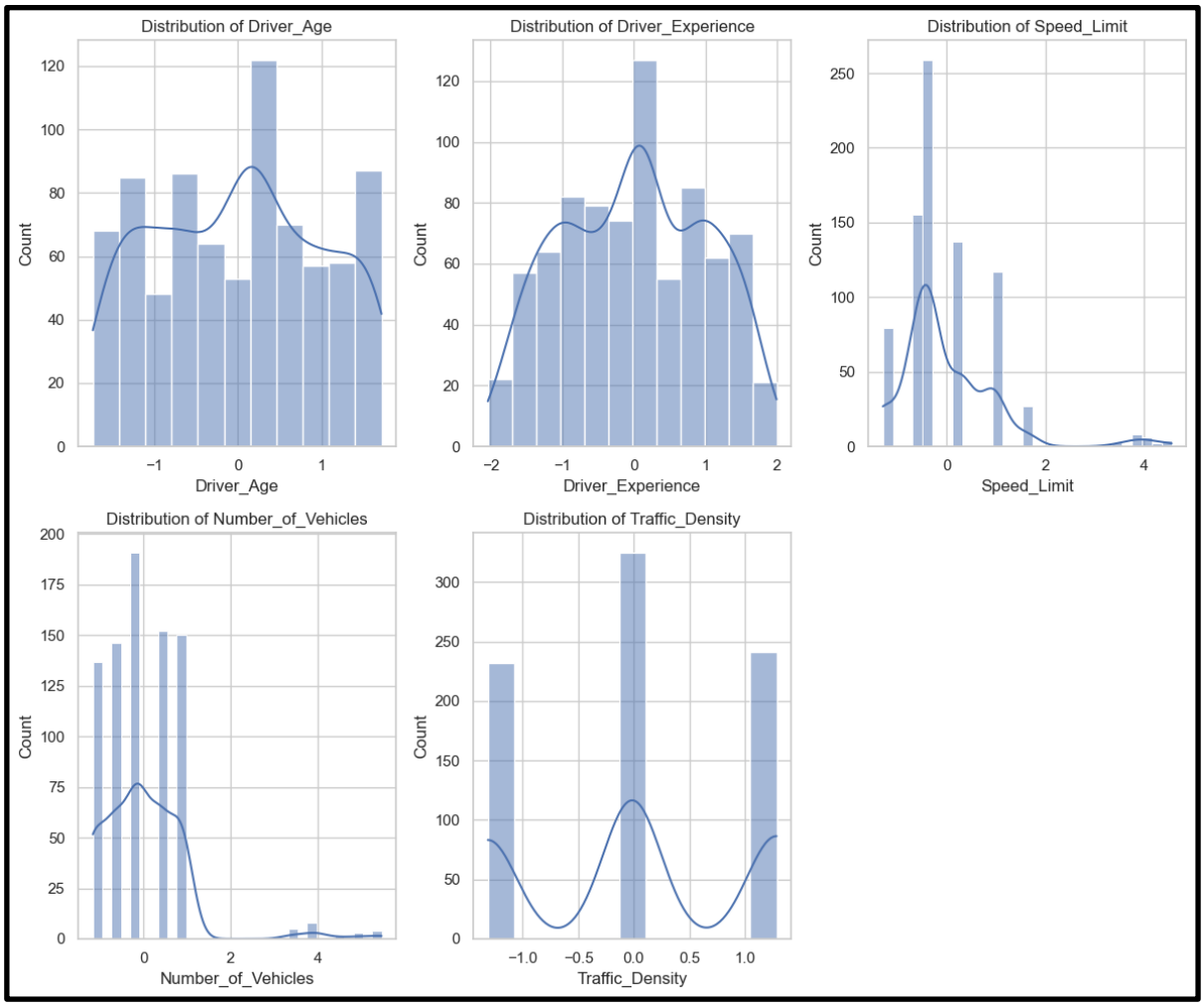

Figure 3: Distribution Plots of Numerical Features

This is a sample plotting of various numerical features like Driver_Age, Driver_Experience, Speed_Limit, Number_of_Vehicles, and Traffic_Density. The interpretation of the visualizations brings the skewness and spread of the data in hand as can be seen in these visualizations, and can be discussed on how to normalize or transform data.

