Road Traffic Accidents - Final Report

Yarin Yerushalmi Levi, Ken Yaggel, Tal Amar, Idan Hershkovitz

June 2023

1 Introduction

The increasing prevalence of road traffic accidents (RTAs) poses significant challenges for both insurance companies and national road safety authorities. In order to effectively mitigate risks and provide better services, it will be helpfull to find the causes which influence accident severity. This project aims to leverage machine learning techniques to analyze a comprehensive dataset obtained from Kaggle.

The primary objective of this project is to develop predictive model that can determine the severity of accidents. By examining various classification models, we seek to identify meaningful relationships within the dataset and derive actionable insights. The outcomes of this project hold substantial implications for insurance companies, enabling them to adapt their insurance policies to better suit individual customers. Additionally, the national department for road safety authority can leverage these findings to identify key risk factors that contribute to more severe accidents, ultimately enhancing road safety measures.

By understanding the underlying patterns and factors associated with accident severity, insurance companies can tailor their policies to the specific needs of customers, resulting in more accurate premium pricing and personalized coverage. On the other hand, the national department for road safety authority can utilize these insights to implement targeted interventions, such as improved infrastructure, educational campaigns, or stricter enforcement measures, to mitigate the risks that lead to more severe accidents.

2 The Data

The dataset used in this project, titled "Road Traffic Accidents," is a valuable collection of records obtained from the Addis Ababa Subcity Police Departments for a Master's research work. The dataset focuses on road traffic accidents that occurred between the years 2017 and 2020. It has been carefully prepared from manual records, ensuring that all sensitive information has been excluded during data encoding.

Key Characteristics:

Size and Structure: The dataset consists of 12,316 instances or records, providing a significant amount of data to analyze. Each instance represents a specific road traffic accident that was recorded by the Addis Ababa Subcity Police Departments.

Features: The dataset is comprised of 32 features, or attributes, that provide insights into various aspects of the accidents. Some of the features include temporal information such as "Time" or the time of day when the accident occurred. Additionally, it includes features related to driver characteristics such as "Age_band_of_driver" and "Sex_of_driver." Other features encompass details about the vehicles involved, road conditions, environmental factors, and casualty information.

Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehicle_driver_relation	Driving_experience	Type_of_vehicle	Owner_of_vehicle	Service_year_of_vehicle
13:55:00	Sunday	18-30	Male	Junior high school	Employee	5-10yr	Other	Owner	2-5yrs

...	Vehicle_movement	Casualty_class	Sex_of_casualty	Age_band_of_casualty	Casualty_severity	Work_of_casuality	Fitness_of_casuality	Pedestrian_movement	Cause_of_accident	Accident_severity
...	Stopping	Pedestrian	Female	5	3	Driver	Normal	Crossing from nearside - masked by parked or s...	Changing lane to the right	Slight Injury

Figure 1: Example of record for our dataset

The imbalanced nature of the target variable, Accident_severity, where there is a lower number of fatal injuries compared to serious or slight injuries, presents a significant challenge in the analysis and modeling process. Imbalanced datasets can potentially lead to biased models that favor the majority class and provide poor predictions for the minority classes.

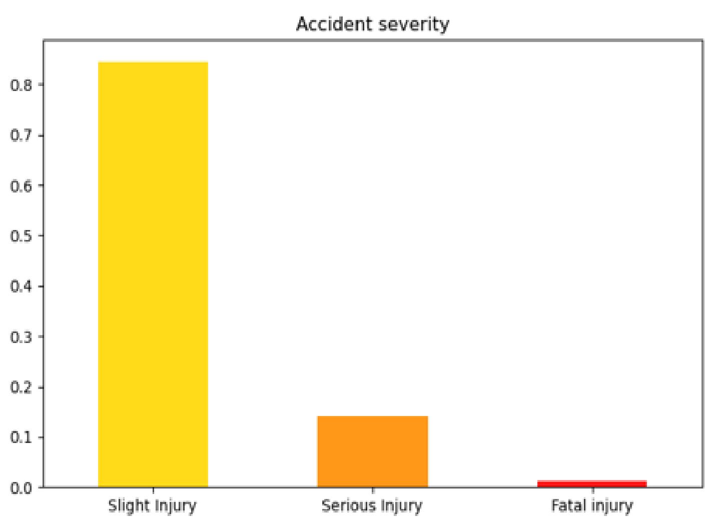

Figure 2: Distribution of target feature

In order to narrow the focus of our research and concentrate on the factors that can be controlled prior to an accident, we made the decision to exclude any features related to casualties from our analysis. Our research objective centers around understanding and predicting accident severity based on the contextual factors leading up to the accident.

3 The Project Pipeline

3.1 Exploratory Data Analysis (EDA)

Conduct an in-depth analysis of the dataset to gain a comprehensive understanding of its characteristics. Explore the distributions, correlations, and relationships between different features. Identify any patterns or trends that may provide insights into accident severity and its contributing factors. Visualize the data using plots, charts, or graphs to facilitate understanding and interpretation.

3.2 Data Preprocessing

Perform necessary data cleaning tasks, such as handling missing values, outliers, and duplicates. Encode categorical variables using techniques like one-hot encoding or label encoding. Split the dataset into training and testing sets to ensure unbiased model evaluation.

3.3 Feature Selection and Engineering

Select relevant features based on domain knowledge, EDA, and correlation analysis.Perform feature engineering if needed, such as creating new features from existing ones that may enhance model performance. Consider techniques like dimensionality reduction to reduce the number of features without losing important information.

3.4 Handling Class Imbalance

Address the class imbalance issue in the target variable, Accident_severity. Apply appropriate resampling techniques like oversampling or undersampling to balance the classes. Alternatively, use class weighting or ensemble methods to handle class imbalance during model training.

3.5 Model Selection and Training

Choose appropriate classification algorithms that are suitable for the problem at hand, such as XGboost, decision trees, random forests. Split the training set further into training and validation subsets for model selection and hyperparameter tuning. Train the selected models on the training set and evaluate their performance on the validation set.

3.6 Model Evaluation:

Assess the trained models using suitable evaluation metrics for imbalanced datasets, such as precision, recall, F1-score, or AUC-ROC. Compare the performance of different models and select the one that exhibits the best overall performance, considering accuracy and performance on all classes.

3.7 Explainable AI (XAI) Techniques

Utilize XAI techniques to enhance the interpretability and explainability of the selected model. Apply methods such as feature importance, SHAP (Shapley Additive Explanations) values, or LIME (Local Interpretable Model-Agnostic Explanations) to understand the model's decision-making process. Analyze and visualize the explanations to gain insights into the features and factors that contribute most to the prediction of accident severity.

3.8 Conclusions and Recommendations

Generate insights and recommendations based on the model's predictions and performance. Communicate the findings, suggestions, and recommendations to relevant stakeholders such as insurance companies and road safety authorities.

