Preamble

This document serves as a formal acknowledgment of the contributions made by each team member towards the completion of ***Indian Road Analytics***. It outlines the specific contributions and the estimated percentage of total effort each member has invested in the project. This preamble is intended to ensure transparency and mutual recognition among team members.

Contributions

Mohammed Allam - 30106564

- Implemented the Impact of Road Conditions on Accident Severity model
- Contributed to writing the report
- Total Contribution: 25%

Kusumpreet Kaur Heer - 30114618

- Implemented the Impact of Geographic Location on Accident Severity model
- Contributed to correlation analysis
- Estimated % of Total Contribution: 25%

Rayner Nyud - 30120995

- Implemented the Impact of Drivers Educational Level on Accident Severity model
- Contributed to writing the report
- Estimated % of Total Contribution: 25%

Kai Ho Chak - 30147119

- Implemented the Impact of Geographic Location on Accident Severity model
- Contributed to correlation analysis
- Estimated % of Total Contribution: 25%

Declaration

We, the undersigned, affirm that the above statement accurately reflects our contributions to the project and the estimated percentages of the total effort. We acknowledge that this document is a true representation of our work and that any misleading or false declarations may have consequences.

Kai Ho Chak | Mohammed Allam | Rayner Nyud | Kusumpreet Kaur Heer

Abstract

This study analyzes the causes of road accidents in India, focusing on environmental, demographic, and vehicular factors. It aims to enhance road safety by identifying key contributors to accidents, such as location, road conditions, and driver education. Utilizing statistical and machine learning techniques, the research delves into a comprehensive dataset to uncover patterns that inform targeted safety measures. The findings indicate significant correlations between accident severity and various factors, offering insights for policy-making and future research. Overall, this study provides a deeper understanding of road accidents in India, contributing to efforts to reduce their incidence and improve public safety.

Introduction

In this report, we undertake a comprehensive analysis of the causes of road accidents in India, focusing on environmental and demographic factors. The significance of this study lies in its potential to improve road safety by identifying key factors contributing to accidents. In India, where road accidents pose a major challenge, understanding these factors is vital for reducing accidents and enhancing public safety, with implications for healthcare, traffic management, and economic productivity.

Previous studies in this domain have utilized data visualization and statistical methods to explore variables such as time, day, driver demographics, vehicle details, and their correlations with road accidents. Common findings include patterns in accidents based on days of the week, the influence of driver demographics on accident severity, and the importance of vehicle maintenance. These studies, primarily employing exploratory data analysis (EDA), have laid the groundwork for policy formulation.

Our project aligns with these studies in its goals and data utilization, particularly in analyzing accident hotspots and the impact of road conditions on accidents. However, we introduce a novel aspect by examining the correlation between drivers' education levels and accident rates, potentially offering fresh insights for educational initiatives and vehicle monitoring in road safety.

Our analysis encompasses several key areas:

- Accident Hotspots Analysis, where we identify and compare the frequency and timing of accidents in specific locations.
- Impact of Road Conditions on Accident Severity, focusing on how various road conditions affect accident outcomes.
- Educational Level Effectiveness, investigating the relationship between drivers' education levels and their involvement in accidents, including the nature and severity of these incidents.

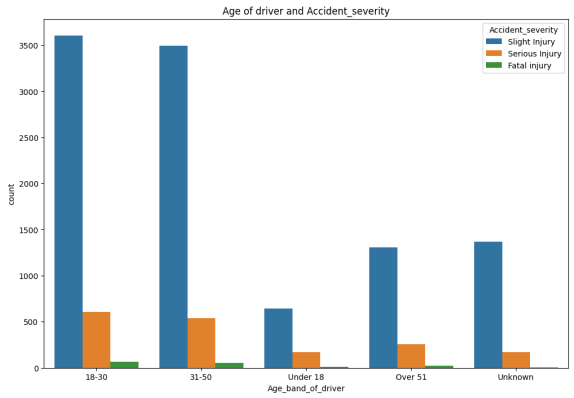
Looking forward, we aim to predict vehicle malfunctions and assess vehicle risks by analyzing factors like age, type, and maintenance history. This future work is geared towards identifying vehicles more prone to accidents and the characteristics common to vehicles frequently involved in accidents.

