Road Accident Factors Analysis

Bhuvan Thirwani, Piyush Gulhane, Harshit Malpani
50565974 50608504 50608809

Abstract

The main aim of the project is to conduct a comprehensive analysis of road accident trends and study the different factors responsible for accidents, with the ultimate goal of improving road safety measures and making informed policy decisions. It focuses on identifying patterns in road accidents and their impact severity, examining the relation between various factors contributing to road accidents. These factors include driver attributes, vehicle conditions, road surface conditions, and environmental factors.

By using data analytics and statistical methods, it tries to establish important correlations between accident severity/cause and the various contributing factors mentioned above. The results from this study will provide valuable insights for intervention in reducing the number of accidents and casualties if an accident occurs .

The project's outcomes will serve as a guiding document for policymakers, transportation authorities, and safety organizations to consider while making future decisions. It helps in designing safety protocols, road design protocols, and more effective policies. These evidence-based suggestions will contribute towards reaching the goal of reducing the frequency and severity of road accidents, ultimately saving lives and improving overall road safety.

1 Introduction

According to the National Highway Traffic Safety Administrations report [1], the US recorded 18,720 people death in motor vehicle traffic crashes during the first half of 2024. Around 9202 Drivers,3267 pedestrians and 2821 Passengers lost their lives. Keeping aside the financial loss, delays, and traffic due to such road accidents, the number of deaths are itself alarming, which calls for the need of scientific study of factors responsible for these accidents, study patterns in road accidents, and understanding Driver Demographics and other contributing factors to it.

The study intends to study factors such as Demographic Factors, Vehicle Conditions, Road Conditions, and environmental conditions that led to these accidents, derive co-relations and patterns using Machine Learning Algorithms like SVM, Decision Trees, Gradient boosting, Naive Bayes, Random Forest, etc algorithms.

2 Data and preprocessing

The data for this study has been taken from KAGGLE. Several data preprocessing steps were applied on the data:

1. **Remove Duplicate Values**: Removing duplicate values is an essential step of data cleaning for any data science project. It helps in reducing the bias where certain data points are represented multiple times. If the duplicate values are not removed, it can skew the results and therefore lead to incorrect conclusions.
2. **Validation**: This step of data cleaning is done to validate that the data in the dataset is useful for the problem we are solving.
3. **Detection and Removal of Outliers**: Outliers in the data can impact the decision-making using the analytics from the data. We should detect and process the outliers.
4. **Handling Missing Values**: In this step of Data Cleaning, we either remove or impute the missing values in the dataset.
5. **Correcting Errors**: In this data cleaning, we identify and fix the errors or inconsistencies present in the data.
6. **Standardize the Data**: Convert all the text in the dataset to lowercase to ensure consistency. This helps in avoiding the situations where same words with different cases are considered different.
7. **Parsing the data**: Convert all the text in the dataset to lowercase to ensure consistency. This helps in avoiding the situations where same words with different cases are considered different.
8. **Feature Engineering**: It is the process of creating, transforming, or selecting the most relevant features (variables) in a dataset to improve the performance of machine learning models. Using the existing columns, we create new features that help in finding new patterns in the data.
9. **Ordinal Encoding**: It is a technique for converting categorical data into numerical values while also maintaining the inherent order of the categories. Categorical data should be converted so that they can be fed to the algorithms that are used on the data.
10. **One Hot Encoding**: One-hot encoding is a method for converting categorical data into a numerical format that machine learning algorithms can use. It transforms each unique category value into a new binary column, ensuring the algorithm treats categories as separate, independent entities. Categorical data should be converted so that they can be fed to the algorithms that are used on the data