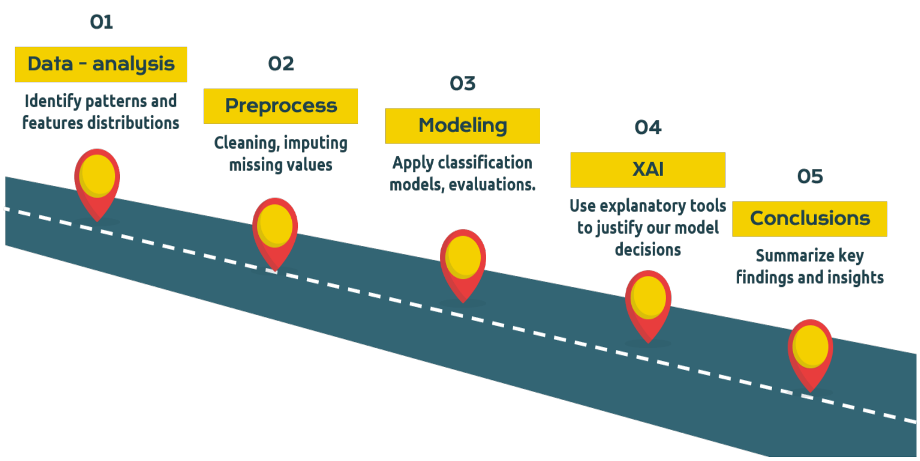

Figure 3: Our project pipeline

4 Data Analysis

In this phase we performed exploratory data analysis (EDA), visualizing the data, and conducting statistical analysis to uncover patterns, relationships, and trends within the dataset. By leveraging these analytical techniques, we aim to gain a deeper understanding of the underlying factors that influence accident severity. We investigated the relationship between different features and accident severity, and the impact of individual factors on the severity of accidents.

Here is our key analysis:

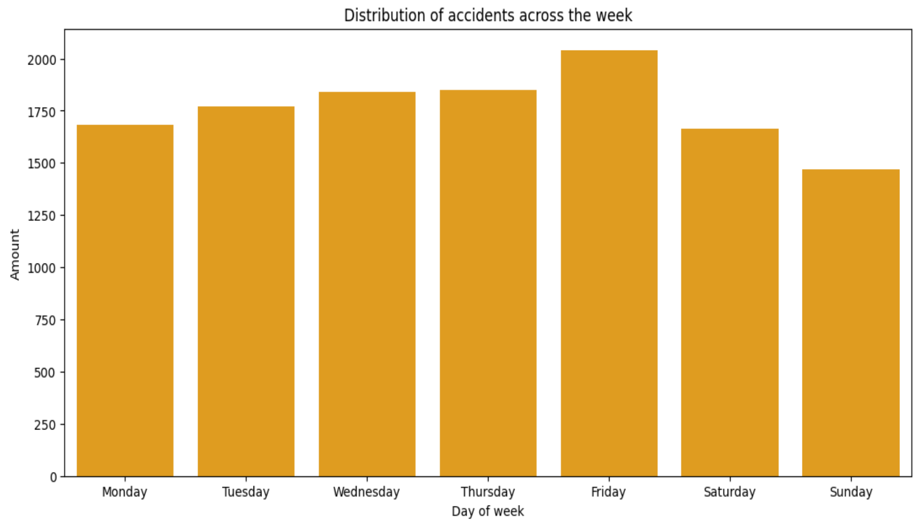

Figure 4: Our project pipeline

Our initial focus in the analysis was to examine the influence of time on the frequency of accidents. Figure 4 provides a clear visualization of this relationship. It reveals that the number of accidents tends to increase as the days progress, peaking on Fridays (which is equivalent to Thursdays for Christians). However, during the weekends, there is a noticeable downward trend in the number of accidents.

This observed pattern aligns with our expectations, considering the typical routines of individuals. It is anticipated that more people engage in driving activities during weekdays for work-related purposes, leading to a higher accident occurrence. Conversely, during the weekends, when people tend to take a break from work and engage in leisure activities, the frequency of accidents decreases.

Moving forward, our investigation focuses on examining the variations in the number of accidents throughout different parts of the day. Figure 5 provides insightful visualizations of these patterns. It is evident that the majority of accidents occur during daylight hours, with a peak around noon, compared to the nighttime hours.

However, a closer analysis reveals an intriguing finding: during the nighttime, there is a higher proportion of serious and fatal accidents. While the overall number of accidents may be lower during this period, the severity of the accidents tends to be more significant.

Continuing our investigation, we further explore the impact of changing light conditions on the severity of accidents. Figure 6 provides a valuable visual

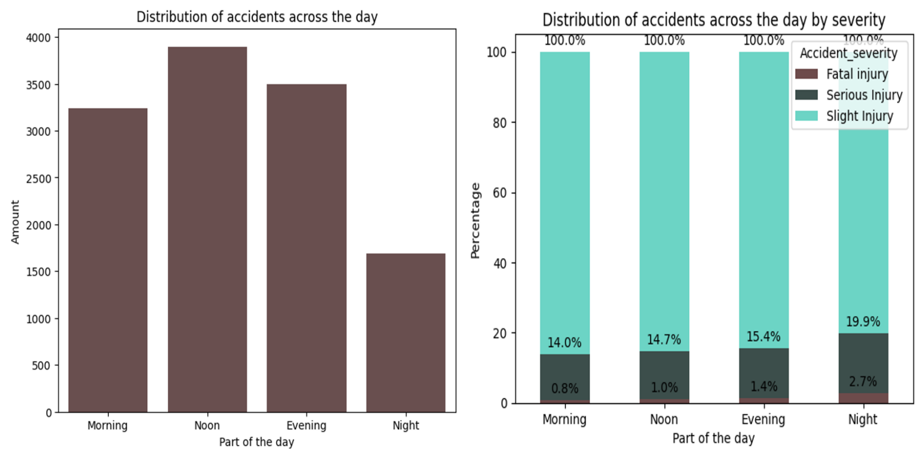

Figure 5: (Left)Distribution of accidents across the day, (Right)Distribution of accidents across the day by severity

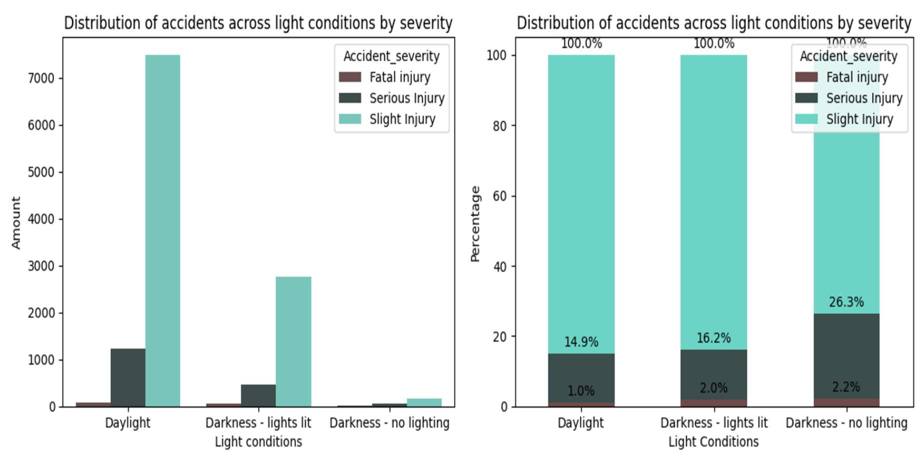

Figure 6: Distribution of accidents across light conditions

representation of this relationship. Notably, as the light conditions become darker (due to the decrease in daylight or transitioning into nighttime), we observe a decrease in the overall number of accidents. This decrease can be attributed to fewer instances of driving during darker periods.

However, a significant insight emerges when we analyze the severity of accidents in relation to light conditions. Despite the reduction in the total number of accidents, we find that in darker conditions, the proportion of serious and fatal accidents is the highest. This finding highlights the heightened risk and potential severity of accidents when visibility is limited or compromised.