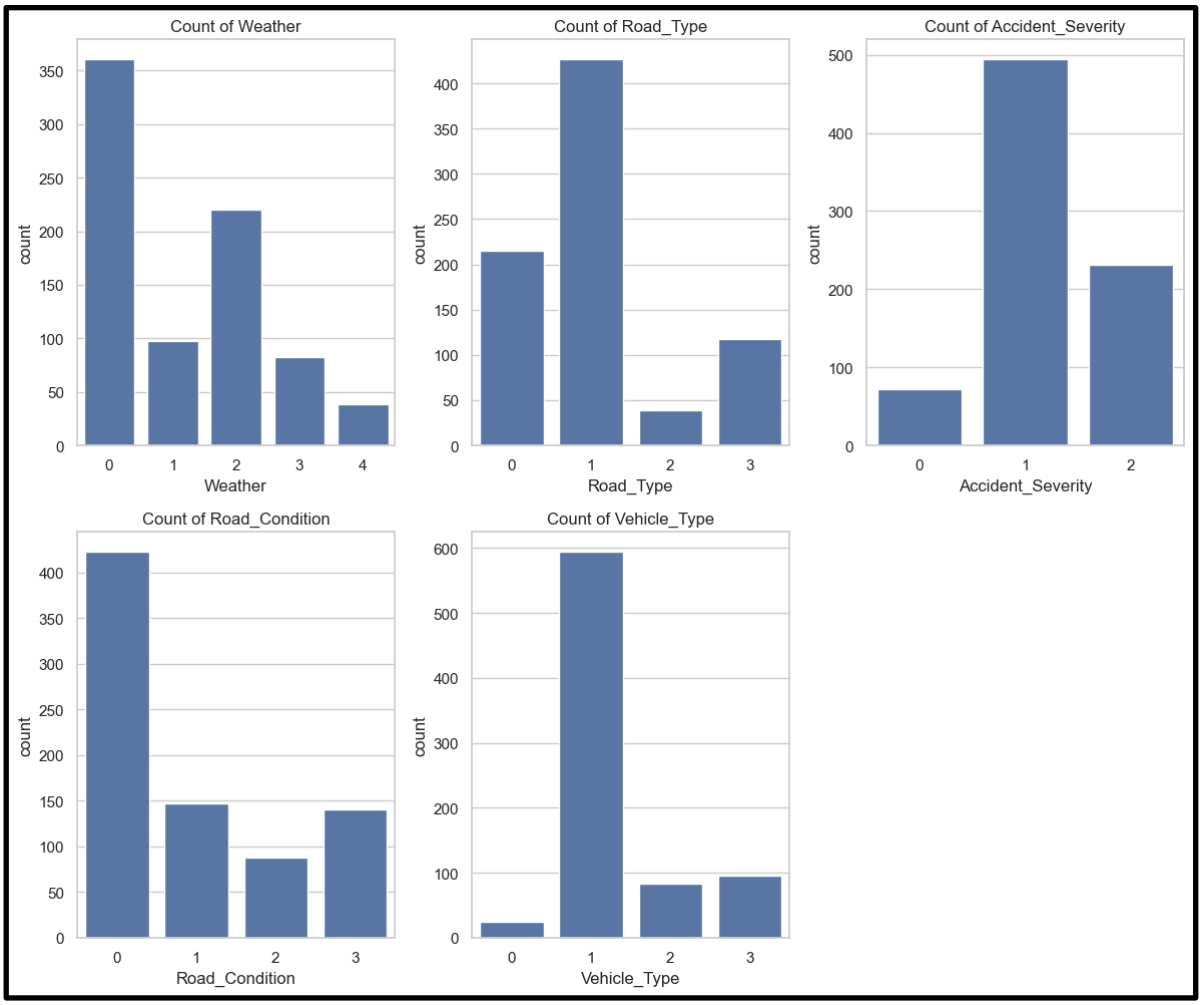

Figure 4: Count Plots of Numerical Features

Above is a plot of count plots for categorical features like Weather, Road_Type, Accident_Severity, Road_Condition and Vehicle_Type. These plots are helpful in identifying class imbalance or skewed distribution of each category and frequency distribution of each category in general.