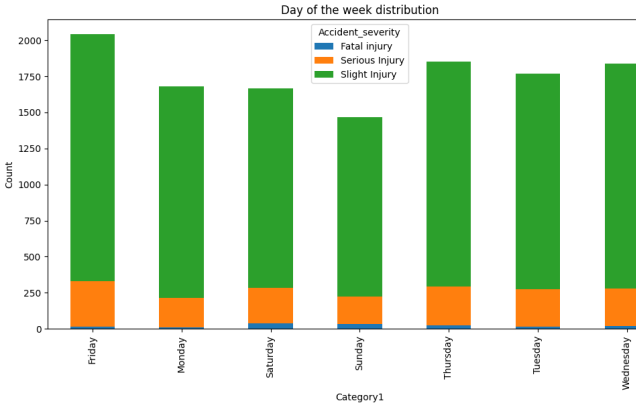
Background and related work

This project focuses exclusively on the severity of road accidents in India, examining critical factors such as geographic location, road conditions, and the role of education in influencing these incidents. By adopting a targeted approach, we aim to uncover how these specific elements interact to exacerbate the severity of road accidents, providing crucial insights for effective intervention strategies.

The study of road accident severity in India is vital, impacting the safety and well-being of countless individuals. Understanding the nuances of how location, road conditions quality, and educational aspects contribute to accident severity is essential to finding ways to reduce accident severity. This knowledge is not just critical for improving road safety but also for shaping informed urban planning and public policy, ultimately fostering safer and more resilient transportation networks across India.

Here are some statistics about the data we're working with. More findings can be found in "general_statistic.ipynb" as attached.





Technical background

Our project harnesses statistical analysis, machine learning, and data visualization to delve into the severity of road accidents in India. We use the Chi-Square test to uncover significant associations among categorical variables influencing accident severity.

P-values guide us in gauging the statistical significance of our findings, informing the inclusion of factors in our machine learning models. Logistic regression helps predict the probability of severe accidents by considering variables like road conditions, while decision trees map out the complex interplay of factors due to their transparency and practicality. To convey our insights, we rely on a suite of visualization tools, crafting an intuitive narrative for stakeholders. Our comprehensive approach aims to provide a nuanced perspective on road safety, aiding in crafting informed interventions.

Review of existing work pertinent to your project

Prior Research

The research on road accident analysis is extensive and focuses on various factors influencing accident rates and severity. Studies commonly examine the relationship between driver demographics (age, sex, education), vehicle types, and accident severity, as well as the impact of road, light, and weather conditions on accidents. Utilizing statistical and machine learning methods, these studies aim to identify patterns and insights to enhance road safety strategies.

Relation to Our Work

Our project builds on existing research by analyzing the combined effects of driver education, road conditions, and vehicle characteristics on accident severity. We incorporate educational levels into predictive models, offering new insights into the impact of education on driving behavior. Utilizing advanced techniques like Logistic Regression and Decision Trees, our comprehensive study aims to deepen the understanding of the multifaceted factors contributing to road accidents in India.

Critical Analysis

Our project addresses gaps in current research by examining the combined effects of education, road conditions, and vehicle types on road accidents. We provide a more comprehensive view of accident causation through advanced machine learning models, aiming for a nuanced understanding and more effective road safety interventions.

Methodology

Experiment setup

1. Software Environment and Tools: We conducted our analysis using a Python environment, with the PySpark, numpy, and seaborn libraries. We also used the ML features within PySpark.

2. Data Loading, Preprocessing and Partitioning in Spark:

A Spark session was initiated to load the dataset into a dataframe from a CSV file. After cleaning and initial data exploration, the dataset was partitioned into a 70-30 split for training and testing.

3. Feature Transformation and running Logistic Regression:

Features were engineered using PySpark's StringIndexer and VectorAssembler, preparing the data for machine learning models. Weighting was applied to the target variable to address class imbalance.

4. Calculating ML model metrics

Once our ML models were trained and tested, we calculated ML model metrics such as f1 score, precision, recall, and accuracy for each model, in order to evaluate the model's performance. We also created confusion matrices to judge what sort of predictions our ML models were making.

