# Road Safety Forecasting :

***A Predictive Modelling of Accidents Severity in Addis ababa.***

## Report compiled by :

1. Bonventure Osoro.
2. Brian Kipkorir.
3. Faith Nyawira.
4. Jacinta Fiona.
5. James Murigi.
6. Sandra Luyali.

## Introduction :

In urban areas, road accidents profoundly common. As urban landscapes grow, accurate accident severity classification gains importance. This empowers urban planners, law enforcement, and policymakers to allocate resources effectively, enhance safety measures, and optimize emergency responses. Our project aims to create a safer city environment by improving accident classification, aiding in informed urban development decisions.

# stakeholders:

**Urban Planning Institutes:** Designs city layout and uses accident severity data to build safer roads.

**Emergency Services:** Fast accident response. Accurate classification aids resource allocation, saving lives.

**Road Safety Authority:**Reduces accidents by creating safety measures, policies based on accident classification

# Objectives of the Project :

1. **Inform Urban Planning Decisions**: Provide actionable insights to the Urban Planning Institute for identifying high-risk zones, improving safety measures, and aligning urban development strategies.
2. **Optimize Emergency Response**: Empower emergency services with rapid, accurate severity assessments to optimize resource allocation and response effectiveness.
3. **Shape Road Safety Initiatives**: Assist the Road Safety Authority in evaluating measures and formulating policies tailored to different accident scenarios.

# Dataset overview:

**Source**: [Addis Ababa Sub city police department records]

**Time Period**: Year 2017-2020

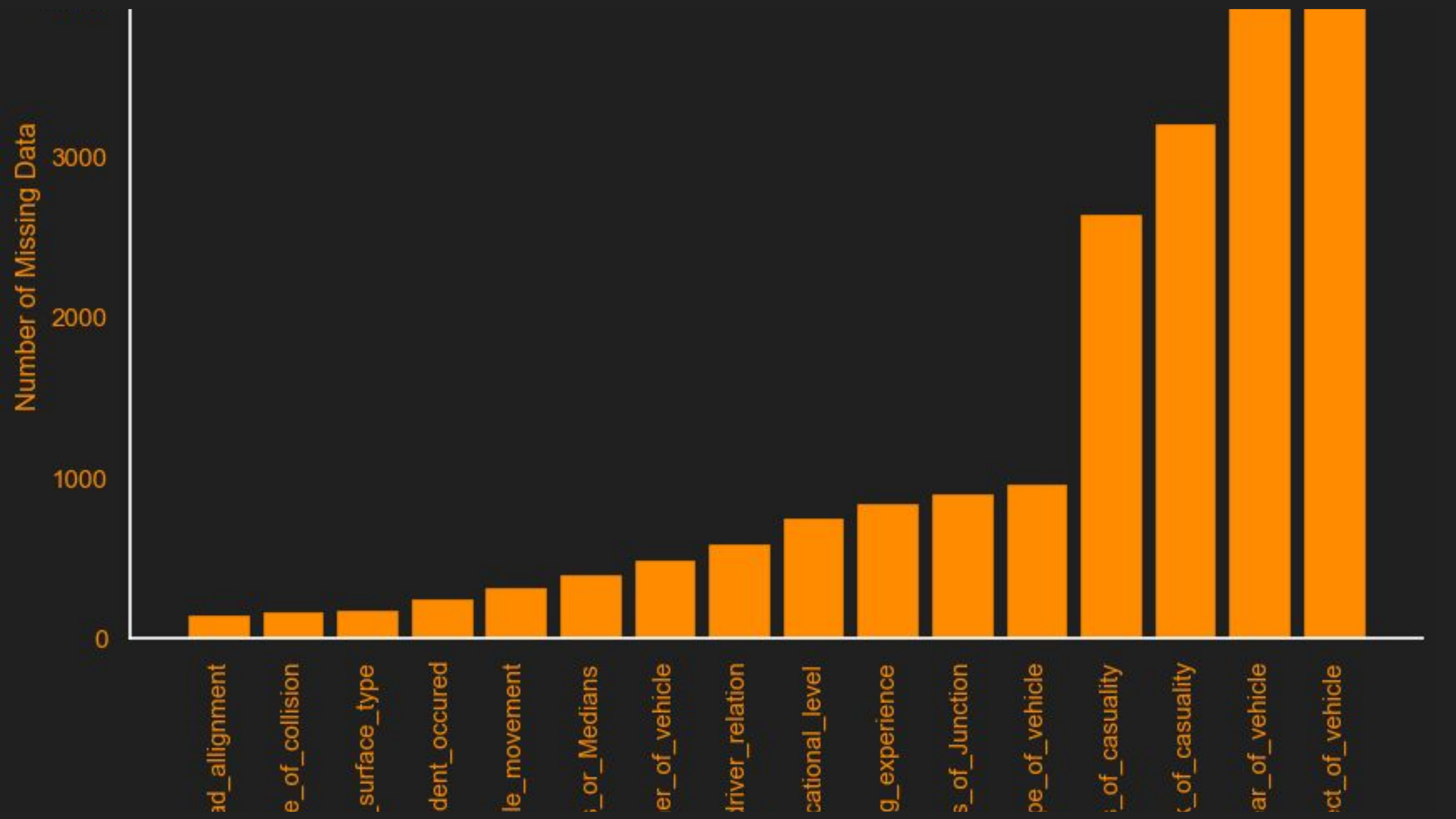
**Number of Instances (Rows)**: 12316

**Number of Features (Columns)**: 32

# Data Cleaning :

1. Handled Missing values : We treated "na", "Unknown", "Unknown or other", and "unknown" as Null and dropped all rows with null values
2. Removed Duplicates : We dropped duplicates
3. We changed ‘Time’ feature to ‘Hour’ (from ‘13:32:00’ to 13)
4. Transformed categorical data using label encoding

**The resulting dataset had 8300 rows and 22 column


Number of Missing Data
3000
2000
1000
0
ad_allignment
e_of_collision
surface_type
dent_occured
le_movement
_or_Medians
er_of_vehicle
driver_relation
cational_level
g_experience
s_of_Junction
pe_of_vehicle
_of_casuality
_of_casuality
ar_of_vehicle
ct_of_vehicle

# Exploratory Data Analysis (EDA)

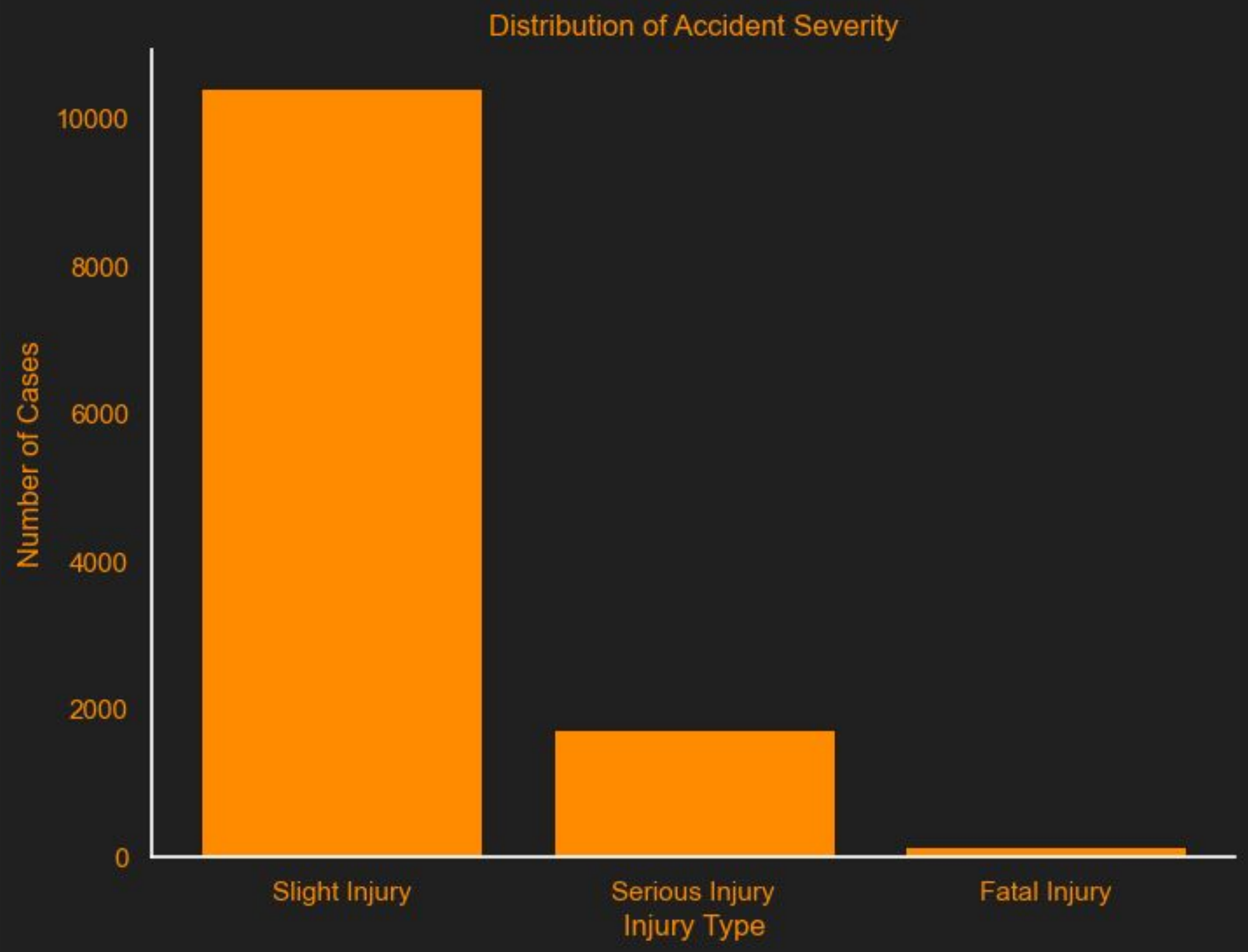
Carried out Bivariate and Univariate analysis.

