Comparative Study of Accident-Experienced and Accident-Free Individuals

A PROJECT REPORT

Submitted by

Mr. Parth Vikas Patil

in partial fulfillment for the award of the degree

of

M.Sc. Statistics Part 2

ACADEMIC YEAR 2024 – 2025

Submitted to

DEPARTMENT OF STATISTICS

RAMNIRANJAN JHUNJHUNWALA COLLEGE OF ART'S,

SCIENCE & COMMERCE (AUTONOMOUS),
GHATKOPAR (W)

RAMNIRANJAN JHUNJHUNWALA COLLEGE OF ART'S, SCIENCE & COMMERCE (AUTONOMOUS), GHATKOPAR (W)

University of Mumbai

(Affiliated to University of Mumbai)

CERTIFICATE

This is to certify that the project entitled **Comparative Study of Accident-Experienced and Accident-Free Individuals** *bonafide work of* ***Mr. Parth Vikas Patil*** *bearing seat no.* ***926*** *during the year 2024-2025 in partial fulfillment of the requirements for the award of Degree Master of Science in Statistics.*

Signature of Internal Guide
(Mr. Jaishankar Singh)

Signature of Co-ordinator
(Mr. Jaishankar Singh)

Seal of the College

Signature of Examiner

ACKNOWLEDGEMENT

A lot of efforts have been taken in this project; however, this work would not have reached its completion and would not have been possible without the kind support and help of many individuals, so we would like to extend our sincere thanks to all of them. We have learned a lot from working on this research project with different participants and gained valuable skills and insights that will help us in our future careers.

First and foremost, we would like to thank the Department of Statistics, Ramniranjan Jhunjhunwala College, Mumbai for providing us with the necessary resources and facilities to conduct this research. The department's commitment to excellence in teaching and research has been a constant source of inspiration throughout our academic journey.

We would like to extend my heartfelt appreciation to our project mentor Mr. Jaishankar Singh, whose unwavering support and guidance have been indispensable in shaping this project. He has provided us with invaluable insights and feedback and has been a constant source of encouragement throughout the process.

We would also like to express our gratitude to Prof. Jaishankar Singh whose expertise and guidance have been invaluable throughout our academic career. He has been a constant source of inspiration, motivation, and encouragement, and his insights and feedback have been instrumental in shaping the direction of this report.

Finally, I would like to express my gratitude to our family and friends for their unwavering support and encouragement. Their love and encouragement have been a constant source of strength throughout my academic journey.

Thank you all for your support, guidance, and encouragement. Without you, this project would not have been possible.

INDEX

Introduction: Comparative Study of Accident-Experienced and Accident-Free Individuals

Road accidents are a pervasive global issue, causing significant loss of life, property damage, and longterm physical and emotional challenges. Understanding the behavioral, environmental, and systemic factors influencing accidents is crucial for effective prevention and intervention strategies.

This study examines the key differences between accident-experienced individuals, who provide data based on real-life incidents, and accident-free individuals, who offer opinions on the same scenarios. By employing a parallel survey design for these two groups, the research explores multiple dimensions—driver behavior, road conditions, emergency healthcare response, and mobile phone usage—that contribute to accident occurrence and severity.

The analysis incorporates statistical techniques to identify and compare patterns within the responses, offering insights into critical behavioral and environmental variables. The study also delves into how external factors, like road type and emergency service delays, influence accident outcomes. By contrasting actual experiences with perceptions, this study aims to provide actionable recommendations to reduce accident rates and mitigate their impact.

This research is envisioned to assist policymakers, urban planners, and road safety advocates in designing targeted interventions that promote safer roads, responsible driving, and efficient emergency healthcare systems.

.

Objectives:

1] Identify key factors of Driver Behavior for accidents.

2] Find relationship between the severity of road accidents and response time of emergency healthcare services.

3] Study the contribution of mobile phone usage (texting, calls, apps) to accidents among drivers, pedestrians.

4] Examine how road types influences accident rates.