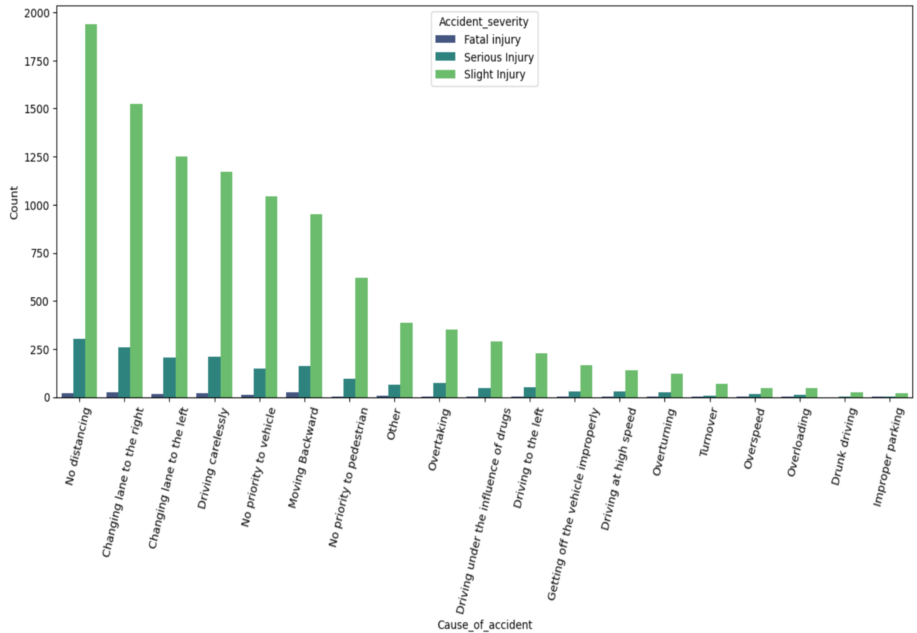

Figure 7: Distribution of accidents across cause of accident by severity

An additional intriguing feature we sought to investigate is the cause of accidents and the specific areas we should focus on. Our analysis reveals, as shown in figure 7, a consistent pattern across all severity levels, indicating that the most prevalent causes of accidents are "No distancing" and "Changing lanes carelessly."

This finding underscores the significance of these factors in contributing to accidents of varying severity. Understanding these prominent causes enables us to prioritize efforts and allocate resources towards addressing these specific areas. By targeting interventions and implementing preventive measures that address "No distancing" and "Changing lanes carelessly," we can effectively reduce the occurrence of accidents and mitigate their potential severity.

Lastly we investigated the change in the distribution of "Number of vehicles involved" across different levels of severity. Our results reported at table 1.

Table 1: Number of vehicles involved statistics

Accident_severity	**Average**	**Standard Deviation**
Slight Injury	2.07	0.67
Serious Injury	1.90	0.78
Fatal Injury	1.80	0.55

As we see, the most dangerous accidents are involving less cars on average.

5 Preprocessing

In order to prepare our dataset for modeling, several preprocessing steps need to be performed. These steps involve handling missing values, encoding categorical variables, and splitting the dataset into training and testing sets. Here's an outline of the preprocessing section:

5.1 Handling Missing Values

We identify the features that contain missing values and analyze the nature of the missing values and determine the appropriate strategy for imputation. For most of the categorical features, where there was not any hidden logic we could have used, we impute missing values with the most common value, which is the mode of that feature. In other cases, we were able to make more suitable choice, like for example when we impute the driving experience in years, we used the most common value among the relevant age group. In cases where missing values cannot be imputed logically or statistically, set them as "Unknown" to indicate their absence or unavailability.

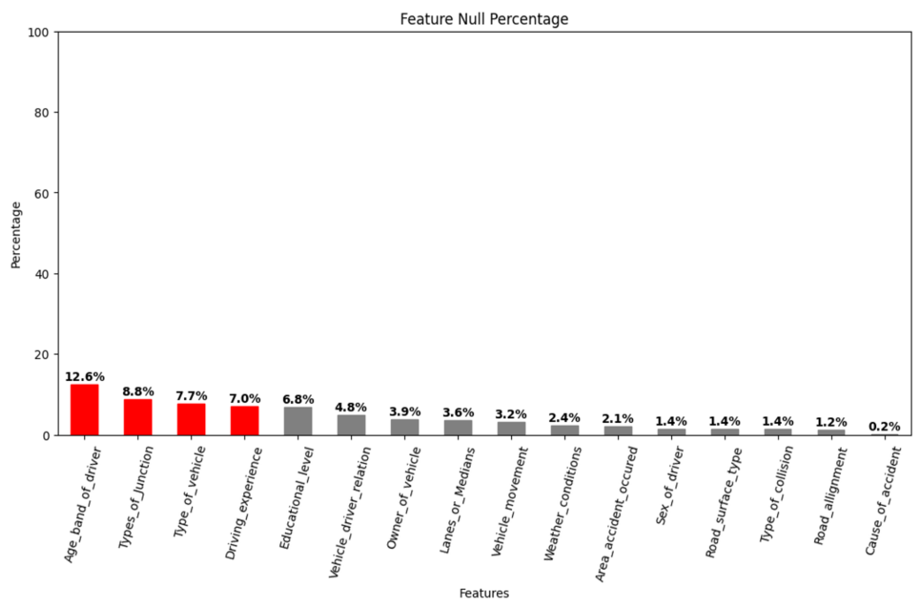

Figure 8: Feature Null Distribution

5.2 Encoding Categorical Variables

We identify the categorical features that require encoding and used two techniques. For the ordinal features such 'Educational_level', we normalize the rank value of each category. On the other hand, for the rest of the categorical features we used one-hot encoding to make it suitable for our machine learning model.

6 Feature Selection

In order to enhance the effectiveness and efficiency of our classification models in predicting accident severity, we performed feature selection. we initially attempted to select numeric features based on their correlation scores. However, upon conducting the correlation test, we obtained poor correlation scores for the selected numeric features. The low correlation scores indicate a weak linear relationship between the numeric features and the target variable, accident severity. Thus, in this project, we employed feature selection with Chi-square test instead

The Chi-square test is a statistical method used to determine the dependency between categorical variables. We applied the Chi-square test to evaluate the relationship between each feature and the target variable, accident severity. This test assesses whether there is a significant association between a categorical feature and the target variable, considering the frequency distribution of different categories within each feature.

The feature selected will be utilized in our classification model.

7 Dealing with the imbalances of the target feature

Addressing the imbalances within the target feature, namely the disparity in the distribution of accident severity categories, is a crucial aspect of our project. The target feature, Accident_severity, exhibits a significant imbalance with a higher frequency of fatal injuries compared to serious or slight injuries. This imbalance poses challenges for accurate model training and prediction. To overcome this issue, we employ various techniques such as undersampling, class weight adjustment, and ensemble methods to balance the distribution of the target feature. By mitigating the impact of class imbalance, we aim to develop models that can effectively capture and predict different levels of accident severity.

In imbalanced datasets, where one class is significantly more prevalent than the others (such as in the case of accident severity where there are more instances of slight injuries compared to serious or fatal injuries), accuracy alone can be misleading. A model that simply predicts the majority class for all instances will achieve high accuracy, even though it fails to accurately predict the minority classes.

Balanced accuracy is a metric that provides a fair assessment of a model's performance, particularly in scenarios where the dataset is imbalanced. It calculates the average accuracy for each class, considering both the minority and majority classes, and provides a balanced representation of overall performance.