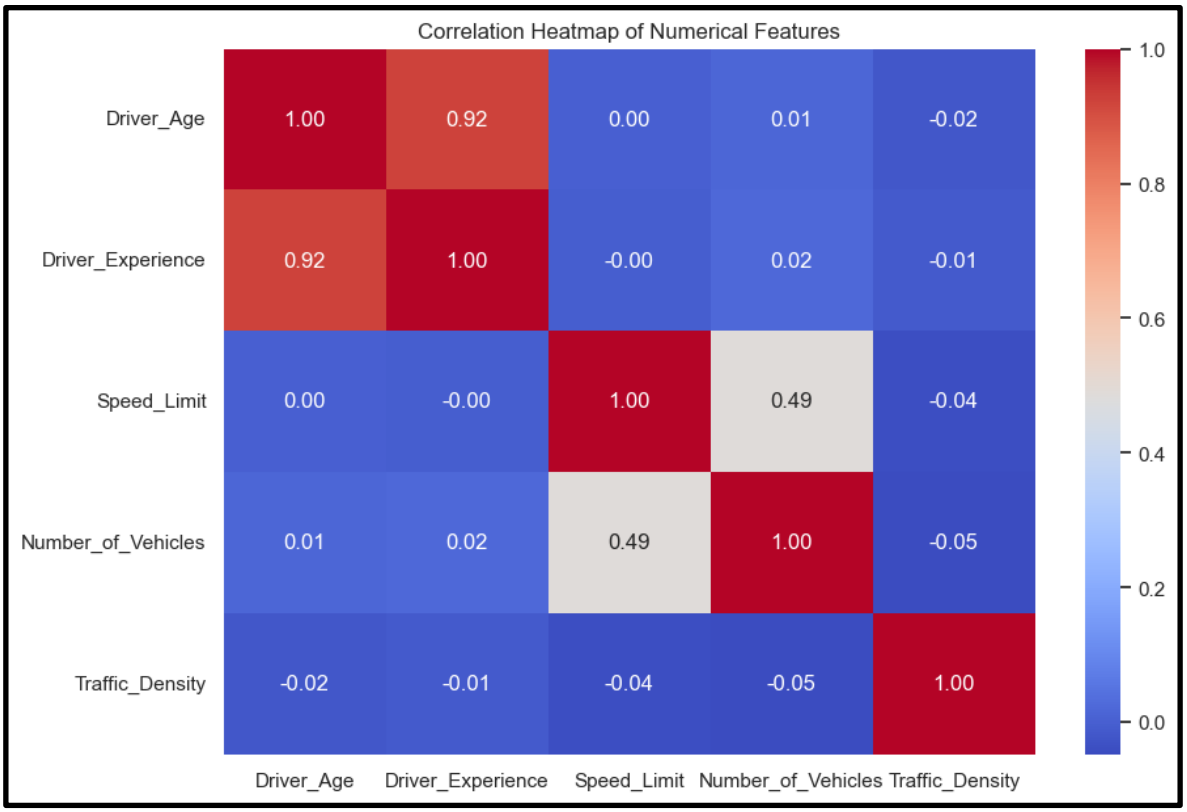

Figure 5: Correlation Heatmap of Numerical Features

The correlation heatmap shows the numerical features within the dataset, namely Driver_Age, Traffic_Density and Speed_Limit for example. Brighter colors indicate stronger correlations and can help indicate features which may affect other features and help identify feature selection decisions.