Bivariate analysis examines variable pairs, using tests like Anova and Chi-Squared to reveal associations and correlation. Univariate analysis assesses individual feature distributions, crucial for model selection.

# Finding 1 : class Imbalance

There is significant class imbalance in the target variable y(‘Accident_severiry’) :

1. Slight Injury cases….10415
2. Serious injury cases….1743
3. Fatal injury cases….158


Distribution of Accident Severity
Number of Cases
10000
8000
6000
4000
2000
0
Slight Injury
Serious Injury
Fatal Injury
Injury Type

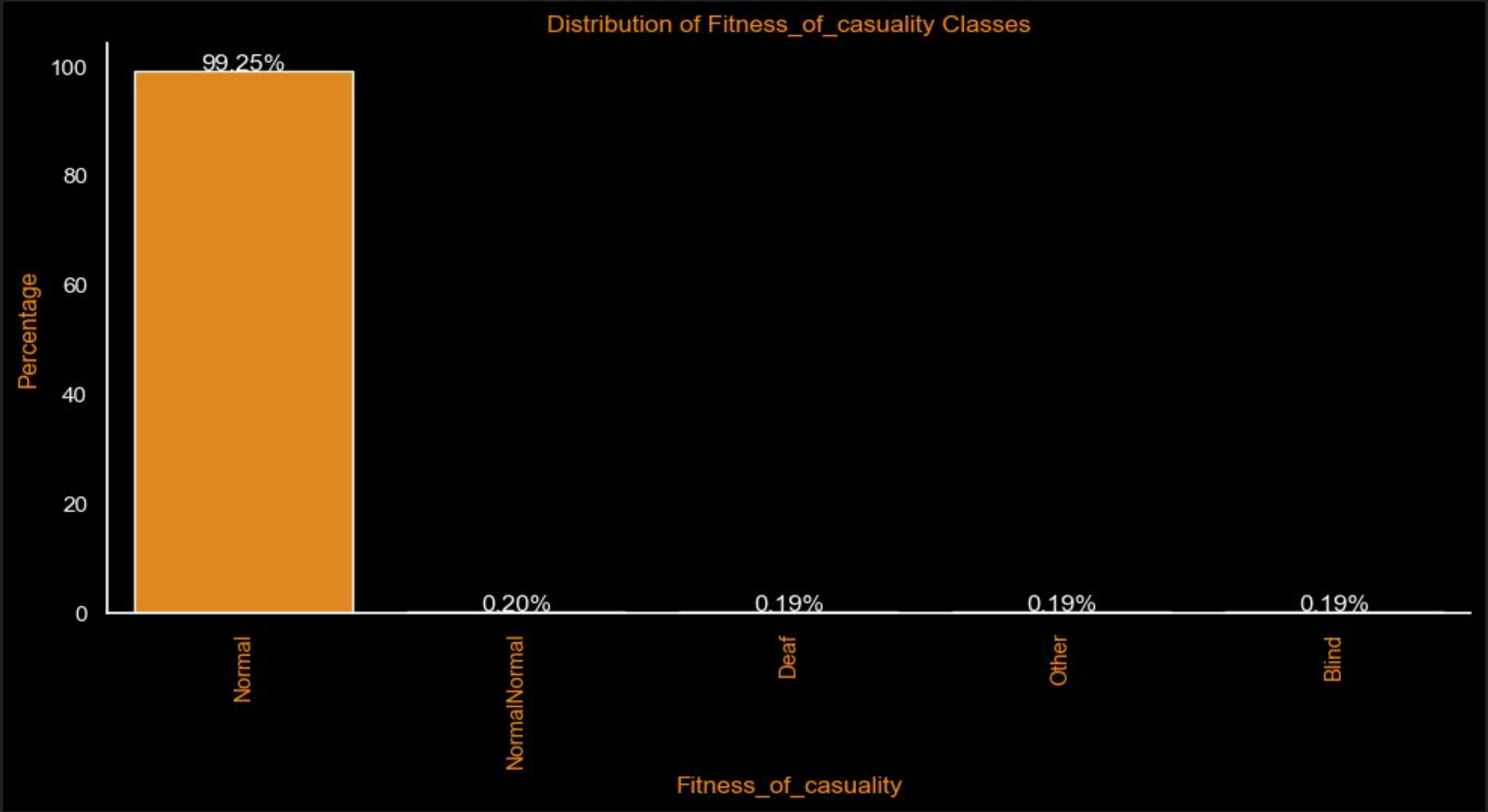
# Finding 2 :

Feature imbalance: Many features show imbalances , impacting potential contributions eg :