Balanced accuracy takes into account the varying class distributions by considering the accuracy of each class individually and then averaging them. It provides a more meaningful evaluation of a model's performance by giving equal importance to each class, regardless of its prevalence. This allows us to assess how well the model performs in predicting all classes, including the minority ones.

8 Modeling

In the modeling phase of our project, we employ a variety of classification algorithms to predict accident severity based on the selected features. Through rigorous evaluation and comparison using metrics like accuracy, precision, recall, and F1-score, we assess the models' performance and determine the most accurate predictors of severity. In order to select the model which we will evaluate we used Pycaret library and taked the top models based on 'Balanced Accuracy'.

Table 2: Pycaret models basic evaluation sorted by 'Balanced Accuracy'

Model	**Accuracy**	**AUC**	**Recall**	**Prec.**	**F1**	**Kappa**	**MCC**	**Balanced Accuracy**
Decision Tree Classifier	0.7494	0.5461	0.7494	0.7586	0.7539	**0.0889**	0.0891	**0.3897**
Extreme Gradient Boosting	0.8399	0.6267	0.8399	0.7802	**0.7909**	0.0886	**0.1255**	0.3575
Light Gradient Boosting Machine	0.8438	**0.6338**	0.8438	0.783	0.7883	0.0702	0.1187	0.3489
Gradient Boosting Classifier	0.844	0.6206	0.844	0.7826	0.7826	0.0423	0.0877	0.3453
Naive Bayes	0.0758	0.5249	0.0758	0.7797	0.1092	0.0041	0.0138	0.3429
K Neighbors Classifier	0.8347	0.578	0.8347	0.7565	0.7803	0.0378	0.0572	0.3421
Random Forest Classifier	**0.8469**	0.6269	**0.8469**	**0.811**	0.7799	0.0239	0.0871	0.3384

We will examine each of the listed models separately despite Naive Bayes and K Neighbors Classifier,which reached relatively worse results.

8.1 Decision Tree Classifier

We used sklearn.tree.DecisionTreeClassifier and optimize the performance of the model, with 'RandomizedSearchCV' technique in the search for the best hyperparameters. By randomly sampling from the hyperparameter space and performing cross-validation, we identified the optimal combination of hyperparameters. Through this process, we determined that setting 'min_samples_split' to 5 and 'max_depth' to 9 yielded the best results in terms of achieving a balanced accuracy metric.

8.2 XGBoost

For the XGBoost model, Using the 'RandomizedSearchCV' technique, we searched for the best combination of hyperparameters. After the search, we found that setting the number of estimators ('n_estimators') to 150, maximum depth ('max_depth') to 6, and learning rate ('learning_rate') to 0.050 provided the best results. These hyperparameters were chosen based on their ability to improve the model's performance in terms of accuracy and predictive power.

8.3 Light Gradient Boosting Machine

In our analysis, we also employed the Light Gradient Boosting Machine (LightGBM) algorithm to predict accident severity. Similar to the previous models, we utilized the 'RandomizedSearchCV' technique to search for the best hyperparameters for LightGBM. After the search, we identified that setting the number of estimators ('n_estimators') to 90, maximum depth ('max_depth') to 7, and learning rate ('learning_rate') to 0.06 yielded the optimal results.

8.4 Gradient Boosting Classifier

Moreover, we utilized the Gradient Boosting Classifier algorithm to predict accident severity. To optimize the model's performance, we employed the 'RandomizedSearchCV' technique to search for the best combination of hyperparameters. After the search, we identified that setting the number of estimators ('n_estimators') to 80, maximum depth ('max_depth') to 4, and learning rate ('learning_rate') to 0.07 provided the optimal results.

8.5 Random Forest Classifier

Our final model for predicting accident severity was the Random Forest Classifier algorithm. Using the 'RandomizedSearchCV' technique, we extensively searched for the best combination of hyperparameters to optimize the model's performance. The search led us to determine that setting the number of estimators ('n_estimators') to 190, minimum samples split ('min_samples_split') to 10, and maximum depth ('max_depth') to 6 yielded the most optimal results.

9 Evaluation

9.1 Train-Test Split

To assess the performance of our models, we divided the preprocessed dataset into training and testing subsets. Crucially, we ensured that the distribution of classes within the target feature was maintained in both sets. By using the target feature to guide the splitting process, we aimed to replicate the distribution of accident severity classes in both the training and testing datasets. This approach allowed us to evaluate the models' performance reliably.

9.2 Models Performance

For evaluating the performance of our models, we initially employed metrics such as balanced accuracy, Weighted recall, Weighted precision, and Weighted F1-score. These metrics provided insights into the overall model performance and its ability to handle class imbalances (Table 3). Additionally, we conducted a hyperparameter analysis to examine the impact of each hyperparameter on the balanced accuracy while keeping other parameters constant. This analysis helped us understand the sensitivity of the models to different hyperparameter settings and identify the optimal configuration.

By utilizing balanced accuracy as an evaluation metric, we can obtain a more accurate and unbiased assessment of the model's performance on imbalanced data, helping us make informed decisions and identify models that are effective in handling class imbalances.

Furthermore, we conducted a cross-class examination to assess the models' ability to accurately detect and classify severe accident cases, specifically Fatal and Serious ones, in comparison to the classification of Slight accidents (Table 4). This analysis aimed to identify which model exhibited superior performance in accurately identifying and distinguishing between different levels of accident severity. By conducting these evaluations and comparisons, we gained valuable insights into the strengths and weaknesses of each model and identified the model with the finest ability to detect and classify severe accident cases.

Table 3: Models performances assesment

Model	**Balanced Accuracy**	**Accuracy**	**Weighted Recall**	**Weighted Prec.**	**Weighted F1**
Decision Tree Classifier	0.4996	0.5175	0.52	0.78	0.61
Extreme Gradient Boosting	0.5459	0.6027	0.60	0.78	0.67
Light Gradient Boosting Machine	0.5302	0.5958	0.60	0.78	0.66
Gradient Boosting Classifier	**0.5582**	0.6031	0.60	**0.79**	0.67
Random Forest Classifier	0.4698	**0.6798**	**0.68**	0.77	**0.72**

Upon analyzing the results presented in Table 4, it is evident that none of our models achieved satisfactory performance in accurately detecting 'Fatal injury. As anticipated, the imbalanced nature of the dataset posed a significant

Table 4: Cross-class examination

Model	**Slight injury**			**Serious injury**			**Fatal injury**		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
Decision Tree Classifier	**0.89**	0.54	0.67	0.20	0.38	0.26	0.03	0.58	0.06
Extreme Gradient Boosting	0.89	0.63	0.74	0.23	0.42	0.30	**0.06**	0.58	**0.10**
Light Gradient Boosting Machine	0.88	0.63	0.73	0.22	0.42	0.29	0.05	0.55	0.09
Gradient Boosting Classifier	**0.89**	0.63	0.74	0.24	**0.43**	**0.31**	0.05	**0.61**	0.09
Random Forest Classifier	0.87	**0.75**	**0.81**	**0.25**	0.27	0.26	0.04	0.39	0.07

challenge in accurately predicting the minority class. Despite our best efforts to address the class imbalance through preprocessing techniques and model selection, the inherent difficulty of the task hindered our models from achieving the desired level of accuracy for 'Fatal injury' detection.