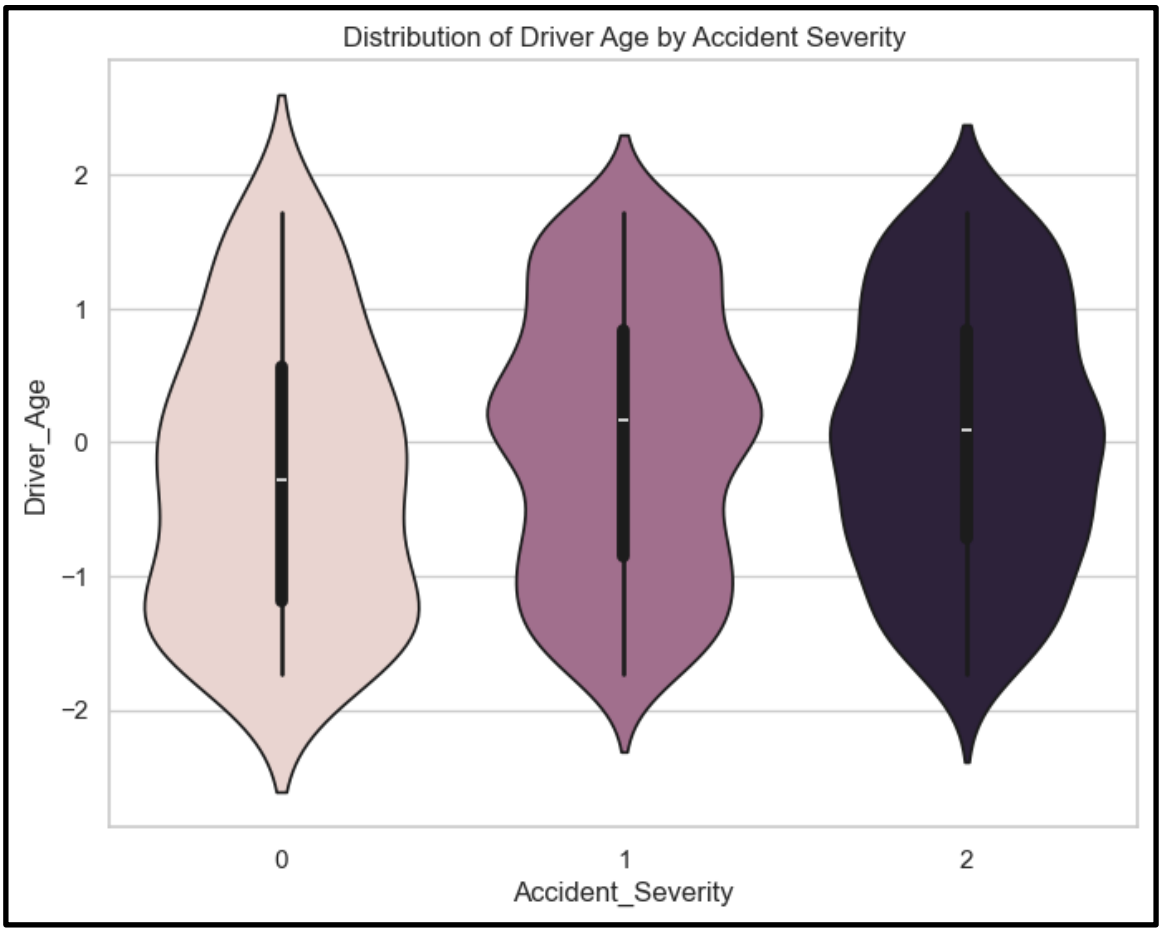

Figure 6: Distribution of Driver Age by Accident Severity

This violin plot shows Driver_Age distribution at different levels of Accident_Severity". It depicts how age of driver affects the severity of crash, whether young or older drivers unhappily are engaged in serious crashes.

3.3 Model Evaluation

```
Training set shape: (638, 13)
Test set shape: (160, 13)
```

Figure 7: Train-Test Split

The shapes of the training and test sets after the split are shown in this figure. Finally, the training set consists of 638 samples and the test set of 160 samples. This helps the model be trained on most of the data but still has a separate set for evaluation.

```
Selected Features using RFE: Index(['Weather', 'Road_Type', 'Traffic_Density', 'Speed_Limit',
       'Number_of_Vehicles', 'Accident_Severity', 'Road_Condition',
       'Vehicle_Type', 'Driver_Age', 'Driver_Experience'],
      dtype='object')
```

Figure 8: Selected Features

The features on this figure represent those which has been selected after Recursive Feature Elimination (RFE) has been applied to the dataset. The most important predictors for accident outcomes are chosen as Features like Weather, Road_Type, Traffic_Density, etc., that would allow to simplify the model by excluding less relevant features.

```
Accuracy: 0.51
Precision: 0.32
Recall: 0.54
F1 Score: 0.40
```