Experimentation factors (e.g., types of ML algorithms used, hyperparameters tuned, details on training/test/cross-validation data set, etc.)

1. Machine Learning Algorithms Used:
We focused on two classic ML models; being Logistic Regression, and Decision Tree. We chose these two models as these work well for non continuous data, so we can't use something like linear regression for these predictions.

2. Training, Test, and Cross-Validation Dataset Details:
The data underwent a 70-30 split into training and testing subsets, and we bypassed cross-validation due to our dataset's modest scale.

3. Data Preparation and Feature Importance:
The dataset was prepared for modeling by selecting features and the target variable (`**"accident_severity_integer"**`).The Decision Tree's output provided insights into the relative importance of different features in predicting accident severity. Our findings, constrained by the dataset size, nonetheless yielded valuable directions for enhancing road safety through targeted predictive analytics.

Experiment process:

1. **Feature Engineering and Preprocessing:**

We Used PySpark's ml.feature to transform categorical variables like "Day_of_week" and "Weather_conditions" into numerical indices.

2. **Model Training and Validation:**

Split the dataset into 70% training and 30% testing sets, ensuring the target variable "accident_severity_integer" was evenly distributed. Trained a multinomial logistic regression model.

3. **Iterative Process and Model Improvements:**

Improved models iteratively by adjusting features; for instance, adding "Types_of_Junction_index" increased model accuracy.

Performance metrics - accuracy, precision, recall, F-score, etc.

1. Accuracy:

- The experiment calculated the accuracy of the model, which is a measure of the correctly predicted instances over the total instances in the dataset.
- In our analysis, accuracy was computed using a custom calculation, (accuracy = (float(total_tp) + float(total_tn))/ (total_tp+ total_fp + total_fn + total_tn)') as well as through PySpark's `MulticlassClassificationEvaluator` with `metricName="accuracy"`.
- Accuracy is a fundamental metric in classification problems and provides a general idea of how often the model is correct.

2. Recall:

- Recall (also known as sensitivity) was computed in our experiment. It measures the proportion of actual positive cases that were correctly identified.
- The recall was calculated both as a micro-average (`**recall_micro**`) and for specific classes, which is crucial for understanding the model's ability to correctly identify each class.

3. Precision:

- Precision, which measures the proportion of positive identifications that were actually correct, was also evaluated.
- Precision was calculated using PySpark's built in 'MulticlassClassificationEvaluator'
- Precision is particularly important in scenarios where the cost of false positives is high.

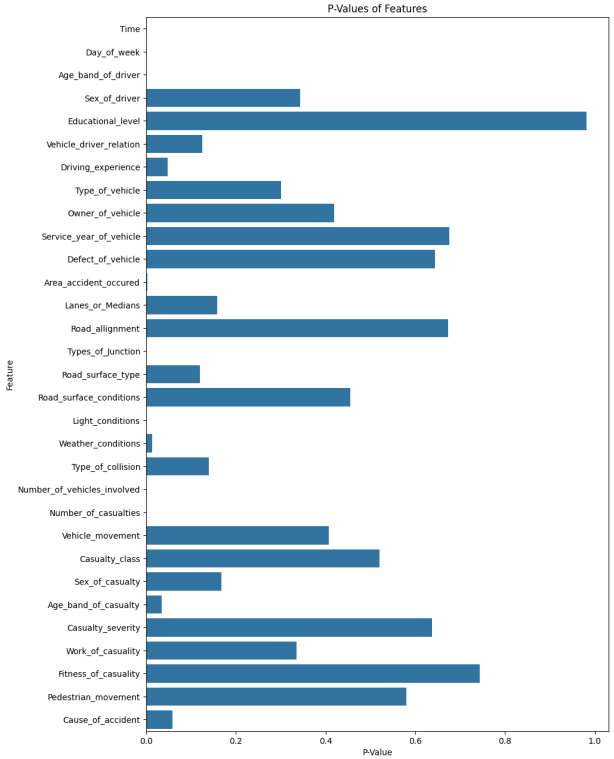
4. F-Score:

- Since f-score penalizes extreme values, F-score reaches its best value at 1 (perfect precision and recall) and worst at 0.
- Since most of our tests have F-score values lie between 0.7 and 0.8, which makes our models a good balance between precision and recall.