9.2.1 Decision Tree Classifier

Our implementation of the Decision Tree Classifier achieved an average performance across various metrics, compere to the models tested. As depicted in Figure 9, we observed that the model's performance reached a relative optimum, indicating that further modifications to the listed hyperparameters would not significantly enhance its performance. The classifeir reached the best precision for 'Slight injury' while the worst for the rest of the classes.

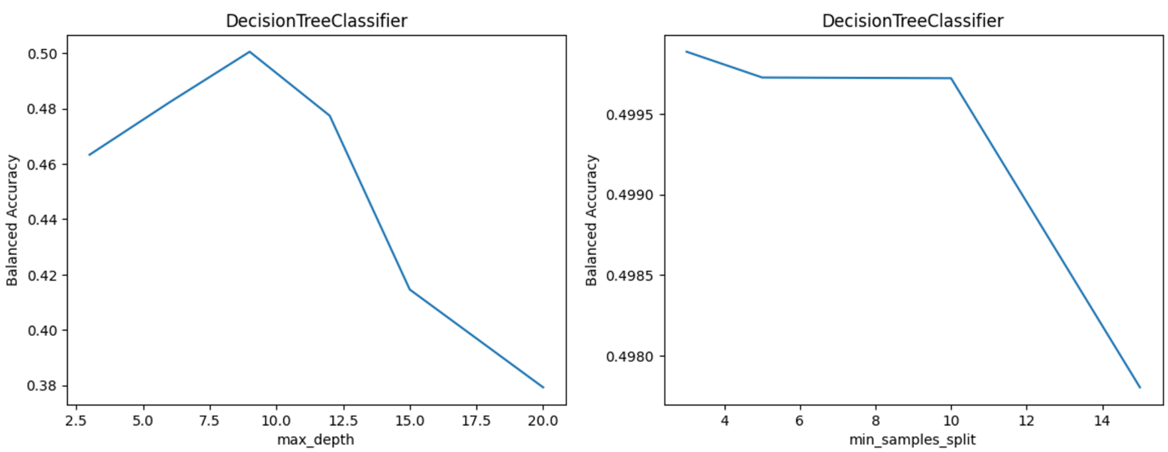

Figure 9: Decision Tree hyperparameters modifications

9.2.2 XGBoost

our implementation of xgboost achieved relatively good balanced accuracy and seems to have the best precision over the fatal class, that we consider as the most important one for our case. Moreover, as we can see from figure 10, optimizing the hyperparameters of the xgboost seems to produce relatively significant benefits in terms of balanced accuracy.

9.2.3 Light Gradient Boosting Machine

Our implementation of the Light Gradient Boosting Machine yielded subpar outcomes, failing to achieve the top position in any of the evaluation metrics we examined. We utilized this model for a comparative analysis against other more successful models, as shown in Table 4.

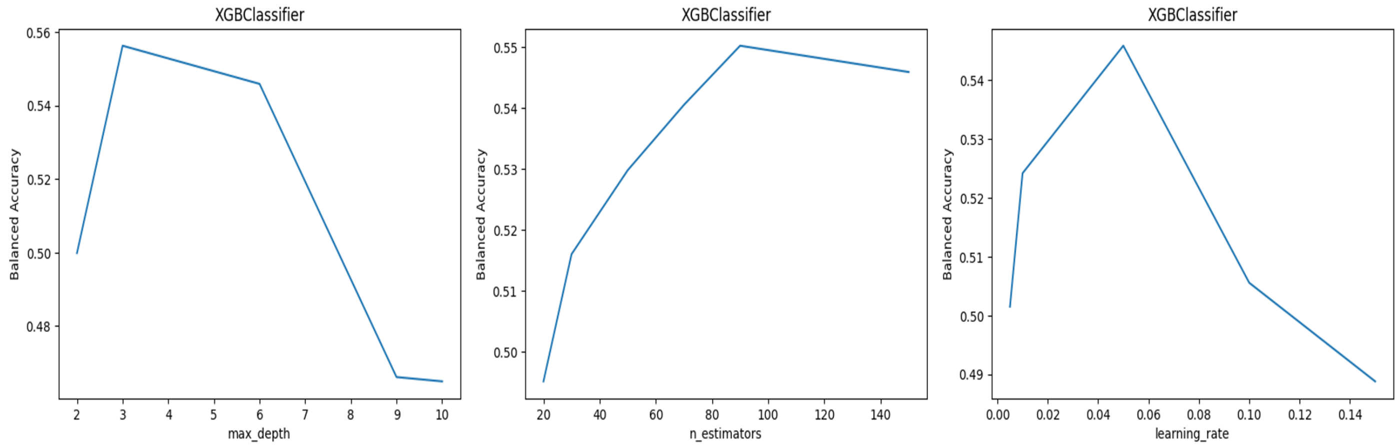

Figure 10: XGBoost hyperparameters modifications

Figure 11 demonstrates the optimal results achieved by the model. Notably, we observed an increase in balanced accuracy as max_depth exceeded 12. However, we need to avoid overfitting, making it impractical to pursue a higher max_depth. Although this model exhibited a significantly faster runtime compared to the other models tested, it is worth noting that all models remained well within the maximum runtime allowed for training. Consequently, the runtime difference becomes irrelevant and does not significantly impact the choice of model.

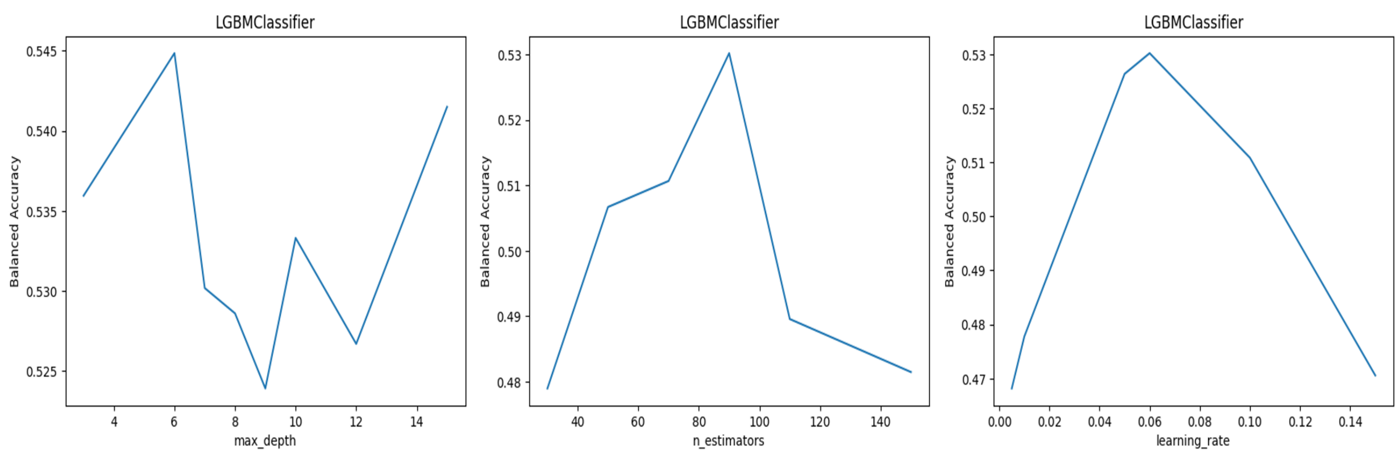

Figure 11: Light Gradient Boosting hyperparameters modifications

9.2.4 Gradient Boosting Classifier

Our implementation of the Gradient Boosting Classifier yielded favorable outcomes. Upon reviewing Table 4, it is evident that the model exhibited high precision scores across all levels of injury prediction. Additionally, the model achieved the highest balanced accuracy, which is a crucial metric given the significant class imbalance in our data.