Methodology:

Steps Involved In Conducting the Survey:

1. Defining our objectives
2. Specifying information needs
3. Identifying primary data sources
4. Designing questionnaires
5. Pilot survey
6. Modifying questionnaires
7. Data collection
8. Data coding and data entry
9. Data analysis
10. Preparation of project report

Identification of population:

The target population for the 'Comparative Study of Accident-Experienced and Accident-Free Individuals' project consists of individuals aged 10-70. This population includes diverse groups such as graduates, post-graduates, and 12^{th} passed and below that of gender Male and Female etc.

Data Collection: Questionnaire was distributed via Google Forms to reach a broad audience efficiently. Questionnaire was also filled by interviewing individuals personally. The study was conducted in Mumbai, Thane, Kalyan and Palghar region.

Sampling technique: The data has been collected using primary data collection methods. The sampling method that has been used is „Convenience" sampling. Convenience sampling is a type of non-probability sampling that involves the sample being drawn from the part of population which is close to hand.

Statistical Techniques

Exploratory Data Analysis (EDA)

Data visualization techniques were used to represent the distribution and relationship between variables effectively.

Ordinal Logistic Regression

Used to assess the effect of socio-demographic factors on the intensity of food cravings (low, medium, high). This method helped identify which demographic factors significantly influence the likelihood of experiencing different levels of food cravings.

Exploratory Factor Analysis

Applied to identify underlying factors that affect food cravings. This technique helps in reducing the complexity of the data by grouping correlated variables, providing insights into the primary drivers of food cravings.

Chi-Square Test

Applied to examine the association between categorical variables such as mobile phone usage and accident occurrence. This statistical test identifies whether observed differences in proportions are significant, providing insights into which behaviors are strongly linked to accident risk.

Apriori Algorithm

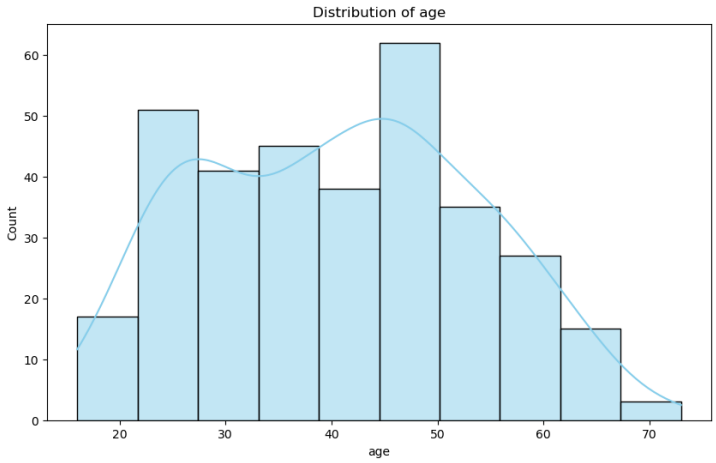
Used to discover frequent patterns and associations between driver behaviors (e.g., speeding, fatigue, phone usage). This rule-based learning technique helps identify combinations of risky actions that commonly occur together and contribute to accident likelihood.

XGBoost (Extreme Gradient Boosting)

Implemented as a powerful classification model to predict accident severity based on multiple behavioral and environmental features. Known for its accuracy and speed, XGBoost provides feature importance rankings, helping identify the most critical factors influencing accident outcomes