Results

Correlation Analysis

In our analytical framework, we employed a two-pronged approach to ascertain the correlation between various features and the severity of road accidents. Initially, we applied the Chi-Square test to quantitatively assess the independence of categorical variables. Features yielding p-values below the significance threshold of 0.05 indicated a potential predictive relationship with accident severity. Except for “time” which is not a categorical value, we acquired 10 features showing statistical significance with the accident severity.



	Feature	Result	P-Value
0	Time	Dependent (reject H0)	0.000000e+00
21	Number_of_casualties	Dependent (reject H0)	2.045418e-67
20	Number_of_vehicles_involved	Dependent (reject H0)	3.013343e-62
17	Light_conditions	Dependent (reject H0)	4.627954e-08
14	Types_of_Junction	Dependent (reject H0)	4.287789e-06
1	Day_of_week	Dependent (reject H0)	4.300319e-06
2	Age_band_of_driver	Dependent (reject H0)	7.452442e-06

11	Area_accident_occured	Dependent (reject H0)	3.205199e-03
18	Weather_conditions	Dependent (reject H0)	1.387240e-02
25	Age_band_of_casualty	Dependent (reject H0)	3.431907e-02
6	Driving_experience	Dependent (reject H0)	4.833484e-02
30	Cause_of_accident	Independent (H0 holds true)	5.875545e-02
15	Road_surface_type	Independent (H0 holds true)	1.202733e-01
5	Vehicle_driver_relation	Independent (H0 holds true)	1.253003e-01
19	Type_of_collision	Independent (H0 holds true)	1.402361e-01
12	Lanes_or_Medians	Independent (H0 holds true)	1.580779e-01
24	Sex_of_casualty	Independent (H0 holds true)	1.677773e-01
7	Type_of_vehicle	Independent (H0 holds true)	3.005278e-01
27	Work_of_casuality	Independent (H0 holds true)	3.354673e-01
3	Sex_of_driver	Independent (H0 holds true)	3.429351e-01
22	Vehicle_movement	Independent (H0 holds true)	4.068349e-01
8	Owner_of_vehicle	Independent (H0 holds true)	4.184744e-01
16	Road_surface_conditions	Independent (H0 holds true)	4.551747e-01
23	Casualty_class	Independent (H0 holds true)	5.203312e-01
	...		
13	Road_allignment	Independent (H0 holds true)	6.734378e-01
9	Service_year_of_vehicle	Independent (H0 holds true)	6.753270e-01
28	Fitness_of_casuality	Independent (H0 holds true)	7.442941e-01
4	Educational_level	Independent (H0 holds true)	9.818024e-01

This statistical validation was complemented by a logical test—a qualitative evaluation where features were chosen based on their contextual relevance to the specific problems under investigation in the Indian road accident scenario.

Post these tests, our team engaged in a deliberative process, applying qualitative criteria to weigh the practical significance of each feature against our investigative goals. This integrative method ensured that the final selection of features held statistical significance and aligned with our study's substantive focus, thereby enhancing the robustness and applicability of our subsequent predictive modeling.

Question 1 - Geographic Location and Accident Severity

	Time	Day_of_week	Age_band_of_driver	Driving_experience	Type_of_vehicle	Area_accident_occured	Road_allignment	Types_of_Junction	Road_surface_conditions	Light_conditions	Weather_conditions	vehicles_involved	Number_of_casualties	Age_band_of_casualty
Chi Sqaure Test		x	x	x		x		x		x	x	x	x	x
Q1 Geographic - Logical Test	x	x			x	x	x	x	x	x	x			
Q1 Geographic - Final Selection	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**			

Features that met the criteria of both the Chi-Square and logical tests, such as **"Day_of_week", "Area_accident_occured", "Types_of_Junction", "Light_conditions", and "Weather_conditions",** were confidently incorporated into our feature set. These variables demonstrated both statistical significance and contextual relevance, underscoring their potential impact on accident severity.