Moreover, Figure 12 showcases that we attained optimal results by selecting appropriate hyperparameters. Further optimization of the hyperparameters would not lead to improvements.

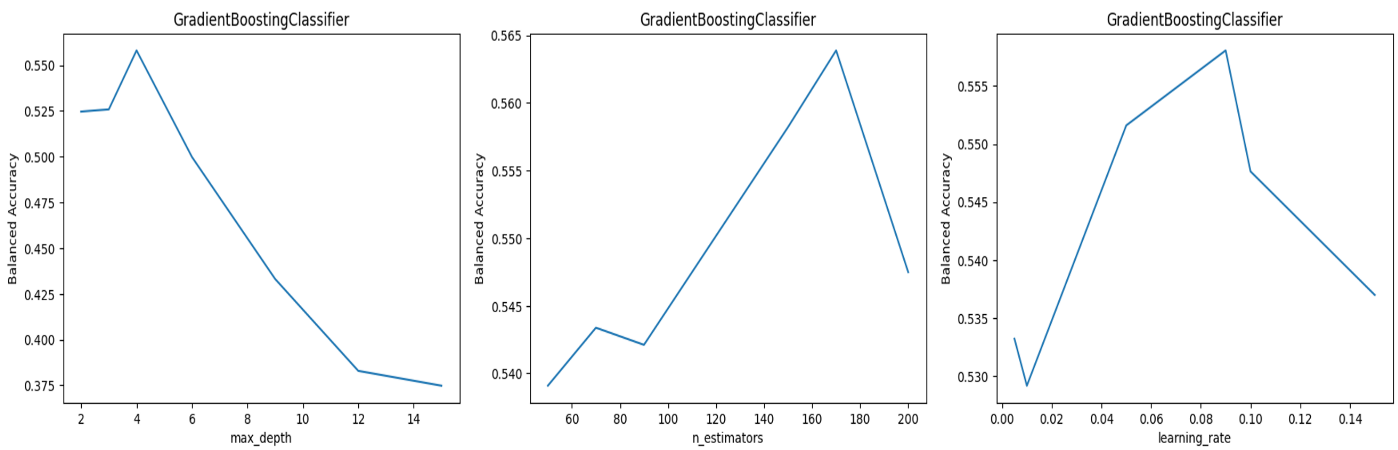

Figure 12: Gradient Boosting hyperparameters modifications

9.2.5 Random Forest Classifier

Although random forest achieved good results over the slight injury class (table 4) it is lacking in terms of the prediction on the fatal class. For our case the fatal injury class as well as the balanced accuracy is crucial because of the grave weight that such incidents hold.

10 Advance Models Analysis

In this section of Advanced Models Analysis, we delve deeper into the evaluation and interpretation of our models' performance. We utilize three powerful algorithms, namely XGBoost, Light Gradient Boosting Machine, and Gradient Boosting Classifier, as they have shown relatively better results compared to other tested models.

To gain a comprehensive understanding of the models' performance, we employ various techniques. Firstly, we present the confusion matrix, which provides insights into the models' predictive capabilities across different classes of

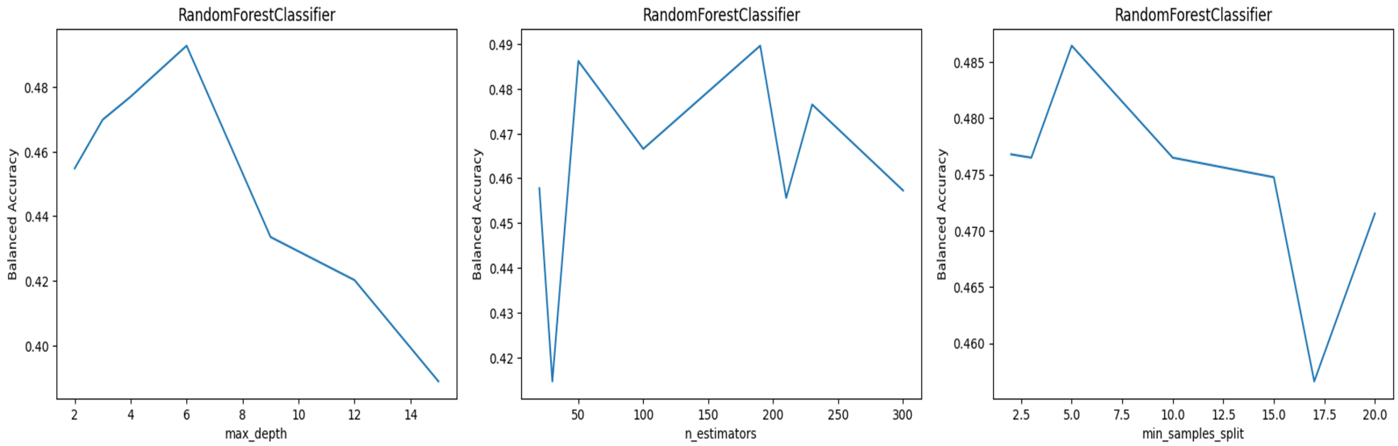

Figure 13: Random Forest hyperparameters modifications

accident severity. Furthermore, we utilize Receiver Operating Characteristic (ROC) curves to evaluate the models' performance in balancing true positive rate and false positive rate across different classification thresholds. in addition, we leverage Explainable Artificial Intelligence (XAI) tools to gain insights into the models' decision-making process. XAI tools provide transparency by highlighting the features that have the most significant impact on the models' predictions.

10.1 XGBoost

Upon analyzing the performance of XGBoost using the confusion matrix (Table 5), we observe that the model accurately classified a majority of the accident severity predictions. However, a closer examination of the precision-recall curve (Figure 14) reveals that the precision for serious and fatal injuries is relatively low. The apparent high performance in the confusion matrix can be attributed to the classification of slight injuries, which exhibits a higher precision.

Furthermore, when considering the ROC curve (Figure 14), we find that the AUC (Area Under the Curve) for the three levels of injuries is greater than 0.5. This indicates that the classification performance is better than random, with the classification of fatal injuries showing particularly promising results.

When examining the SHAP analysis, we gain insights into the influential features for classifying injury severity. Firstly, it is evident that the 'Number_of_vehicles_involved' feature holds the greatest impact across all injury severities. This finding aligns with the intuition that accidents involving more vehicles tend to result in a higher likelihood of injuries, considering the increased number of people involved.

Secondly, the 'Time_bin' feature emerges as significant, indicating that the timing of the accident plays a role in determining the severity of injuries sustained. This suggests that accidents occurring during specific time periods may have distinct effects on the resulting injuries.

Lastly, the 'Types_of_junction_No_junction' feature exhibits notable influence on the classification. Notably, its impact on classifying fatal injuries stands out prominently. This observation is logical since accidents happening on open roads, without junctions, tend to be more hazardous due to typically higher speeds compared to accidents within junctions.

Overall, the SHAP analysis sheds light on the key features contributing to injury severity classification, emphasizing the influence of variables such as the number of vehicles involved, the timing of the accident, and the presence or absence of junctions.