3 Problem Statements :

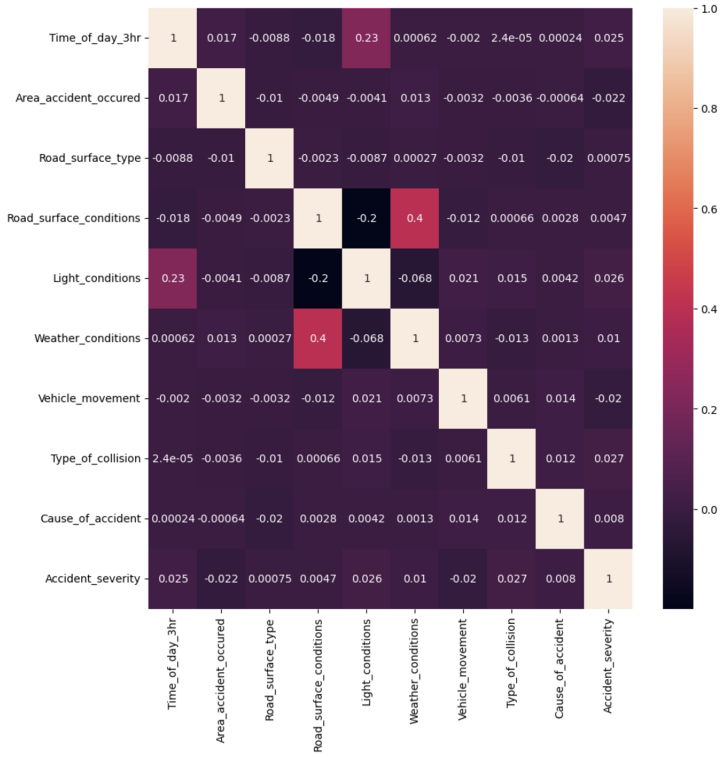
- How do driving experience, road surface conditions, and educational level affect the severity of accidents
- Analyzing how the fatality is related to various factors such as light conditions, weather conditions, type of collision, and day of the week in traffic accidents. Finding patterns and correlations that can suggest road safety strategies.
- Not all vehicles are involved in road accidents equally. Some vehicles have a higher tendency to be involved in any road accident

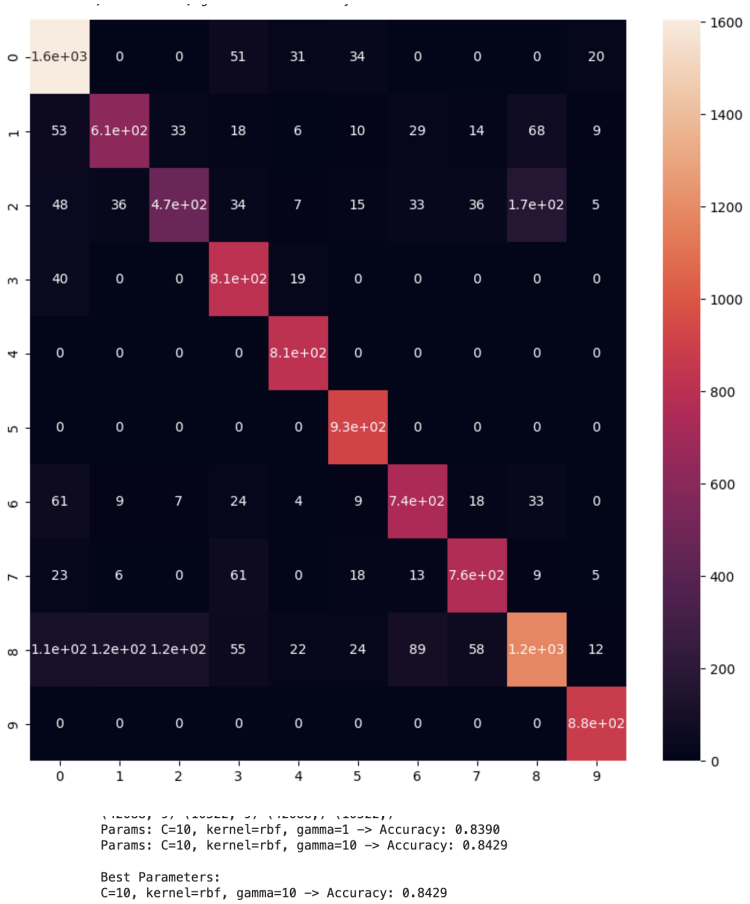
- Does the service period of the vehicle and ownership of the vehicle have any correlation with the accidents
- What is the Impact of area, type of road cross-section, type of roads, and road alignment on different types of Accidents?
- What is the impact of Environmental factors, Light(visibility) impact, Road surface, time of the day, etc?