Additionally, there were features that, despite not meeting the statistical threshold in the Chi-Square test, were deemed essential based on logical reasoning and were thus retained. The inclusion of these variables was justified by specific rationales pertinent to the nuanced context of road safety and accident analysis.

- **“Time”**: This variable is integral to our core objective of identifying temporal patterns and specific locations with high accident frequencies. By examining the time data, we can develop targeted strategies to improve road safety measures during peak periods of accident occurrences.
- **“Type_of_vehicle”**: Different vehicle types may influence the severity of accidents. For instance, smaller vehicles may have more maneuverability to evade accidents, whereas larger vehicles, such as lorries, may be more susceptible due to their size and handling characteristics.
- **“Road_allignment”**: This factor is a crucial aspect of the geographic location analysis. The configuration and curvature of the road can significantly impact the likelihood and severity of accidents, making it a key variable for our spatial analysis.
- **“Road_surface_conditions”**: This feature captures the quality and characteristics of the road surface, which can directly affect vehicle traction and, consequently, accident severity. Poor surface conditions can increase the risk of accidents, especially in adverse weather conditions.

We selected features like **"Age_band_of_driver"** and **"Driving_experience,"** which showed relevance in the Chi-Square test but not initially in logical assessments. The age of a driver influences driving habits, with younger drivers potentially speeding in rural areas and older drivers facing challenges in busy urban environments. Similarly, seasoned drivers may navigate congested areas more skillfully, highlighting the nuanced impact of experience on driving safety.

Features like **"Number_of_vehicles_involved," "Number_of_casualties," and "Age_band_of_casualty,"** though significant in the Chi-Square test, were excluded for irrelevance to our core analysis. Our focus is on factors leading to accidents, not on their aftermath. Including these variables would not enhance our predictive model, emphasizing the importance of aligning features with specific analytical objectives.

Question 2 - Road Condition and Accident Severity

	Day_of_week	Age_band_of_driver	Driving_experience	Area_accident_occured	Road_allignment	Types_of_Junction	Road_surface_type	Road_surface_conditions	Light_conditions	Weather_conditions	Type_of_collision	Number_of_vehicles_involved	Number_of_casualties	Age_band_of_casualty
Chi Sqaure Test	x	x	x	x		x			x	x		x	x	x
Q2 Road Condition - Logical Test					x		x	x	x	x	x	x		
Q2 Road Condition - Final Selection					x		x	x	x	x	x	x		

Similarly, we included features like **"Number_of_vehicles_involved," "Light_conditions," and "Weather_conditions"** in our model, as they passed both the Chi-Square and logical tests.

Some features didn't meet the Chi-Square test's statistical criteria but were kept based on logical considerations. Their retention is supported by specific justifications as follows.

- **“Road_allignment”**: Different road alignments impact the driver's ability to navigate safely. Sharp curves, steep grades, and escarpments can increase the likelihood of severe accidents due to challenges in maneuvering, high speeds, or under poor visibility.

- **“Road_surface_type”**: The material and condition of the road surface affect traction and vehicle control. For instance, gravel or earth roads might be more prone to causing vehicle skidding or loss of control compared to well-maintained asphalt roads.

- **“Road_surface_conditions**”: Wet or Damp, Dry, Flooded, Snow, these conditions directly affect tire grip and braking efficiency. Wet, flooded, or snowy roads can significantly increase the risk of accidents and potentially lead to more severe outcomes due to reduced vehicle control.

- **“Type of Collision”**: Rollover, Collision with Pedestrians, Vehicle with Vehicle Collision, etc.: The nature of the collision gives insights into the dynamics of the accident. For example, rollovers and collisions with pedestrians often result in more severe injuries compared to other types of collisions.

Features including **"Day_of_week," "Age_band_of_driver," "Driving_experience," "Area_accident_occurred," "Types_of_Junction," "Number_of_casualties," and "Age_band_of_casualty"** were excluded from our analysis.