Table 5: XGBoost confusion matrix

		Predicted		
		Slight injury	**Serious injury**	**Fatal injury**
Actual	**Slight injury**	1319	499	266
	Serious injury	159	148	42
	Fatal injury	9	4	18

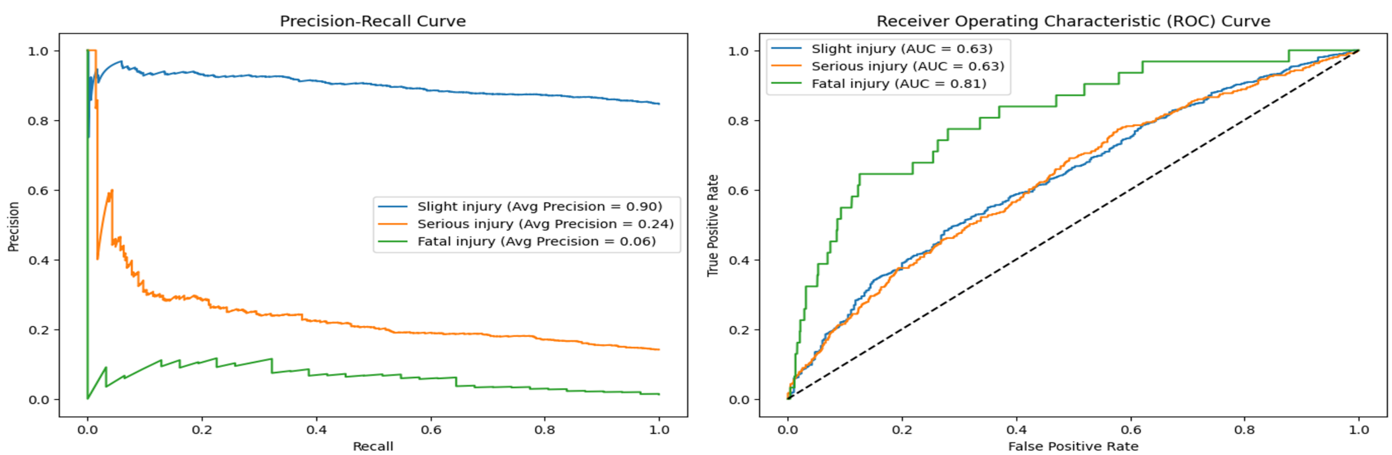

Figure 14: XGBoost Roc and PR REC curves

10.2 Light Gradient Boosting Machine

The SHAP feature importance in this case closely resembles the distribution shown in Figure 15, with the first 10 features appearing in the same order. However, when focusing on the results for the fatal class, the LGB classifier exhibited similar classification patterns to XGBoost (table 5). Although it misclassified 10 observations as slight compared to XGBoost's 9 misclassifications,

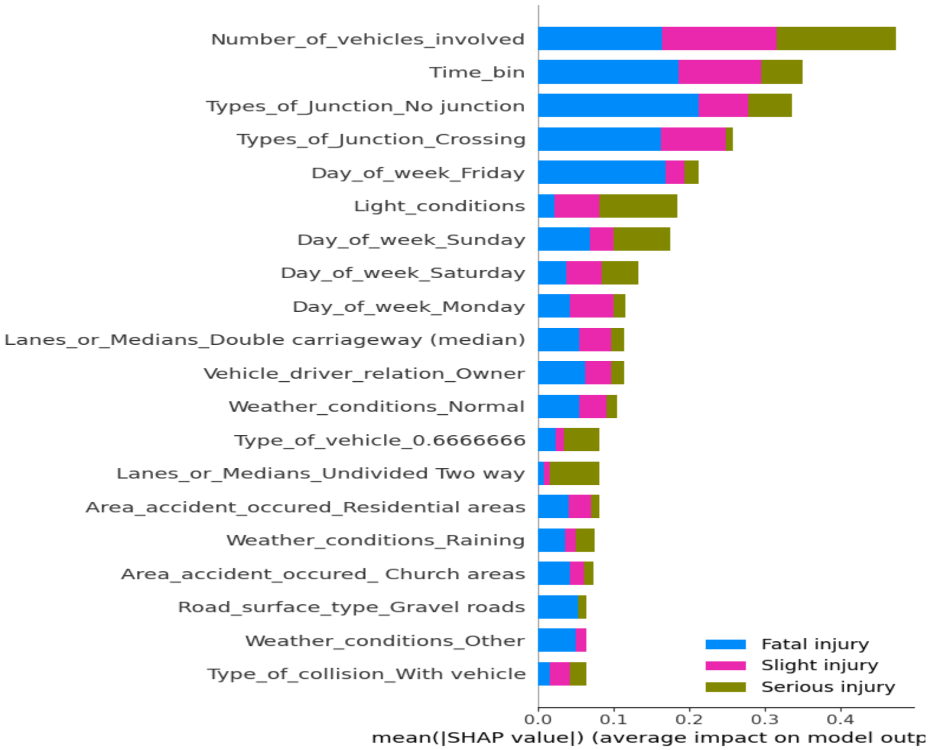

Figure 15: XGBoost SHAP analysis

LGB performed relatively worse in the other two classes.

Interestingly, these results do not strongly reflect in the precision-recall and ROC curves, as LGB actually displayed slightly higher AUC values as compares to the XGBoost (figure 14).

Table 6: Light Gradient Boosting confusion matrix

		Predicted		
		Slight injury	**Serious injury**	**Fatal injury**
Actual	**Slight injury**	1306	511	267
	Serious injury	161	145	43
	Fatal injury	10	4	17

10.3 Gradient Boosting Classifier

The gradient boosting model demonstrates a slightly superior performance when compared to both the XGBoost and light gradient boosting models, particularly in the serious and fatal classes. However, it falls slightly behind the XGBoost