Figure 9: Evaluation Metrics

And here are the figures containing the key evaluation metrics, such as accuracy, precision, recall and F1 score for the trained logistic regression model. The overall accuracy of the model is 0.51 and has 0.32 precision meaning that there is some work to be done in improving the model.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.50      0.59       112
           1       0.32      0.54      0.40        48

    accuracy                           0.51       160
   macro avg       0.52      0.52      0.49       160
weighted avg       0.60      0.51      0.53       160
```

Figure 10: Classification Report

A classification report gives a sense of how the model performs in terms of what is variously known as precision, recall, and F1 score for each class (injury/no injury). However, it shows the strengths and weaknesses of the model, especially, its higher recall in predicting injury outcomes.

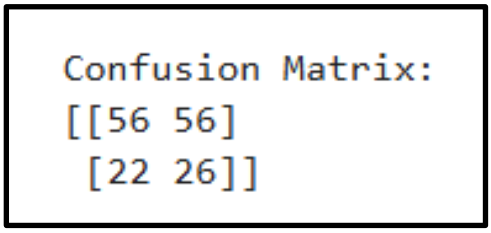

Figure 11: Confusion Matrix

This explains through a confusion matrix the number of true positives, which are the number of examples predicted to be positive but they are positive in reality, and the number of false positives, i.e. number of examples predicted to be positive but in reality they are not, etc. It shows the performance of the model in giving a detailed outcome of "no injury".

3.4 Interpretation of Results

Numerous key features to the performance of the logistic regression model. Outcomes of accidents were predicted by Driver_Age and Traffic_Density. The feature's higher coefficients indicate that younger drivers (either less experienced drivers in general or drivers younger than 17) were more likely to be involved in an accident. The impact of Traffic_Density was also important, as there were more accidents and with greater traffic congestion more accidents could occur in particular, and there could be more and more accidents involving injury. Another important variable was Weather and Road_Type which was important for determining the severity of the accident with poor weather conditions and specific road types evincing higher accident severity. But some features such as Vehicle_Type or Speed_Limit had weak associations perhaps due to the observation of little variability in the dataset.