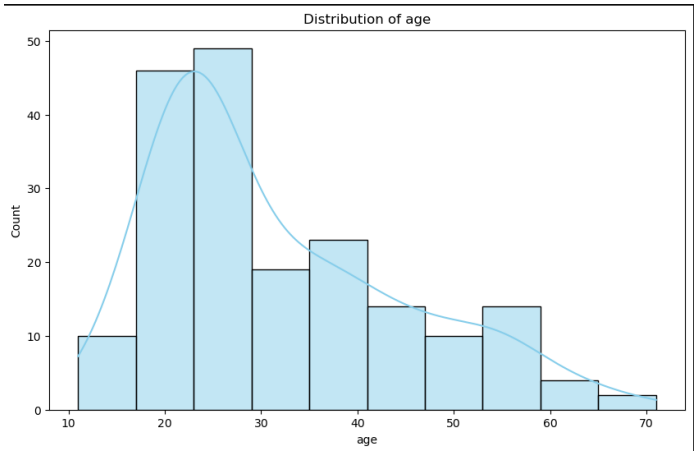
Statistical Analysis

Exploratory Data Analysis



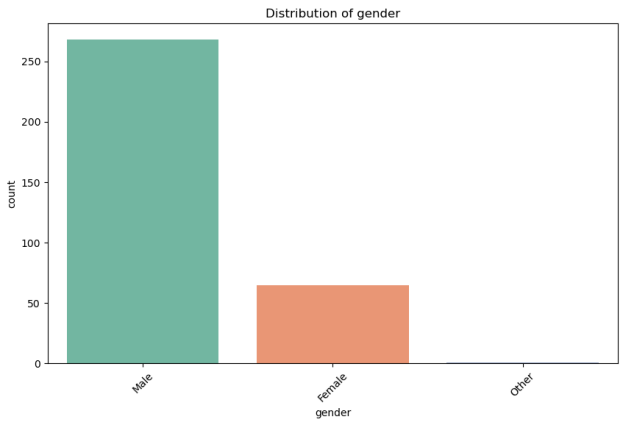
For Experienced

Age distribution of a population, displayed as a histogram. The x-axis marks ages ranging from 10 to 70, while the y-axis shows the count of individuals, scaling up to 50. A majority cluster in the age group of 20-30, and the count declines progressively with age. A smooth density curve accompanies the histogram for additional visual insight.



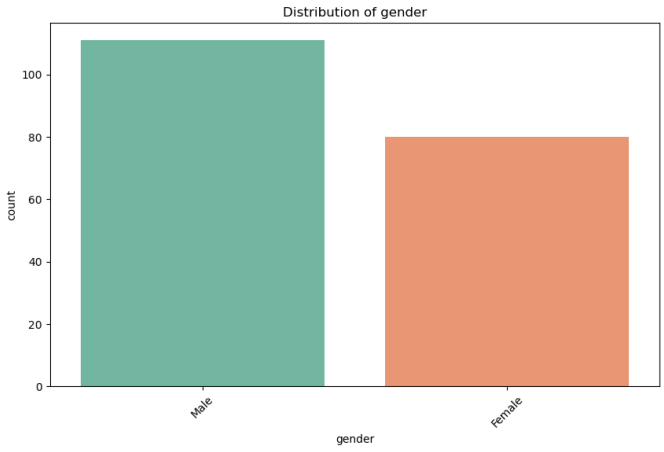
For non-experienced

The histogram displays the distribution of age among respondents. Most individuals fall in the 18–30 age range, with a noticeable peak around the early 20s. The distribution is right-skewed, indicating a higher concentration of younger individuals and a gradual decline in older age groups. The KDE curve reinforces this skewness.



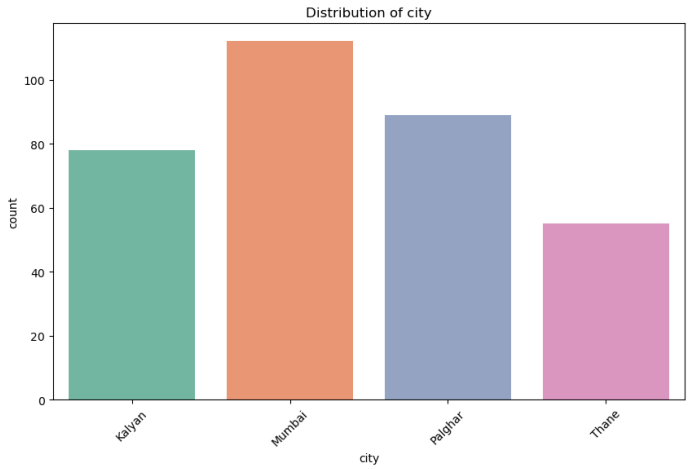
For experienced

The bar chart titled "Distribution of Gender" illustrates the gender composition of the survey respondents. It shows that a larger proportion of participants are male, with over 100 male respondents, compared to around 80 female respondents. This slight gender imbalance may reflect the demographic realities of road usage or survey reach and could influence perceptions and experiences related to road safety captured in the study.