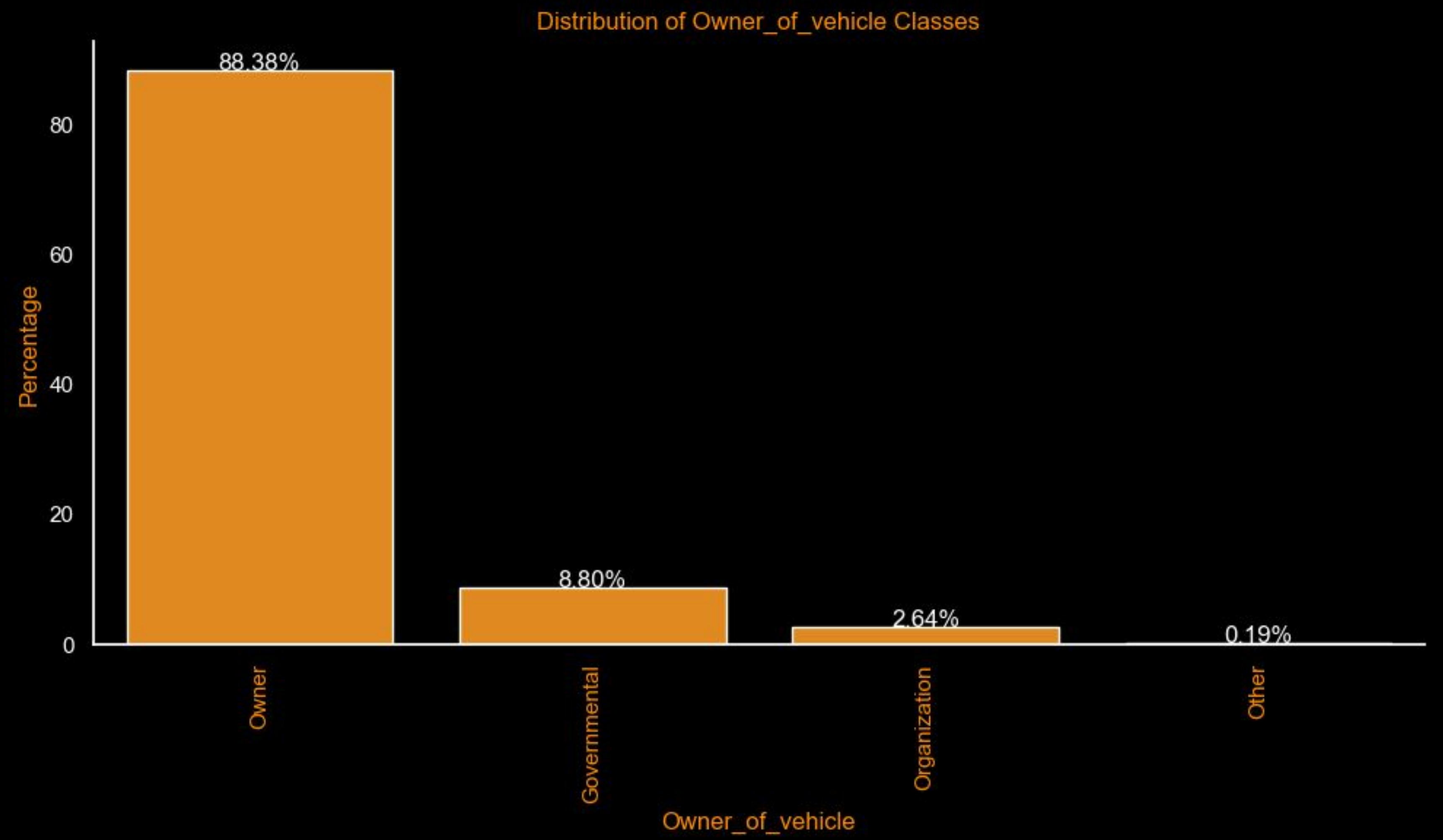
Fitness of casualty

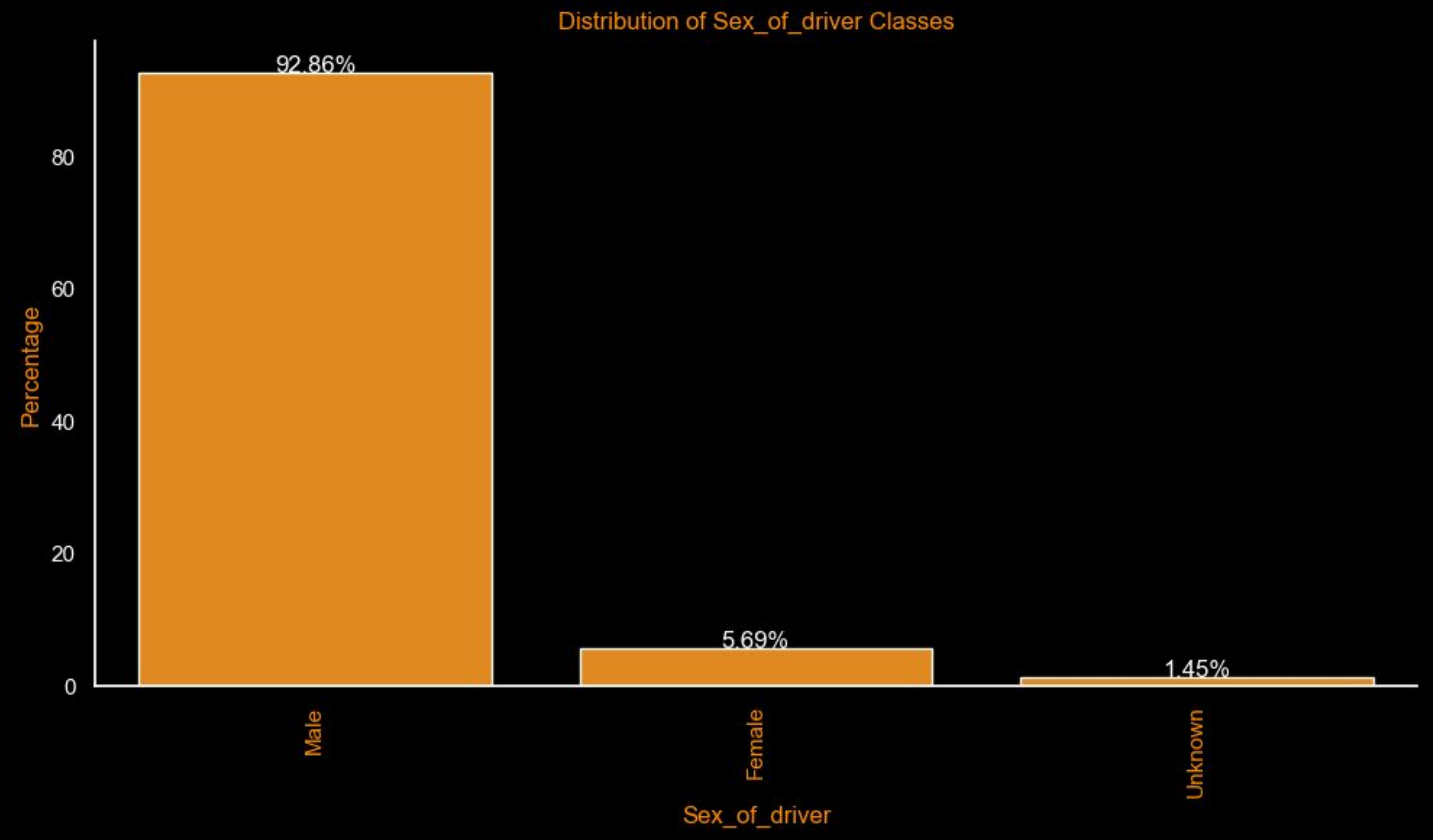
Sex of driver

Owner of the vehicle



Distribution of Fitness_of_casuality Classes
Percentage
Fitness_of_casuality
0
20
40
60
80
100
99.25%
0.20%
0.19%
0.19%
0.19%
Normal
NormalNormal
Deaf
Other
Blind


Distribution of Owner_of_vehicle Classes
88.38%
8.80%
2.64%
0.19%
Percentage
80
60
40
20
0
Owner
Governmental
Organization
Other
Owner_of_vehicle



Distribution of Sex_of_driver Classes
92.86%
5.69%
1.45%
Percentage
0
20
40
60
80
Male
Female
Unknown
Sex_of_driver

# Modeling :

STRATEGY:

1. Create baseline model

2. Create more advanced models using:

 Model_1 (Decision trees) ,Model_2 (Logistic Regression), Model_3 (KNN classifier), Model_4 (Random Forests)

3. select best model.

4. Hyperparameter tuning.(GridSearchCv)

 Model_5 (best model hyperparameter tuning)

5. Feature selection and use of ensemble methods. ( Model_6)

# Baseline Model :

The baseline model the most frequent class (mode) of the target variable for all instances. This is the same as someone guessing the most common target variable every time

Baseline model accuracy = 0.5 or 50 %

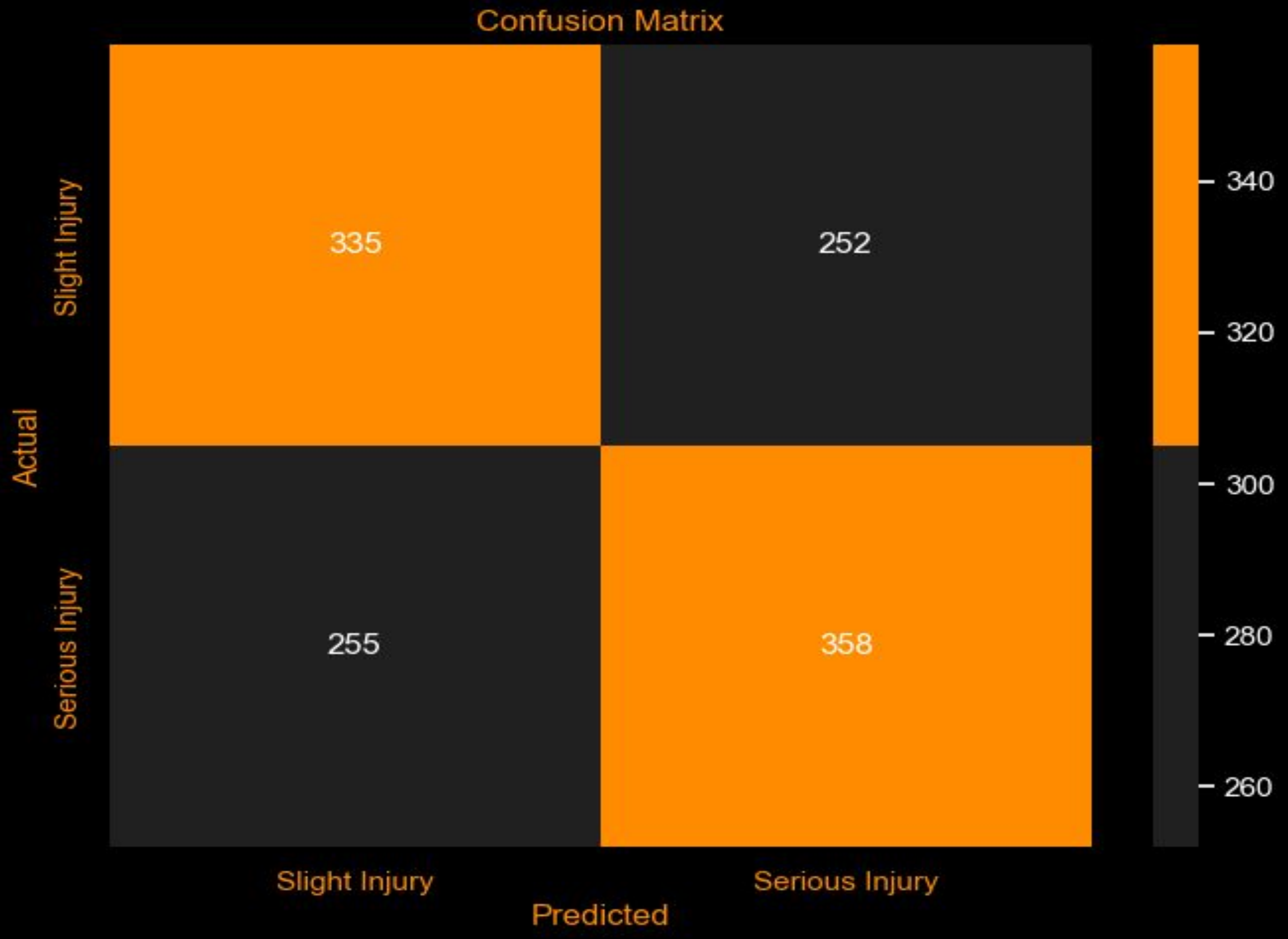
# Model_1 (Logistic Regression)

1. Accuracy: The overall accuracy of the baseline model is 58%

Slight Injury Class: The model achieved a precision of 57% and a recall of 57% for predicting "Slight Injury" cases. Its F1-score for this class is also 57%

Serious Injury Class: The precision for predicting "Serious Injury" instances is 59%, while recall is 58%. The F1-score for this class is 59%.