These results align with traffic safety studies which have observed that younger drivers and high traffic volumes associate with greater likelihood of accident. The results from the model show the necessity for such interventions, especially in improving road conditions, enforcing speed limits and providing additional training for young drivers.

3.5 Model Limitations

However, logistic regression is suitable for binary classification. The model is made under the assumption that there's a linear relation of the predictors on the outcome, which may not exactly fit accurate relations of the features with the outcome. Added to that, the dataset had more instances of "no injury" than "injury," which is referred to as class imbalance.

The likely result of this imbalance was lower precision in predicting injury outcomes. Moreover, the model can miss linear and non-linear interactions as well as higher-order relations of the features that might have enhanced prediction accuracy. Finally, Driver_Age and Driver_Experience feature correlations, as well as any others, may have decreased model efficiency and contributed to multicollinearity.

4. Model Visualizations

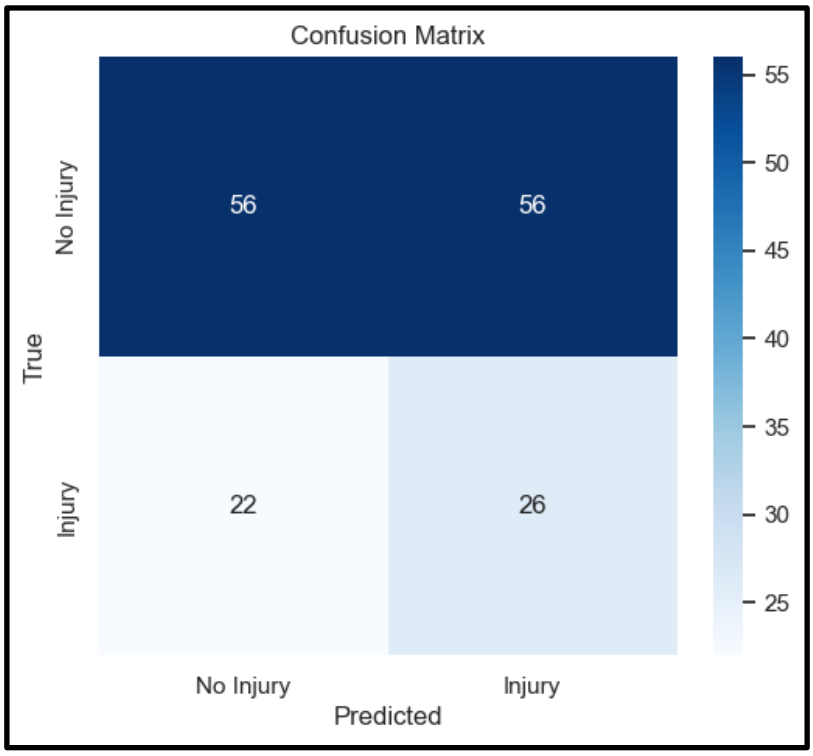

Figure 12: Confusion Matrix

A visualization for the performance of the logistic regression model goes as the confusion matrix. It provides the counts of true positives, true negatives, false positives, and false negatives. Here,

the model made 56 true negatives (no injury) and 56 false positives. In fact, for the injury category, it correctly predicted 26 injuries, but missed 22 injuries.

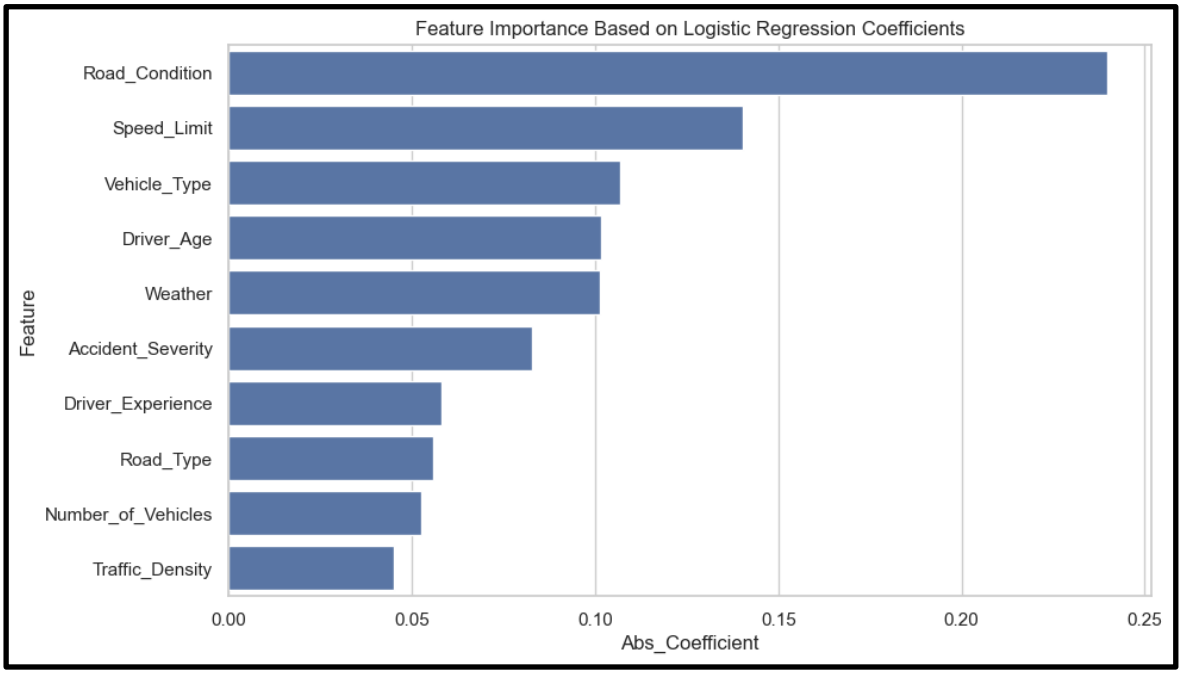

Figure 13: Feature Importance Plot

Each feature is shown as importance to predicting either the accident outcome or not in terms of the magnitude of the logistic regression coefficient. Road_Condition, Speed_Limit and Vehicle_Type are the most influential features among them. The absolute coefficients related to these features are largest, hence, indicating that they have a huge impact on predicting accident severity.