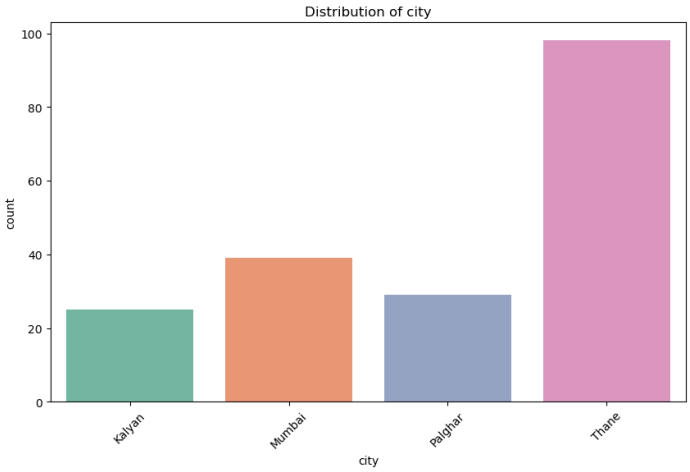
For non experienced

The bar chart shows the distribution of gender among respondents. There are more male respondents (around 110) compared to female respondents (around 80). This indicates a higher male representation in the dataset. The gender labels are displayed on a tilted x-axis for better readability.



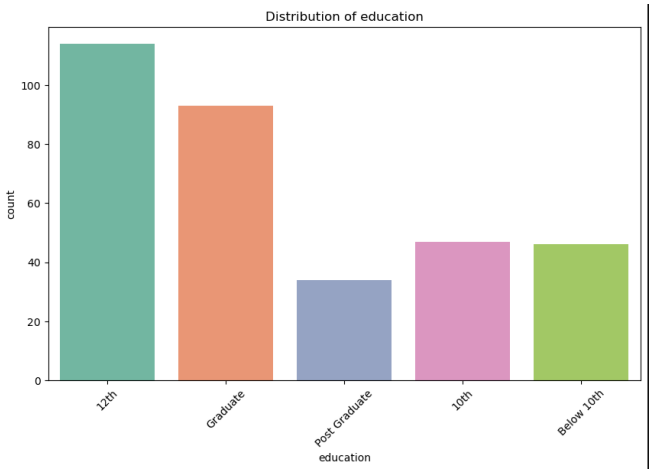
For experienced

The bar chart titled "Distribution of City" shows the number of survey respondents from different cities. Thanhas the highest representation, with nearly 100 participants, followed by Mumbai, Palghar, and Kalyan. This suggests that the data is primarily influenced by responses from Thane, which could reflect the city's higher engagement or accessibility during the survey process.



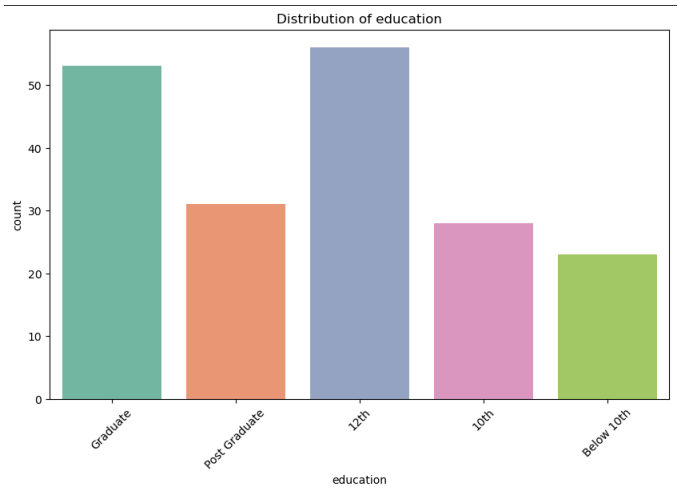
For non experienced

The bar chart presents the distribution of respondents by city. Thane has the highest number of participants, close to 100, followed by Mumbai with about 40. Palghar and Kalyan have fewer respondents, around 30 and 25 respectively. This indicates a major concentration of data from Thane.