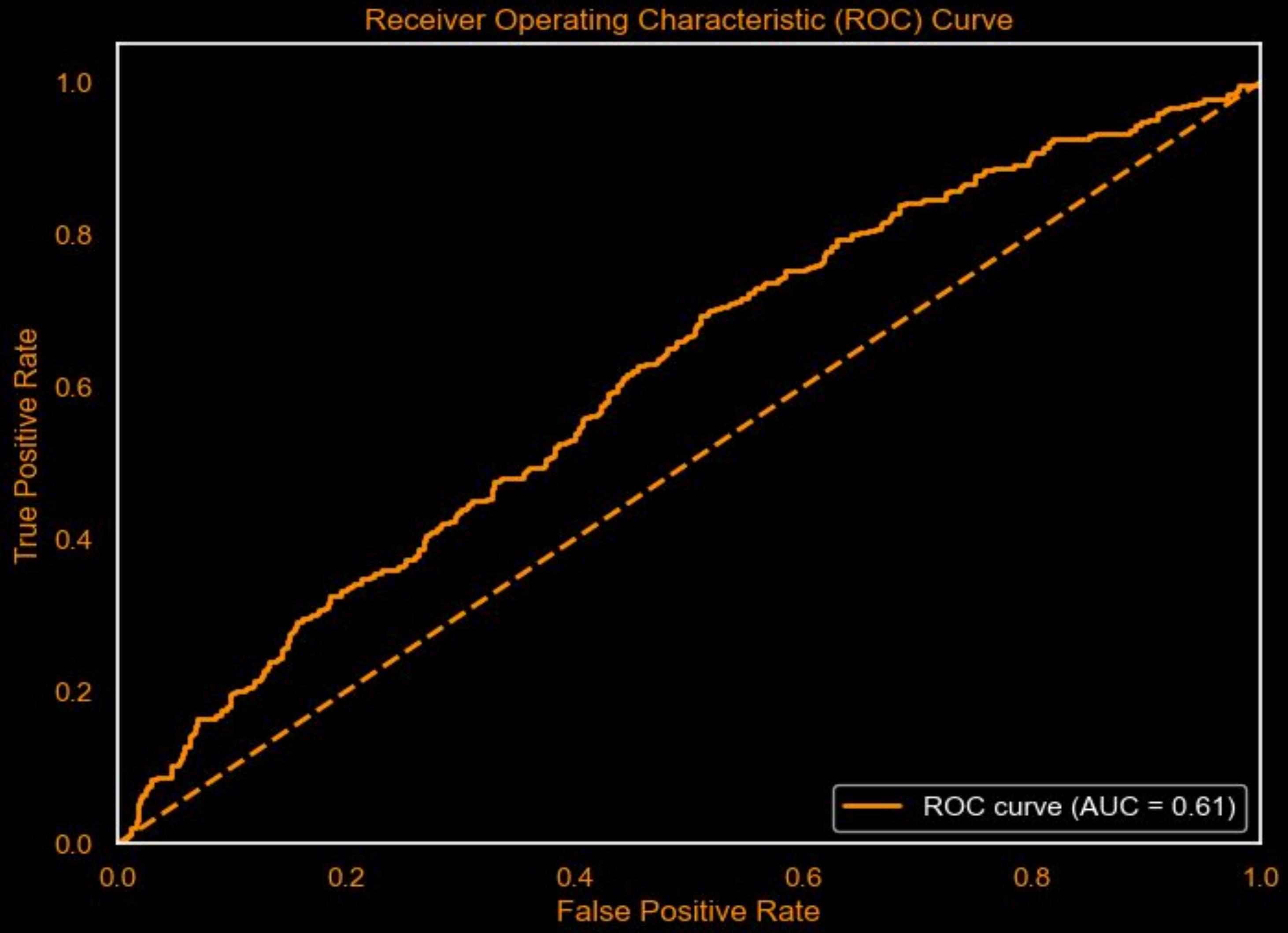
2. Balanced Results: The macro average of precision, recall, and F1-score for both classes is 58%, showing a balanced performance across the two classes.
3. Weighted Average: The weighted average, which considers class imbalances, is also 58%.



Confusion Matrix
335
252
255
358
Slight Injury
Serious Injury
Actual
Slight Injury
Serious Injury
Predicted
340
320
300
280
260

# Model_1 Evaluation:

- Correctly predicted 358 Serious Injury cases.
- Correctly predicted 335 Slight Injury cases.
- Incorrectly predicted 252 Serious Injury cases.
- Incorrectly predicted 255 Slight Injury cases.


Receiver Operating Characteristic (ROC) Curve
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (AUC = 0.61)

# Analysis:

1. Moderate Discrimination: AUC of 0.61 shows moderate accuracy in distinguishing between Slight Injury and Serious Injury cases.
2. Above Random: It performs better than random guessing (AUC > 0.5).
3. Room for Improvement: A higher AUC (closer to 1.0) would suggest improved accuracy.

# Model_2(K-Nearest Neighbour)

Accuracy: The model achieves 65% accuracy, indicating overall correctness in predicting accident severity.

Precision: "Slight Injury": 0.63, "Serious Injury": 0.67. Effectively minimizes false positives for both classes.

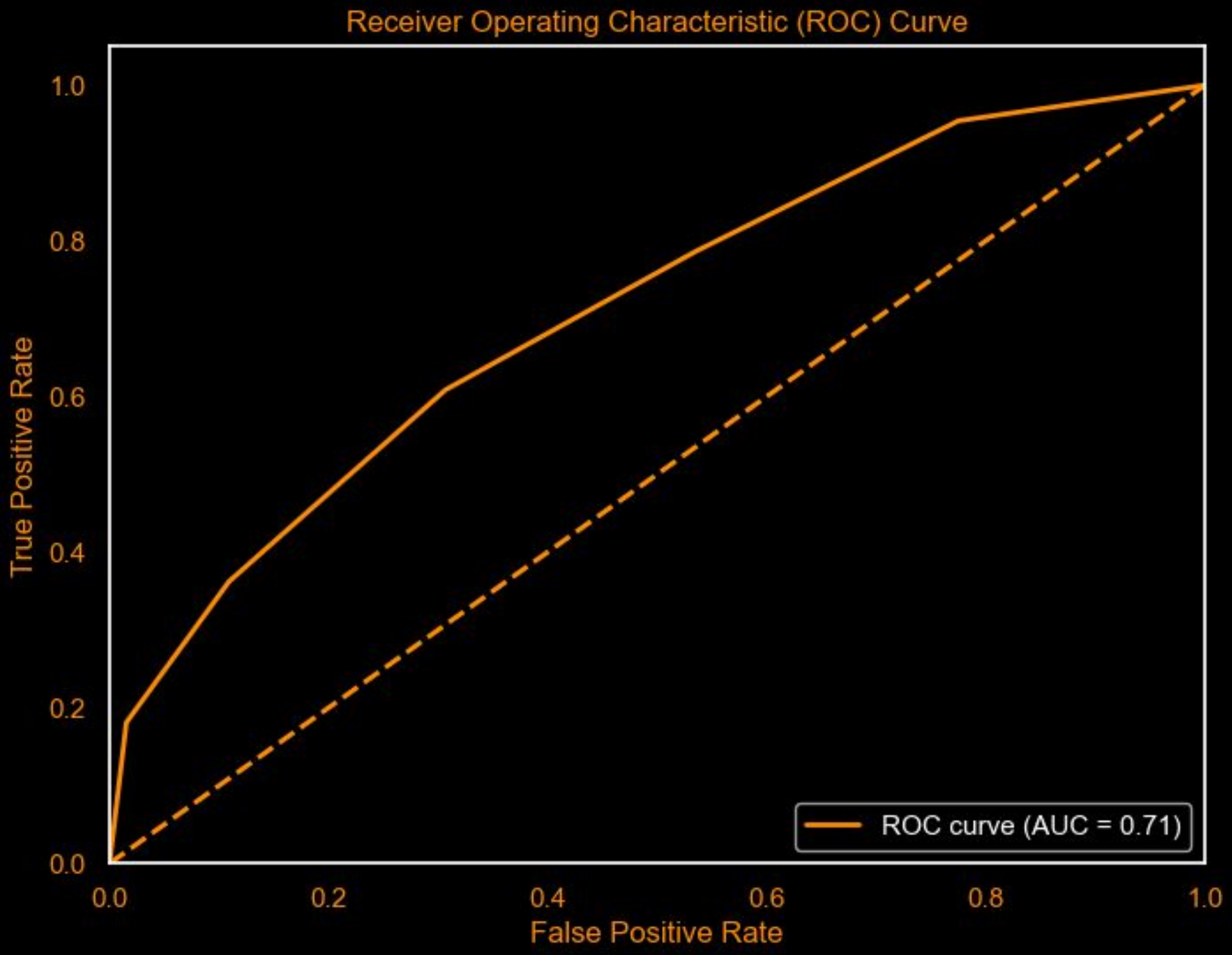
Recall (Sensitivity): "Slight Injury": 0.69, "Serious Injury": 0.61. Captures instances of both classes, slightly favoring "Slight Injury."