4 Methods

4.1 SVM

It is a supervised machine-learning algorithm used for classification and regression. It finds the hyper-plane which divides the points belonging to different groups in a high dimensional space. Data Points that are inseparable in lower dimensions can be separated in higher dimensions using kernel transformations.
To Determine the effect of environmental factors on the Type of collision in accidents we consider various factors like Time of the day, Light conditions, weather conditions, Road Conditions, Road surface conditions, etc, and find the correlation between these factors, we can the relation using a Heat Map.



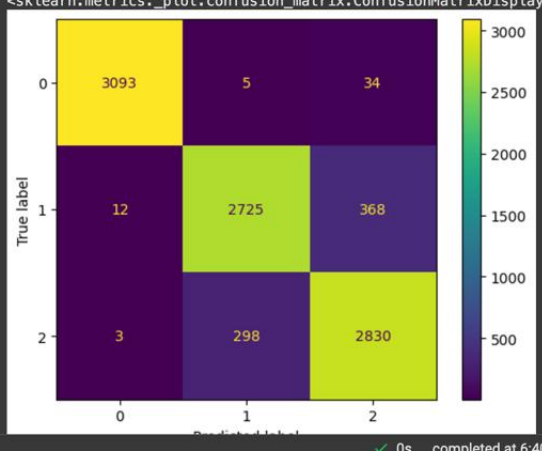


We Were able to achieve an accuracy of about 84 % in the prediction of the type of collision By understanding these factors and predicting results in advance can help drivers to understand vulnerable conditions to take care of while driving under similar situations and be cautious to avoid collisions. It will allow authorities to devise plans to tackle the environmental conditions and spread awareness about dangerous road sections, low visibility regions, appropriate road signs to warn drivers, take measures to improve lighting conditions, frequently clean snow on the road, etc.

4.2 Extra Trees Classifier (ETC)

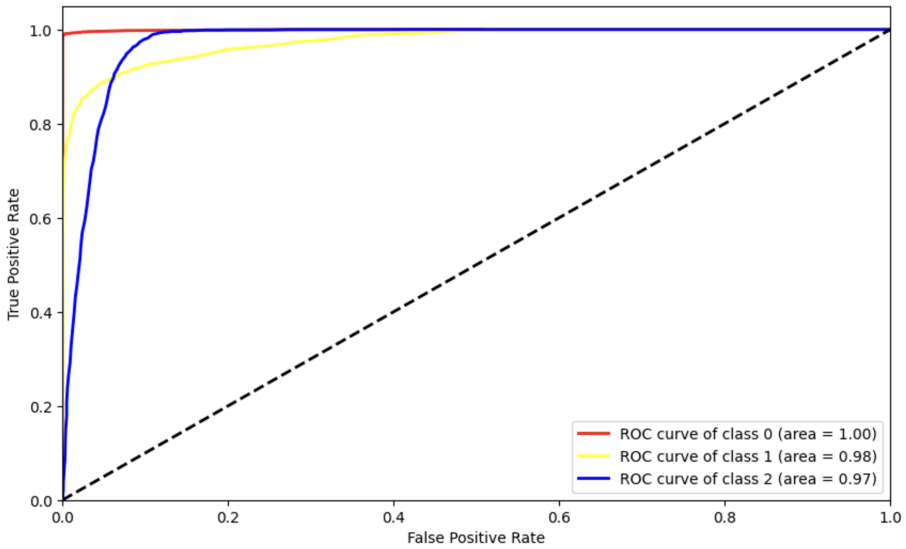
To analyze the effect of driving experience, educational level, age group, etc on the severity of an accident we trained a Decision tree model over data. This ensemble model, similar to Random Forest, is highly effective for classification tasks, especially with datasets having complex interactions. It improves model robustness through random sampling and has high accuracy on your dataset, demonstrated by its optimal results. We were able to derive helpful features from the model with higher precision.