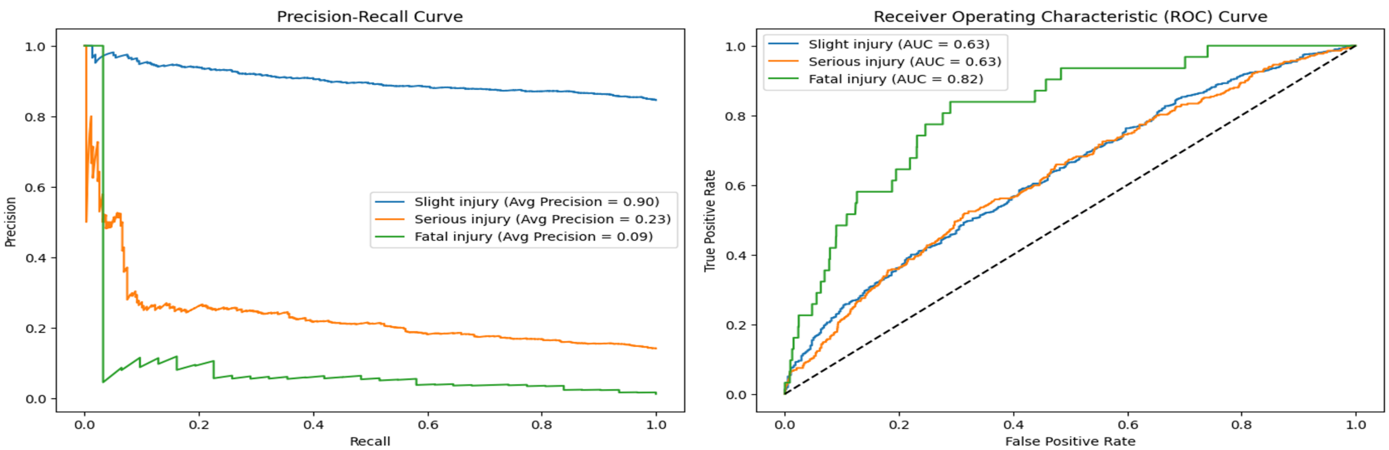

Figure 16: Light Gradient Boosting Roc and PR REC curves

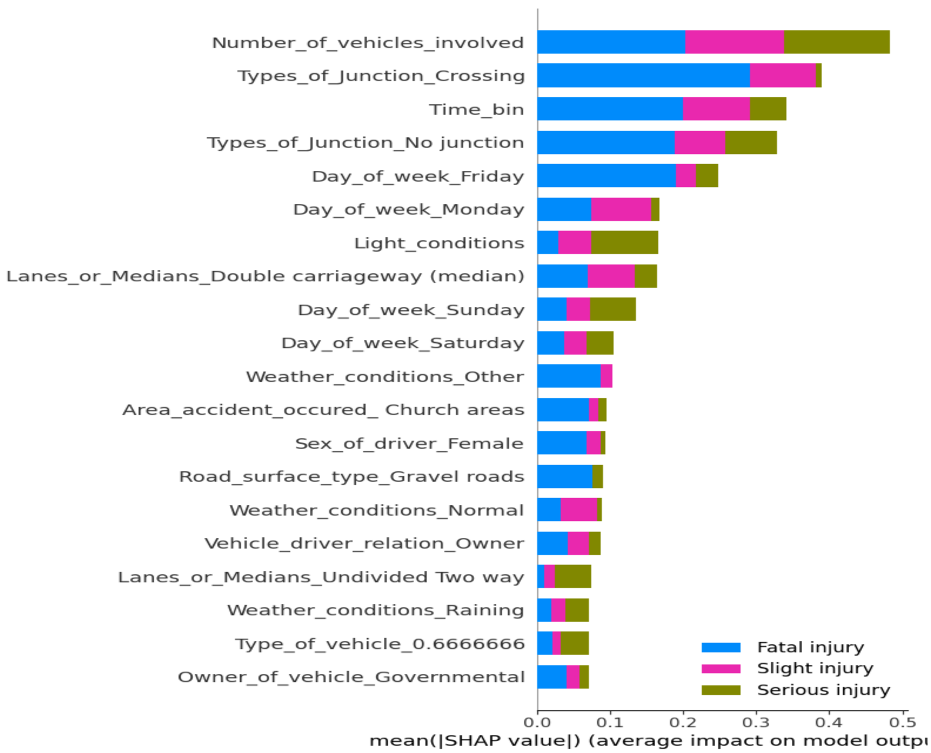

Figure 17: Light Gradient SHAP analysis

model (as shown in Table 5) with 1316 correct classifications compared to the XGBoost's 1319.

This trend is also apparent in the ROC and PR curves, where the gradient boosting model achieves marginally higher AUC scores than the XGBoost (as depicted in Figure 15). Additionally, it is important to note that the SHAP feature importance for the gradient boosting model was not included in the analysis. This omission was due to the limitation of the SHAP package, which currently does not support the calculation of feature importance for multiclass gradient boosting classifiers.

Table 7: Gradient Boosting confusion matrix

		Predicted		
		Slight injury	**Serious injury**	**Fatal injury**
Actual	**Slight injury**	1316	462	306
	Serious injury	152	151	46
	Fatal injury	6	6	19

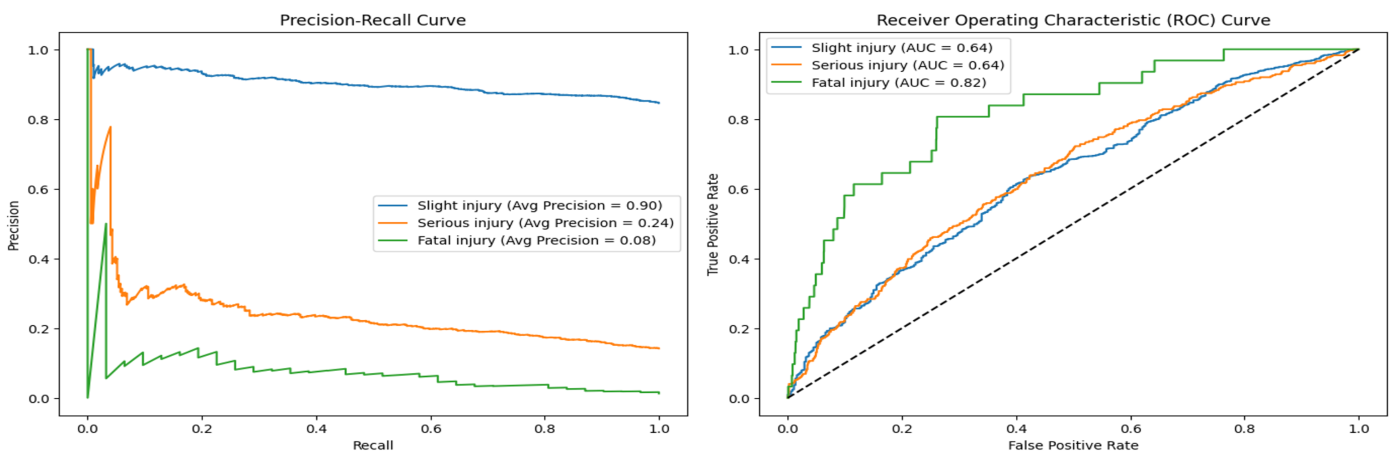

Figure 18: Gradient Boosting Roc and PR REC curves

11 Future work

1. Addressing the Imbalance: To improve the performance of the classification problem, it is crucial to gather a larger number of samples related to fatal accidents. By obtaining more data on fatal accidents, we can enhance the model's ability to accurately detect and predict this class. This additional information will help in gaining deeper insights into the patterns and factors contributing to fatal accidents, leading to more effective preventive measures.

2. Exploring Complex Models: The utilization of more sophisticated models, such as neural networks (NN), can potentially enhance the detection of dangerous driving behaviors and identify critical environmental factors. By leveraging the power of neural networks, we can potentially uncover nuanced insights that might be missed by simpler models.

3. Incorporating Non-Accident Data: To gain a comprehensive understanding of the causes leading to accidents, it is important to include information on drivers who did not encounter accidents. By collecting and incorporating data on drivers who did not experience accidents, we can analyze their driving behavior, environmental conditions, and other factors to identify potential causes and risk factors that contribute to accidents.

12 Conclusions

Overall, our analysis has revealed that the environmental aspects play a significant role in the model's decision-making process. Therefore, we strongly recommend that the national road safety authorities prioritize interventions targeting these factors. On the other hand, we encountered challenges in identifying deterministic patterns related to dangerous drivers, which limited our ability to provide specific suggestions in this regard.

Despite the complexities of the imbalance classification problem, we have put forth considerable effort to identify the appropriate models and methods. We even attempted to train a model specifically for distinguishing slight injuries from all other classes, but unfortunately, we did not achieve significant improvements in performance. Nevertheless, our exploration has led us down various paths and enabled us to gain valuable insights, enhancing our data science expertise and refining our analytical skills.

In conclusion, while the classification problem of imbalanced accident severity presented its challenges, we have made commendable progress in attempting to overcome these hurdles. Our findings emphasize the importance of environmental factors and call for focused attention from road safety authorities. The journey of this project has expanded our understanding and proficiency in data science, setting the stage for further advancements in addressing complex problems within the field of road safety analysis.