F1-Score: "Slight Injury": 0.66, "Serious Injury": 0.64. Balanced trade-off between precision and recall for both classes.


Confusion Matrix
407
180
240
373
Slight Injury
Serious Injury
Actual
Slight Injury
Serious Injury
Predicted
400
350
300
250
200

# Analysis:

1. Correctly predicted 373 Serious Injury cases.
2. Accurately identified 407 Slight Injury cases.
3. Incorrectly predicted 180 Serious Injury cases as Slight Injury.
4. Missed 240 Serious Injury cases.

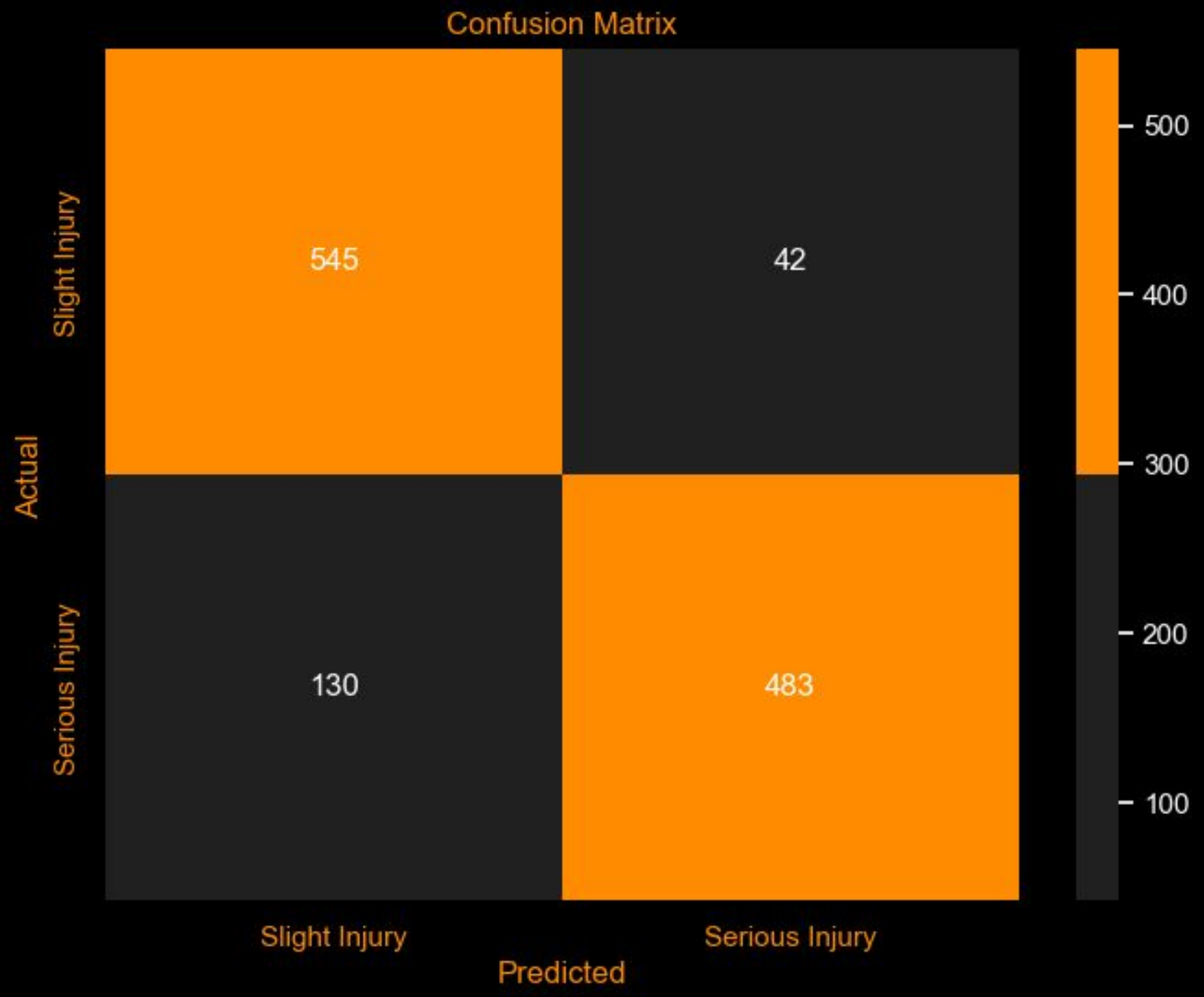

Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.71)

## Analysis :

1. AUC of 0.71 suggests Model_2 reasonably distinguishes between classes.
2. It ranks "Serious Injury" instances higher than "Slight Injury" instances about 71% of the time.
3. Model_2 captures patterns well and reliably separates the classes.

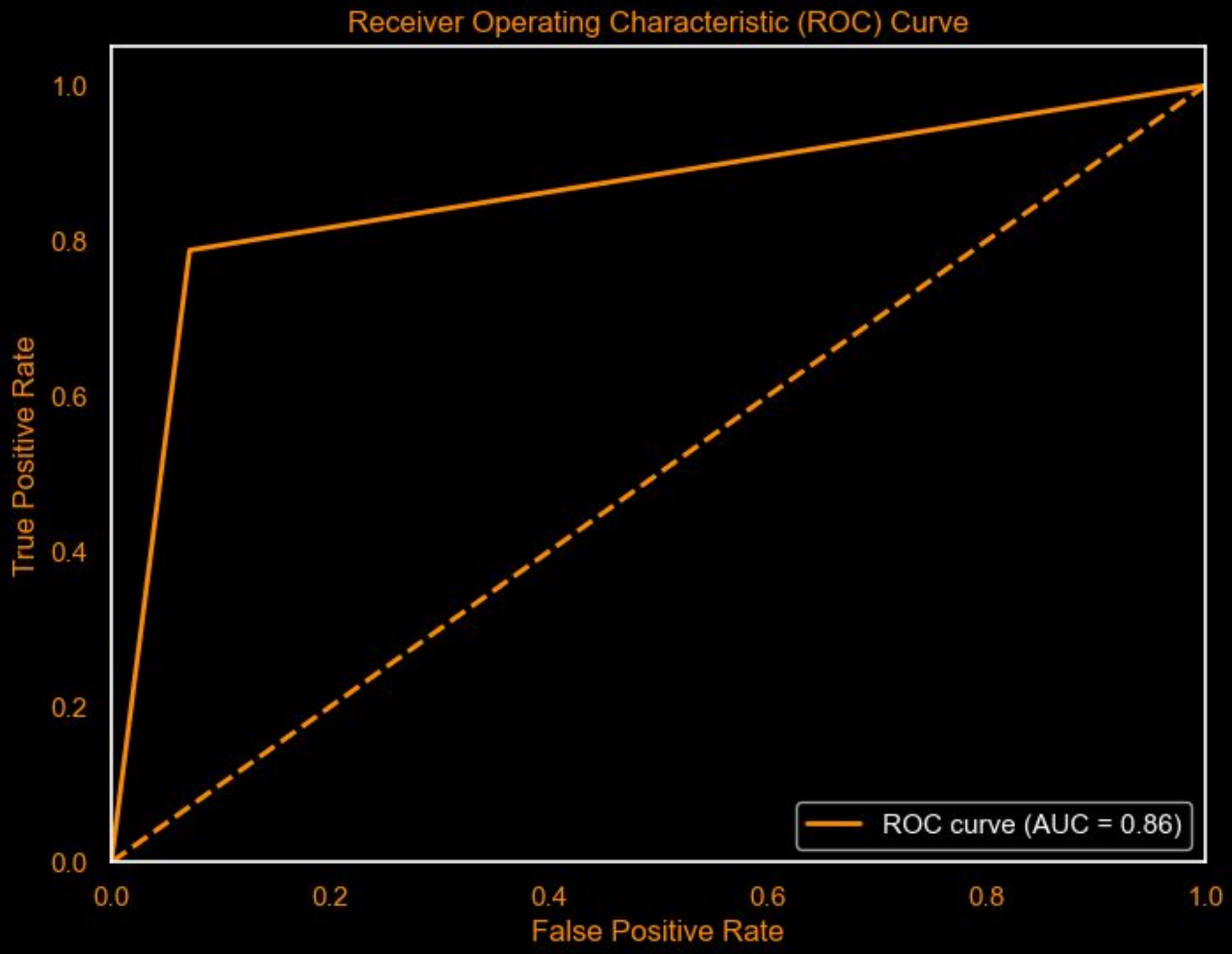
# Model_3(Decision Trees):

1. Precision (Slight Injury):0.81 (81% of predicted "Slight Injury" cases are correct).
2. Recall (Slight Injury):0.93 (93% of actual "Slight Injury" cases were predicted).
3. F1-Score (Slight Injury): 0.86 (balance between precision and recall).
4. Precision (Serious Injury): 0.92 (strong performance).
5. Recall (Serious Injury):0.79 (captured 79% of actual "Serious Injury" cases).
6. F1-Score (Serious Injury): 0.85 (effective precision-recall balance).