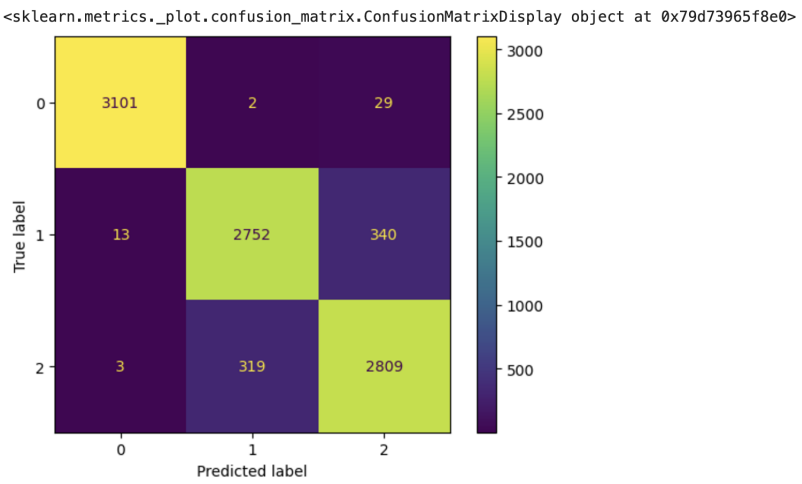
```
--- Model Name: ETC Scenario: Oversampled_without_weights ---
(21856, 135) (21856,) (9368, 135) (9368,) None
------------------  Grid Search Output  ------------------
Best Score: nan
Best Hyperparameters:  {'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split':
------------------   Prediction Starts    ------------------
MCC: 0.9622253996106025
Precision Score : 0.9748348269805301
Recall Score : 0.9748078565328779
F1 Score : 0.9748074300652042
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3145
           1       0.97      0.96      0.96      3128
           2       0.96      0.97      0.96      3095

    accuracy                           0.97      9368
   macro avg       0.97      0.97      0.97      9368
weighted avg       0.97      0.97      0.97      9368

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at 0x7de1bb614250>
```



4.3 Random Forest

Using Random Forest, we are trying to predict the severity of the accidents based on ownership of the vehicle. Training the model over data with fine-tuning of the parameters, we get the following result and confusion matrix.





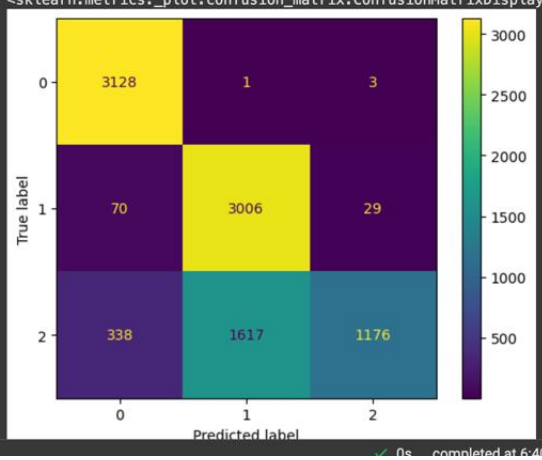
We plot the ROC curve for the NB classifier.

```
--- Model Name: RandomForest Scenario: Oversampled_without_weights ---
(21856, 135) (21856,) (9368, 135) (9368,) None
------------------  Grid Search Output  ------------------
Best Score: nan
Best Hyperparameters:  {'class_weight': None, 'n_estimators': 100}
------------------   Prediction Starts    ------------------
MCC: 0.9630128019070904
Precision Score : 0.9753156428459905
Recall Score : 0.9753415883859948
F1 Score : 0.9753272650990514
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      3145
           1       0.97      0.96      0.96      3128
           2       0.97      0.97      0.97      3095

    accuracy                           0.98      9368
   macro avg       0.98      0.98      0.98      9368
weighted avg       0.98      0.98      0.98      9368
```