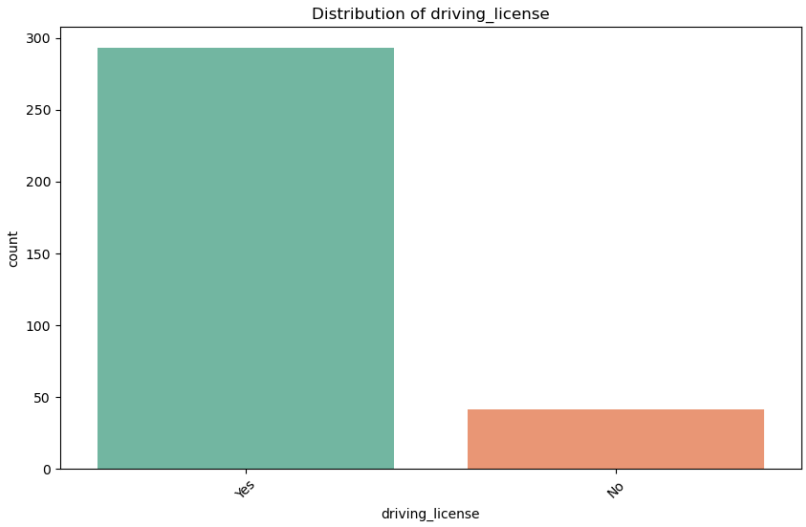
For experienced

The bar chart displays the distribution of education levels among the respondents.
Most individuals have completed their 12th-grade education, making it the highest group.
Graduates form the second-largest group, showing a strong representation of college-educated individuals.
Postgraduates, those with education up to 10th grade, and below 10th are less in number.
This suggests that the sample mostly consists of moderately to well-educated individuals.



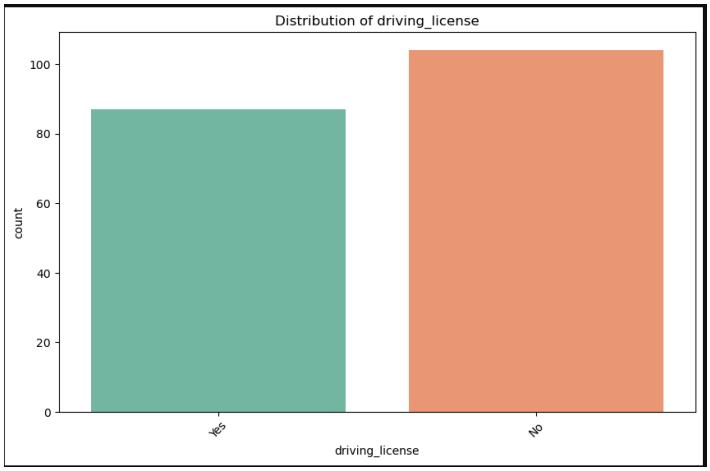
For non-experienced

This bar chart depicts the distribution of educational levels in a population. It compares counts across five categories: Below 10th, 10th, 12th, Graduate, and Post Graduate. The highest count is observed for 12th-grade education, followed by Graduate and Post Graduate levels, while the lowest is for Below 10th-grade education. This visualization provides insights into educational attainment trends.