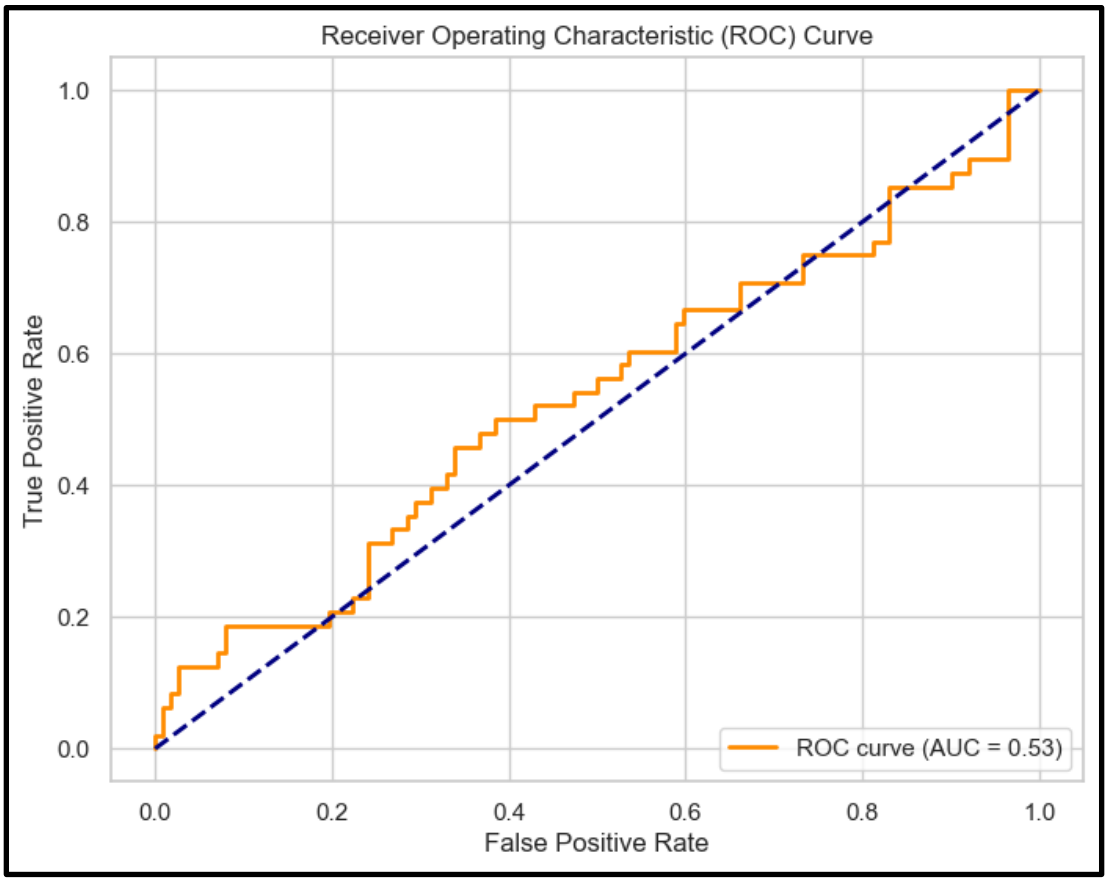

Figure 14: ROC Curve

The Receiver Operating Characteristic (ROC) curve shows the relationship between the true positive rate and the false positive rate. The model performance according to the curve is to distinguish between injury and no injury. The model can discriminate between the two classes with an accuracy of just marginally better than random guessing (AUC of 0.53).

5. Conclusion

Summary of Findings

Using the logistic regression model the accuracy of 51% is obtained with reasonable features like Road_Condition, Speed_Limit, and Vehicle_Type plays a major role in the outcome of accident. Overall, it offered a higher recall in the prediction of injuries, but performance could be improved.

Implications of the Study

By identifying high-risk factors, the model can provide assistance to improving traffic safety through either, specific interventions such as road safety improvement and education of drivers.

Future Work

In enhancement of prediction accuracy, Random Forest or Neural Networks are also worth exploring. Moreover, the dataset size can be increased and the task of handling class imbalance can be addressed to improve the result further.

References

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H.M., Almoshaogeh, M., Farooq, D. and Ahmad, M., 2021. Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International journal of injury control and safety promotion*, *28*(4), pp.408-427. https://www.tandfonline.com/doi/abs/10.1080/17457300.2021.1928233

Kumar, A., 2022. *Unbalanced data classification Using genetic programming* (Doctoral dissertation, Bennett university). http://lrcdrs.bennett.edu.in/handle/123456789/2016

Saran, N.A. and Nar, F., 2025. Fast binary logistic regression. *PeerJ Computer Science*, *11*, p.e2579. https://peerj.com/articles/cs-2579/

Song, X., Liu, X., Liu, F. and Wang, C., 2021. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International journal of medical informatics*, *151*, p.104484. https://www.sciencedirect.com/science/article/pii/S1386505621001106