4.4 KNN

KNN works well for smaller datasets or cases with fewer distinguishing patterns. Despite being sensitive to class imbalance, KNN serves as a useful comparison for model performance, especially for understanding how proximity-based classification can capture patterns in accident data. But this weakness is being taken care of in the code. Using this we predict the accident severity based on a variety of factors such as demographic factors along with certain road conditions. We get better accuracy results in predicting the accident severity.

```
--- Model Name: KNN Scenario: Oversampled_without_weights ---
(21856, 135) (21856,) (9368, 135) (9368,) None
------------------  Grid Search Output  ------------------
Best Score: nan
Best Hyperparameters:  {'n_neighbors': 1}
------------------   Prediction Starts    ------------------
MCC: 0.8944874397814306
Precision Score : 0.9336893907567986
Recall Score : 0.9268787361229718
F1 Score : 0.9255973099467952
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      3145
           1       0.86      0.98      0.91      3128
           2       0.99      0.80      0.89      3095

    accuracy                           0.93      9368
   macro avg       0.93      0.93      0.93      9368
weighted avg       0.93      0.93      0.93      9368

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at 0x7de19a807b80>
```

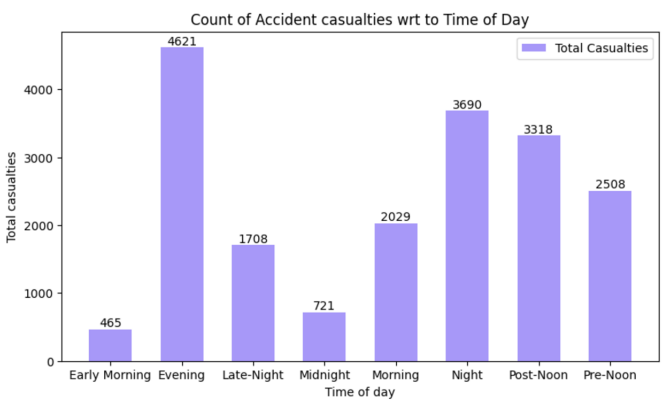


0s completed at 6:40 PM

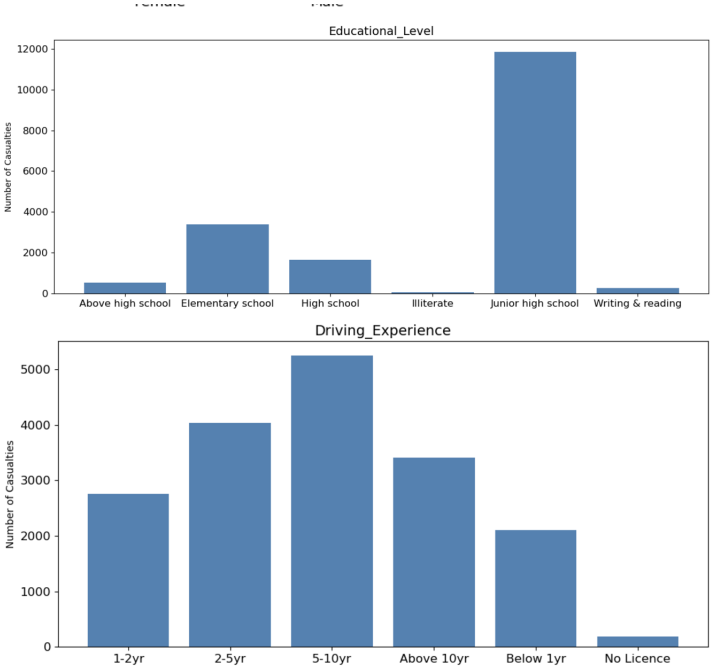
5 Analysis

From the above implementation, we can draw the following inferences.