Confusion Matrix
545
42
130
483
Slight Injury
Serious Injury
Actual
Slight Injury
Serious Injury
Predicted
500
400
300
200
100

# Analysis :

1. Correctly predicted 483 instances of "Serious Injury" out of 613.
2. Accurately identified 545 cases of "Slight Injury" out of 587.
3. Incorrectly predicted "Serious Injury" instead of "Slight Injury" 42 times.
4. Missed 130 instances of "Serious Injury."


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.86)

## Analysis:

- AUC of 0.86 indicates the model's strong ability to rank a randomly selected "Serious Injury" instance higher than a "Slight Injury" instance around 86% of the time.

- This high AUC suggests the model's predictions closely match the true labels, effectively distinguishing between varying accident severity levels.

- An AUC of 0.86 showcases the model's excellent classification performance and its significant capability to differentiate between the two classes.

# Model_4 (Random Forest):

- Precision: 90% accuracy in classifying instances as "Slight Injury."
- Recall: 90% ability to identify actual "Slight Injury" cases.
- F1-Score: 90% balance between precision and recall.
- Precision: 91% accuracy in predicting "Serious Injury."
- Recall: 90% capture of actual "Serious Injury" instances.
- F1-Score: 90% harmonization of precision and recall.


Confusion Matrix
Actual
Slight Injury
Serious Injury
530
57
61
552
Predicted
Slight Injury
Serious Injury
500
400
300
200
100

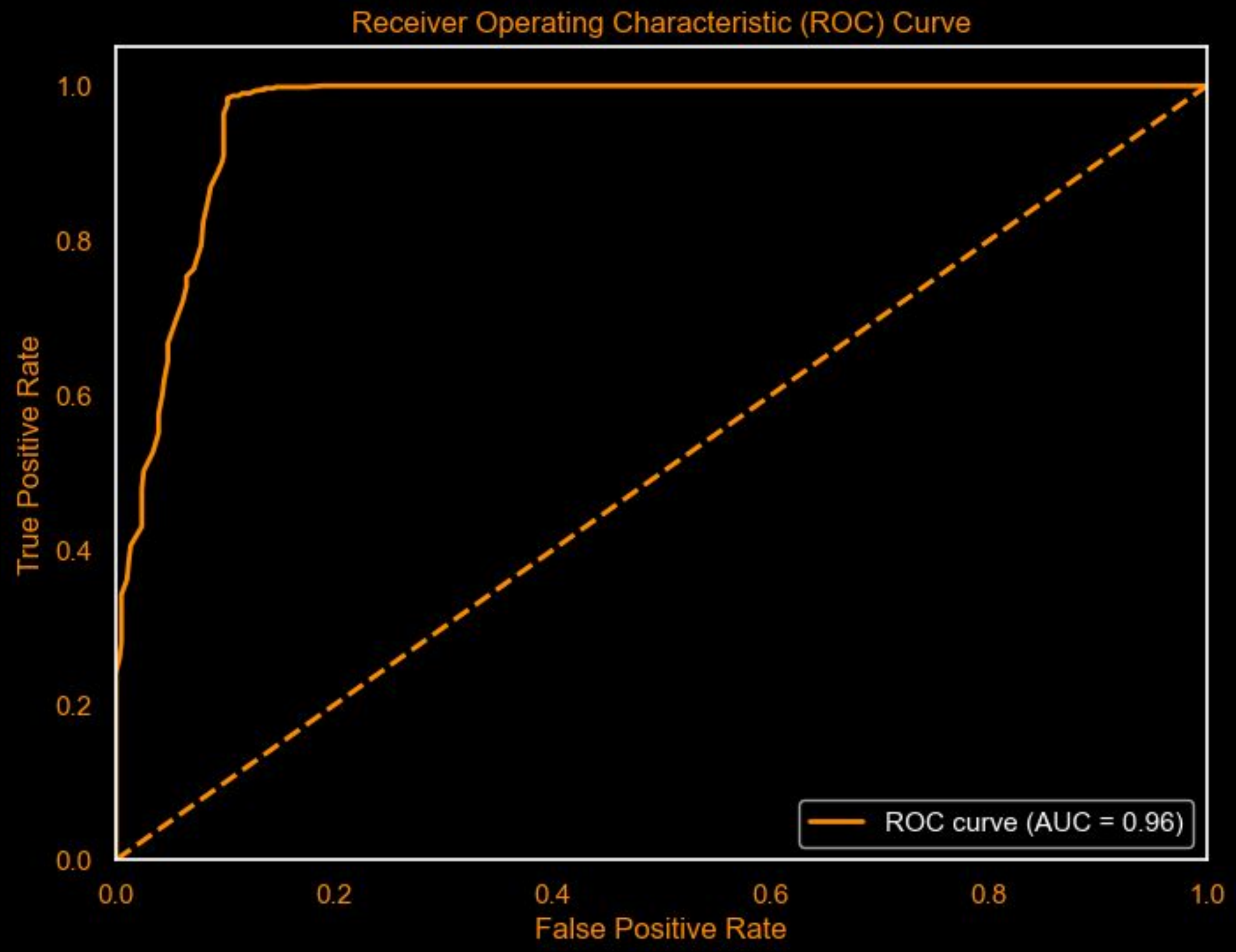
# Analysis:

- The model achieved 552 True Positives, accurately predicting instances of "Serious Injury" out of 613 actual cases.

- It accurately identified 530 True Negatives, instances of "Slight Injury" out of 587 actual cases.

- There were 57 False Positives, instances where "Serious Injury" was incorrectly predicted for "Slight Injury."

- The model missed 61 instances of "Serious Injury" (False Negatives).


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.96)

# Analysis :

An AUC of 0.96 indicates that the model distinguishes between "Slight Injury" and "Serious Injury" instances exceptionally well.

This high AUC value suggests that the model is highly proficient in ranking a randomly selected "Serious Injury" instance higher than a "Slight Injury" instance around 96% of the time.

AUC reflects the model's capability to accurately discriminate between different levels of accident severity and make well-informed predictions.

# Model_5 (Random Forest Hyperparameter tuning) :

Now that we have established that Random Forest classifier is our best model,..we are going to tune it's hyper parameters and do further feature selection to increase its accuracy

for this model we decided to add the following hyper parameters to the randomtree function:

'max_depth': None

'min_samples_leaf': 1

'min_samples_split': 2

'n_estimators': 200

# Analysis :

- Precision: Model's precision is 0.93 for "Slight Injury" and 0.92 for "Serious Injury."