Question 3 - Education Level and Accident Severity

	Day_of_week	Age_band_of_driver	Educational_level	Driving_experience	Area_accident_occured	Types_of_Junction	Light_conditions	Weather_conditions	Number_of_vehicles_involved	Number_of_casualties	Age_band_of_casualty	Cause_of_accident
Chi Sqaure Test	x	x		x	x	x	x	x	x	x	x	
Q3 Education - Logical Test		x	x	x								x
Q3 Education - Final Selection		x	x	x	x							x

We incorporated **"Age_band_of_driver" and "Driving_experience,"** as they cleared both the Chi-Square and logical assessments. Their statistical significance and practical relevance underscore their role in influencing the severity of accidents. Driving experience can be

considered a form of education, so it's important to include it in our analysis. Age could also be helpful if a certain age band is more educated than the others.

Additionally, we retained **"Education_level" and “Cause_of_accident,”** despite their lack of statistical significance in the Chi-Square test. This decision was grounded in logic, aligning with our goal to explore the link between drivers' educational backgrounds and their likelihood of causing accidents, a factor crucial for designing targeted educational measures. The cause of an accident can be education/driving experience related, so we chose to include it.

However, we chose not to include features like **"Day_of_week," "Area_accident_occurred," "Types_of_Junction," "Light_condition," "Weather_conditions," "Number_of_vehicles_involved," "Number_of_casualties," and "Age_band_of_casualty"** in our analysis. Despite their statistical relevance, they do not align with our focus on understanding accident causes rather than consequences. Their inclusion would not have enhanced the model's ability to meet our specific research aims, where we are aiming to use educational-related factors to predict accident severity.

Key Findings for Q1

For Question 1, our main focus was to pinpoint specific geographic locations and temporal patterns where accidents occur most frequently, enabling targeted measures to enhance road safety. We chose various features as stated above, and tried to fine tune the model by adding new features and changing various parameters. These findings can be witnessed in the shared Python file. Upon applying a Logistic Regression model on our features, based on accident severity, we got an accuracy of 0.8477, recall of 0.6867 and F-score of 0.7587. Which made the model a good balance between precision and recall. Upon applying the Decision Tree Classifier on our model, the Accuracy seems to be 0.8466 and the precision is around 0.7592.

Key Findings for Q2

In the analysis of how road conditions affect accident severity (Q2), our dataset encompassed 12,316 incidents, detailed across 32 variables. We meticulously encoded categorical variables numerically to suit machine learning algorithms. Our first approach with a Logistic Regression model yielded a promising accuracy of 85.84% and a recall of 86.78%. However, this model disproportionately favored predicting minor injuries. We attempted to rectify this imbalance by incorporating class weights, which, while well-intentioned, overly adjusted the model's sensitivity and subsequently reduced accuracy to 71.48%. In contrast, the Decision Tree Classifier demonstrated more balanced performance, achieving an accuracy of 84.64%, with notable variables such as the number of vehicles and road alignment proving to be significant predictors of accident severity.

Key Findings for Q3

For Q3, focusing on the influence of educational factors on accident severity, we applied similar preprocessing steps to the identical dataset, this time accentuating features related to educational background, driver's age, and experience. The modified Logistic Regression model, now with class weights to balance the distribution of severity categories, garnered an accuracy of 63% with an impressive recall of 87%. This indicated the model's proficiency in capturing severe accidents, despite a lower precision of 40%. The Decision Tree model, on the other hand, outperformed in accuracy, with a score of 84%, and showed a superior F1 score of 0.78. This hinted at a more refined balance in the model, adeptly identifying various levels of accident severity without compromising on either the recall or the precision.

Conclusion and suggestions

Overall, the analysis of the 3 questions, and the development of our three ML models shows that for each question, our models do indeed work to predict accident severity based on the factors we decided using statistical analysis. Our ML models all had similar results based on ML metrics, showing that the identified factors do play a part in accident severity. It is also difficult to predict a serious or fatal accident, as the data set for those is small, however with class balancing we have shown that it is still possible.

The analysis of Q1,Q2 and Q3 using our ML methods shows us that overall, geographic locations and temporal patterns, drivers experience and education levels, and also road conditions can be used in order to make a prediction to see how accident severity is affected based on these factors, which can be seen in our ML metrics and confusion matrices.