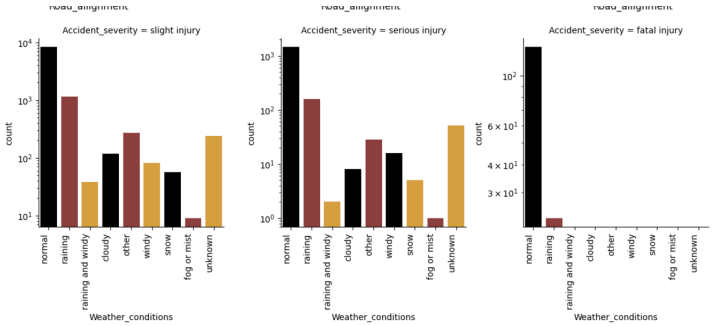
- Days of week and time of day have influence on the count of accidents, it may be due to large people commuting and being in a hurry to reach the office in time, resulting in more carelessness and an increased number of accidents on weekdays. This is evident from accidents in office areas in weekdays.



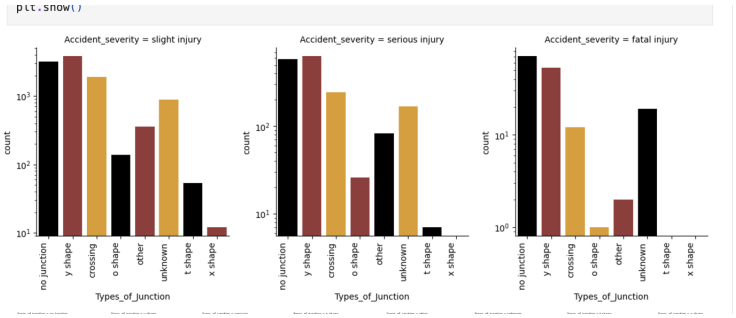
- Age of Driver and Driving experience are also influential factors influencing the driving style that leads to accidents. Young drivers who recently learned to drive are careless and account for a major share of accidents due to overspeeding.



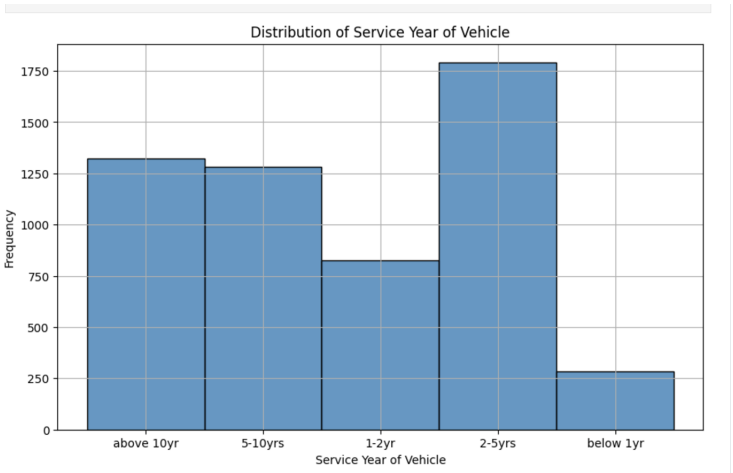
- Environmental Factors like Light Conditions along with road conditions affect accidents. Due to lack of visibility, drivers are not able to see the road properly combined with bad road conditions like water, snow on the road makes a lethal combination. The driver needs to drive carefully in such a situation. Light Condition - Daylight Conditions and Normal it is the most important feature for severity of accidents. That means, In Daylight and normal conditions, most of the accidents occur. It is obvious as at Night, people tends to avoid travelling.



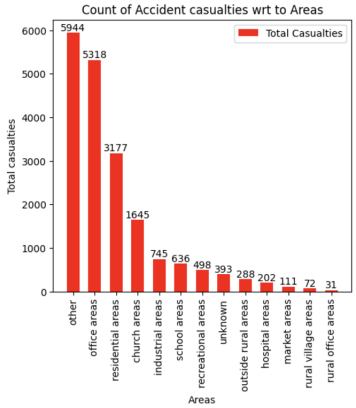
- Weather conditions: From overall perspective, there is almost very low effect of weather conditions but weather conditions normal and weather conditions raining are the main cause when considering only these 22 features for this problem statement.
- Road engineering needs to be improved, a large number of accidents occur at the Y-crosssection of the road due to the inability to spot another incoming vehicle. Many vehicles collide with cars parked roadside this shows a requirement for proper planning while designing roads.