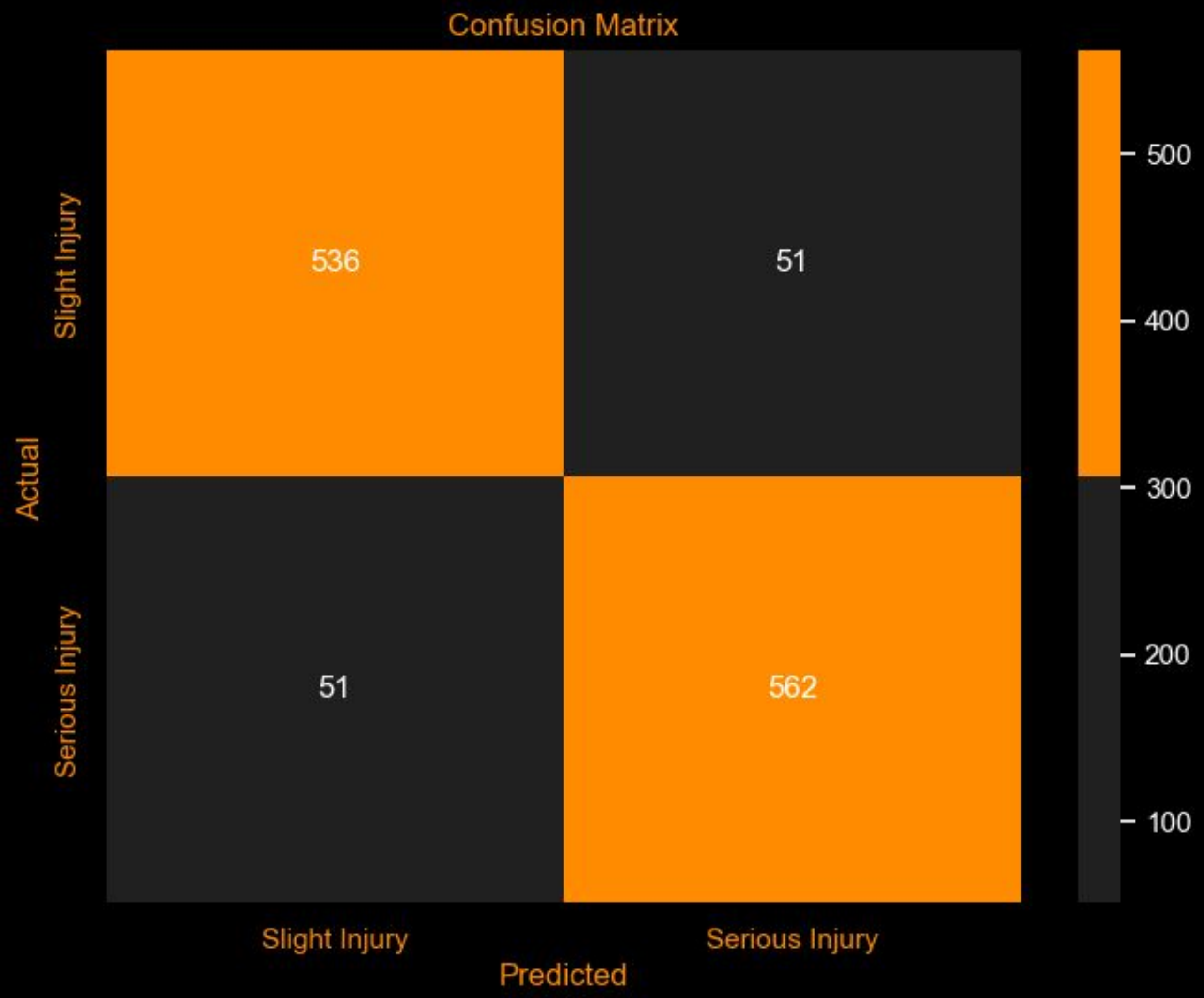
- Recall: Recall is 0.91 for "Slight Injury" and 0.93 for "Serious Injury."

- F1-Score: F1-score is 0.92 for both "Slight Injury" and "Serious Injury."

- Accuracy: Overall accuracy is 0.92.

- Macro Avg: Macro-average F1-score is 0.92, considering class averages.

- Weighted Avg: Weighted-average F1-score is 0.92, considering class distribution.


Confusion Matrix
Slight Injury
Serious Injury
Actual
536
51
51
562
Slight Injury
Serious Injury
Predicted
500
400
300
200
100

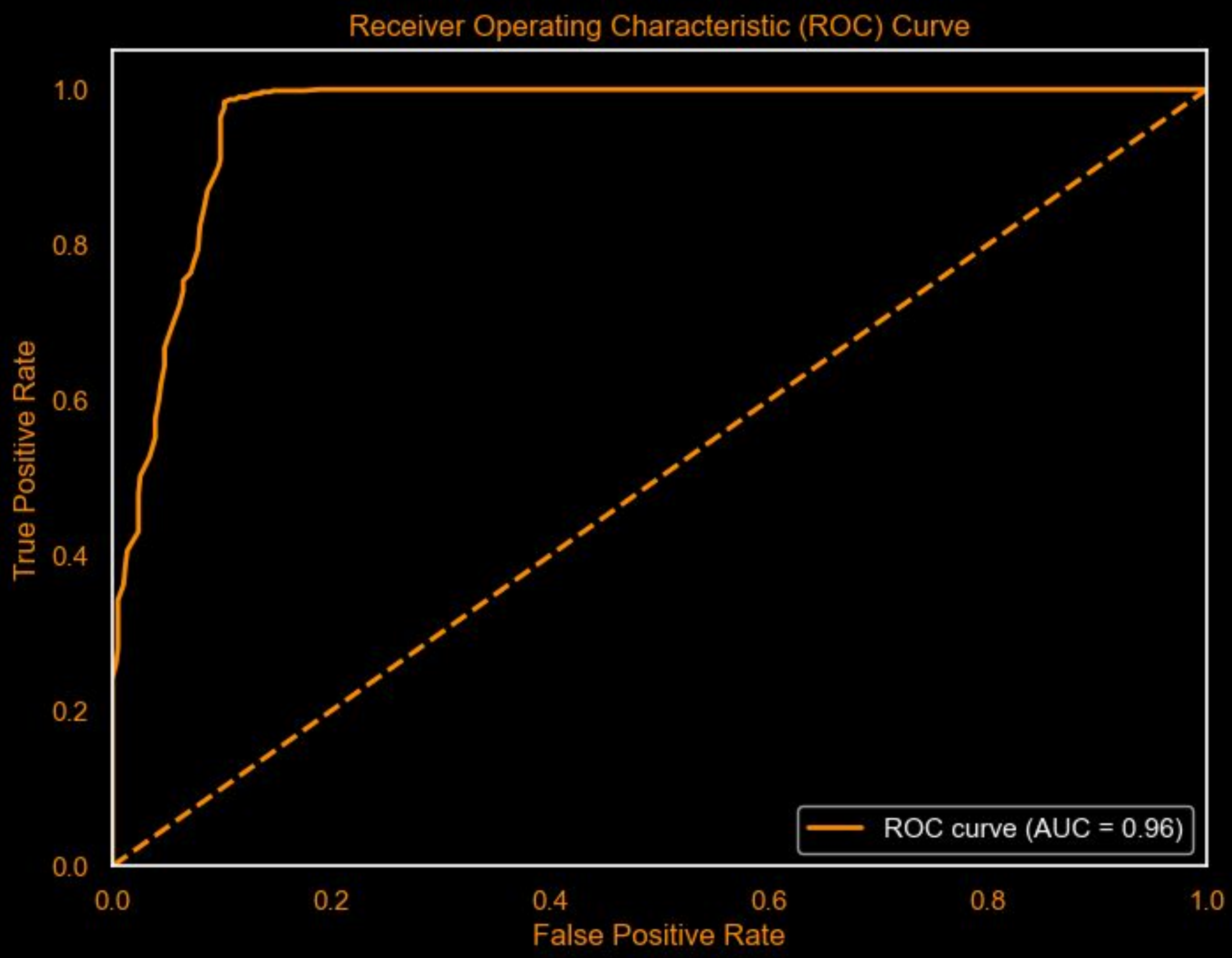
# Analysis :

572 instances of "Serious Injury" correctly predicted.

536 instances accurately identified as "Slight Injury."

51 instances incorrectly predicted as "Serious Injury" for "Slight Injury."

41 instances of "Serious Injury" missed by the model.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.96)

# Analysis :

The AUC score of 0.96 highlights the model's outstanding capability in distinguishing between "Slight Injury" and "Serious Injury" instances.

With a high AUC value of 0.96, the model consistently prioritizes a randomly selected "Serious Injury" instance over a "Slight Injury" instance around 96% of the time.

This AUC score underscores the model's reliability in effectively differentiating between different accident severity levels, showcasing its strong predictive performance

## Model_6(Feature Selection and ensemble methods):

for our last model we decided to improve on model_5 by employing feature selection...we used correlation to find the top 10 columns with high absolute correlation to 'Accident_severity