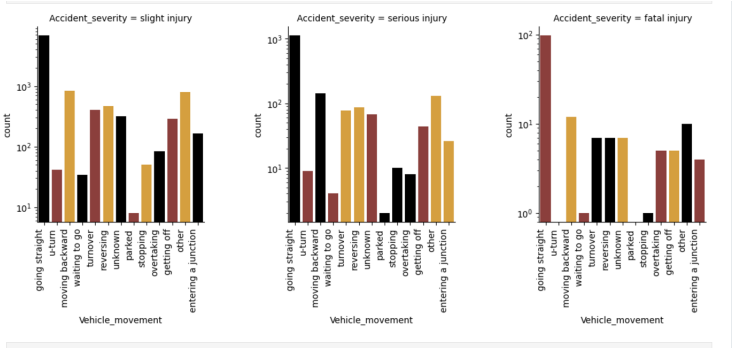
- Vehicles with delayed servicing periods are more prone to accidents. From the data, we can see fewer accidents for vehicles that had maintenance in the past year than the others.



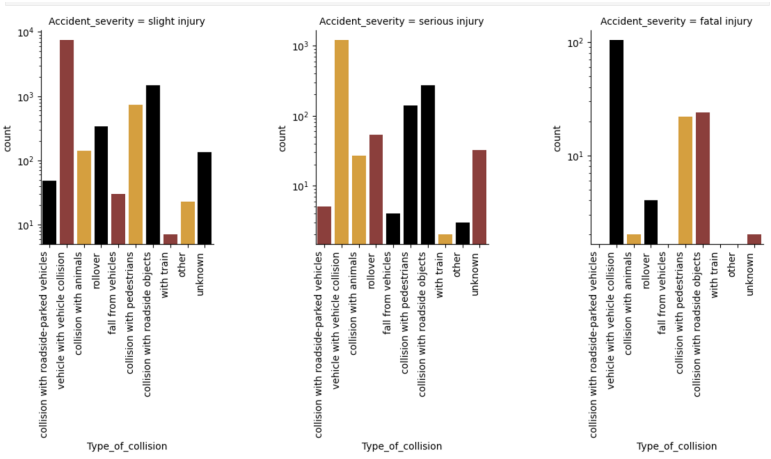
- Areas with offices, schools, and urban areas experience higher accidents as compared to rural and other areas



- Driving Discipline is important and it entirely depends on human behavior, violating traffic rules causes unnecessary events and leads to casualties.



- Type of the collision with another vehicles and road side objects is ranked 2nd among factors for the accident severity that means accident occurrance is moderately affected by collisions with another vehicle and with road side objects



6 Suggestions

From the Analysis of the Road Accidents Dataset, we can make some suggestions to improve the situation and assist government authorities in reducing Road accidents and casualties.

- Weekend days are a major cause of accidents, Increase the number of police checkpoints and patrols on weekends, especially in areas with high accident rates.
- Authorities should Implement mandatory road safety and awareness programs in junior high schools explaining the importance of safe driving.
- The government should introduce incentives for regular vehicle servicing as servicing of the vehicle is the 4th most important feature across Models
- Require or incentivize drivers with 5-10 years of driving experience to participate in defensive driving courses every few years. Curb driving without a driving license.
- Install road signs to alert drivers of accident-prone spots and street lights wherever needed.
- Vehicles should make improvements to their braking system, tire quality, and safety measures like airbags to save the lives of passengers in the worst case.

7 Conclusion

From the study, we identified numerous factors that contribute to the increased accident count. We studied the correlation between the factors helping us derive conclusions to make improvements. The analysis can help in designing roads based on past data to avoid accident-prone spots and provide proper light and drainage systems to avoid water collection on roads and many such conditions.

References

1]https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813661