# Analysis :

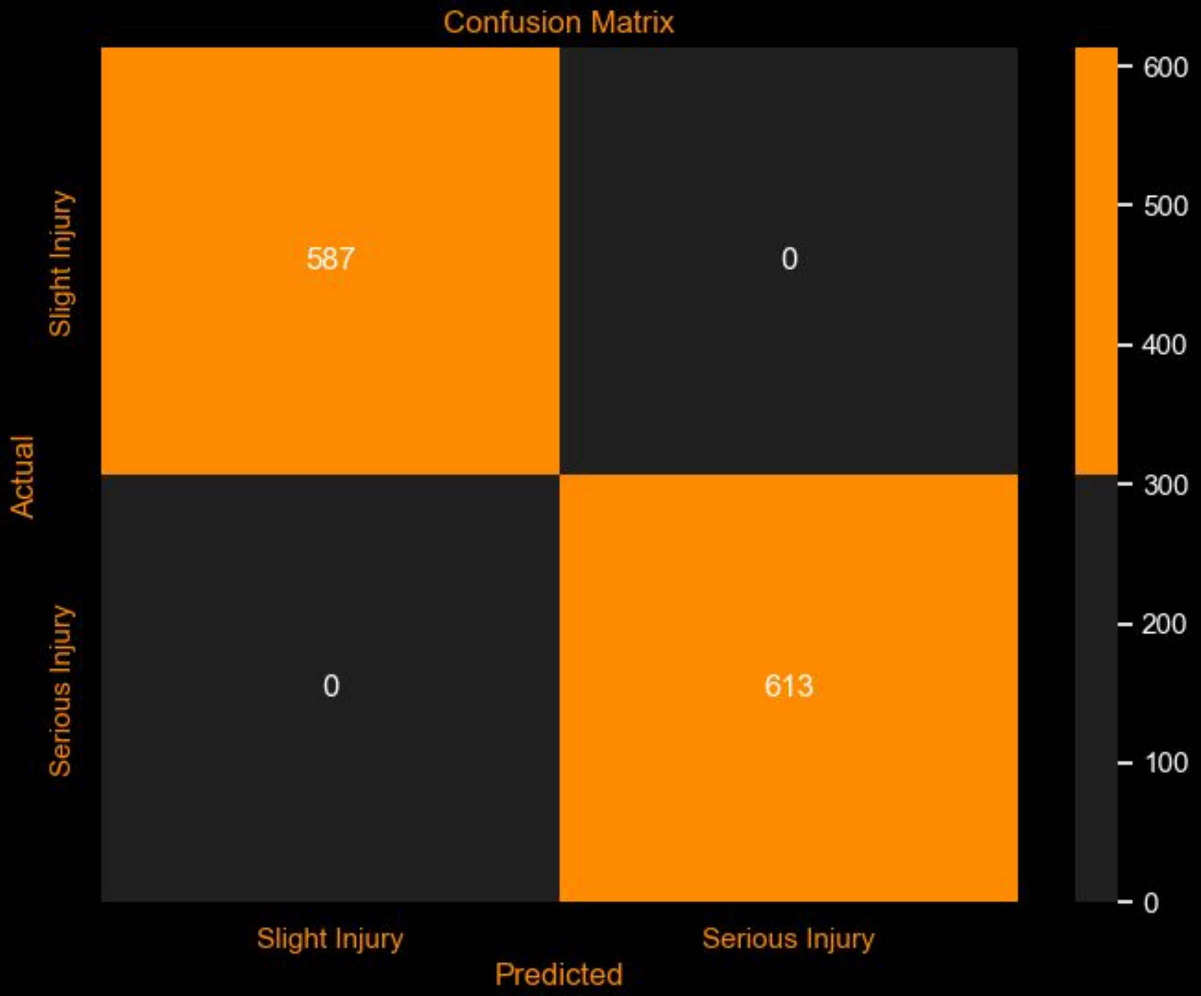
Precision:Both classes ("Slight Injury" and "Serious Injury") achieve a perfect precision score of 1.00, ensuring accurate positive predictions.

Recall: The model attains a flawless recall of 1.00 for both classes, correctly identifying all instances.

F1-Score: With an impeccable F1-score of 1.00 for both classes, the model maintains an ideal balance between precision and recall.

Accuracy:The model's overall accuracy is a perfect 1.00, reflecting correct classification of all instances.

Macro Avg and Weighted Avg: The macro-average and weighted-average F1-scores, both at 1.00, indicate consistent excellence across classes and their distribution.


Confusion Matrix
Actual
Slight Injury
Serious Injury
587
0
0
613
Predicted
600
500
400
300
200
100
0

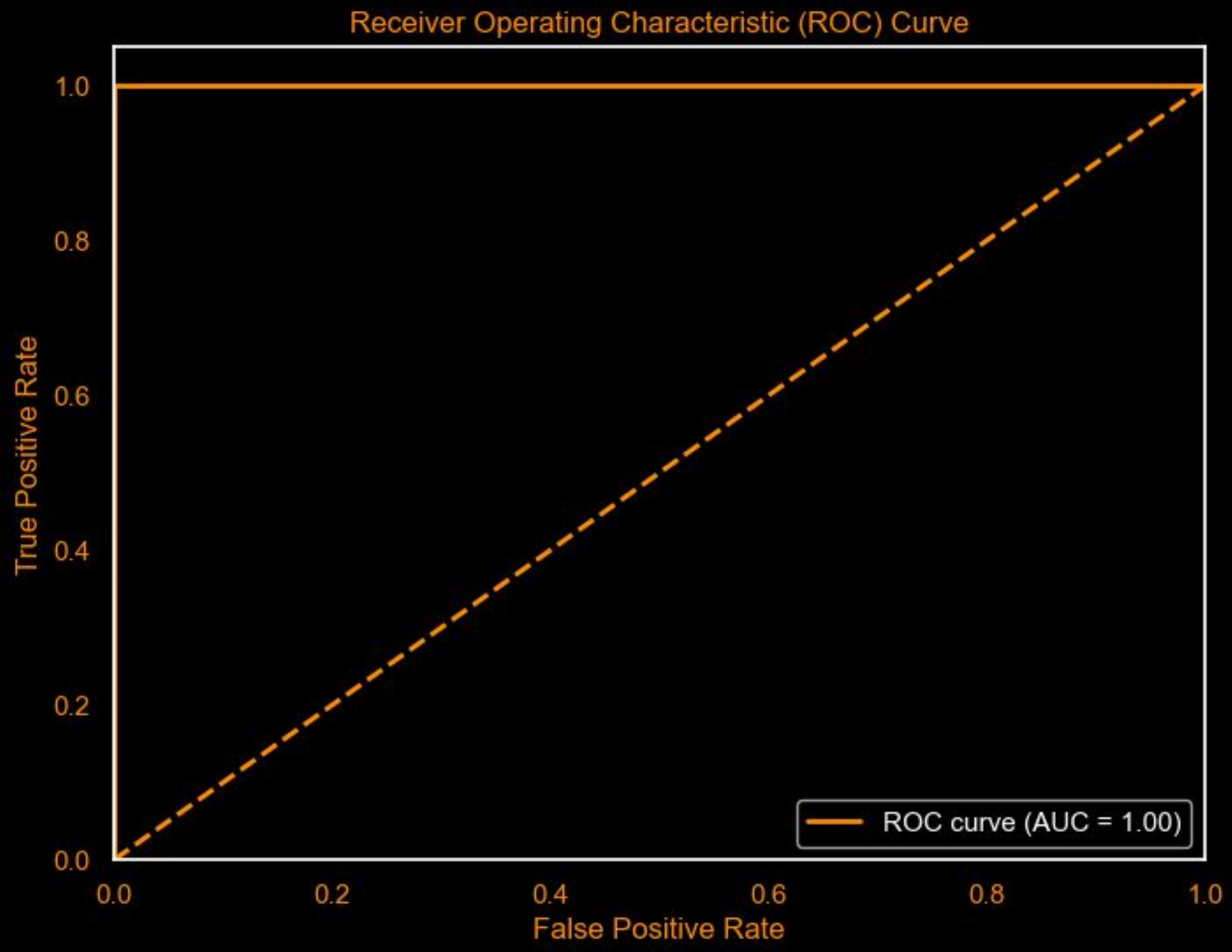
## Analysis:

- True Positives (TP): 613 instances of "Serious Injury" correctly predicted out of 613.

- True Negatives (TN): 587 instances of "Slight Injury" accurately identified out of 587.

- False Positives (FP): No instances wrongly predicted as "Serious Injury."

- False Negatives (FN): No instances missed predicting "Serious Injury."


Receiver Operating Characteristic (ROC) Curve
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (AUC = 1.00)

## Analysis :

An AUC (Area Under the Curve) score of 1 indicates that the model's ROC curve perfectly separates the classes, achieving a flawless distinction between "Slight Injury" and "Serious Injury" instances. This implies that the model has a 100% true positive rate and 100% true negative rate, making it a near-perfect classifier.

# Summary:

Our project predicted "Accident_severity" in Addis Ababa using advanced machine learning techniques. Key features like "Day_of_week," "Hour," and "Area_accident_occured" significantly influenced predictions. Models like Logistic Regression, K-Nearest Neighbors, Decision Trees, and Random Forests showed varying accuracies, with Random Forest reaching 0.92 and our final model achieving 1. Practical implications include optimized urban planning, efficient resource allocation, and targeted awareness campaigns for road safety. Our project advances accident severity prediction and promotes safer roads through data-driven insights.

# Recommendations to Addis Ababa Urban Planning Institute:

1. **Day_of_week and Hour Insights:** Analyze accidents based on "Day_of_week" and "Hour" to identify high-risk periods. Devise urban development and traffic strategies accordingly.

2. **Area_accident_occured and Road_allignment:** Utilize these features to pinpoint accident-prone areas. Enhance infrastructure and optimize road layout to ensure safer travel.

3. **Driving Experience and Educational Level:** Foster collaboration with educational institutions to enhance driver education, with a focus on "Driving_experience" and "Educational_level." Promote safer driving practices.

# Recommendations to Addis Ababa Emergency Services:

1. **Accident_severity and Casualties/Vehicles:** Allocate resources based on severity and predicted casualties. Prioritize high-severity incidents for efficient emergency response.

2. **Type of Vehicle and Owner:** Collaborate with manufacturers and vehicle owners to enhance vehicle safety standards and maintenance practices.

3. **Location Data and Time Patterns:** Utilize location-based insights and time patterns for optimized deployment of emergency services during peak accident times.

## Recommendations to Addis Ababa Road Safety Authority:

1. **Weather_conditions, Light_conditions, and Type_of_collision:** Formulate safety guidelines addressing adverse weather and collision scenarios.

2. **Pedestrian Movement and Lanes:** Improve pedestrian safety and optimize road design for safer crossings.

3. **Vehicle Driver Relation, Sex, and Cause of Accident:** Create targeted awareness campaigns promoting responsible driving behaviors.