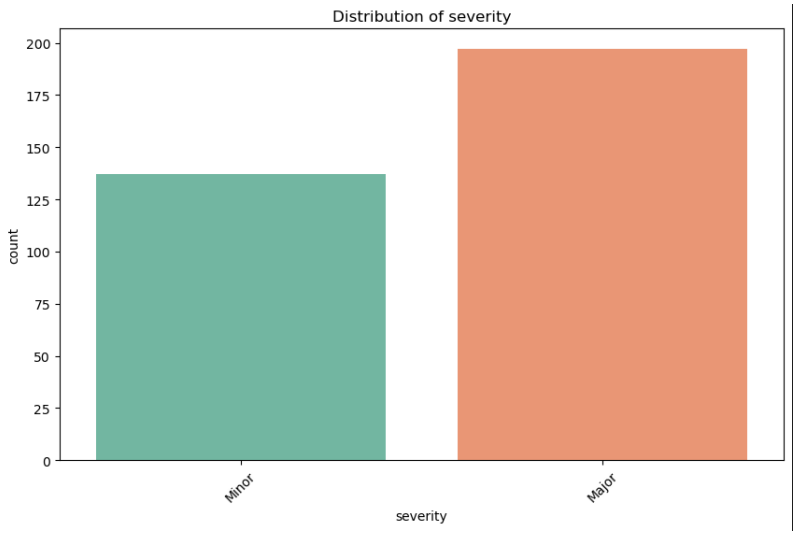
For experience

The bar chart shows the distribution of respondents based on whether they have a driving license. A large majority of individuals reported having a driving license, while only a small portion did not. This indicates that most of the participants are legally eligible to drive.



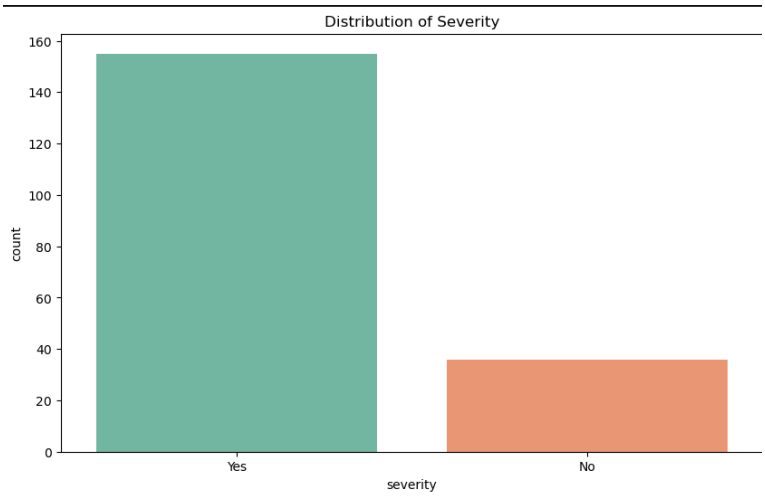
For non-experience

This bar chart compares the count of individuals based on driving license possession. The "Yes" category represents individuals with a driving license and has slightly over 80 people, while the "No" category, for those without a driving license, exceeds 100. It provides a clear visual showing that more individuals lack a driving license than those who have one.



For experience

The bar chart shows the distribution of accident severity among respondents.
Major accidents were reported more frequently than minor ones, indicating a higher occurrence of serious incidents in the data.
This highlights the importance of focusing on preventive measures for severe road accidents.



For non-experience

This bar chart titled "Distribution of Severity" shows the frequency of responses for two categories: "Yes" and "No." The "Yes" category, with a count of approximately 155, dominates over the "No" category, which has about 40 responses. It highlights a significant contrast in the distribution of severity between the two groups.

Objective 1

Identify key factors of Driver Behavior for accidents.

(Chi-Square test & Logistic Regression)

Chi-square test

Chi-Square Test is a statistical method used to determine if there is a significant association between two categorical variables. It is commonly used for analyzing data organized in contingency tables. This test compares observed frequencies with expected frequencies assuming the variables are independent.

Step 1: State Hypotheses

H0 (Null Hypothesis): There is no significant association between the categorical variables.

H1 (Alternative Hypothesis): There is a significant association between the categorical variables.

Step 2: Calculate Chi-Square Statistic

Use:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- OOO = Observed frequency
- EEE = Expected frequency

Step 3: Determine Degrees of Freedom

$$df = (r - 1) \times (c - 1)$$

Where r= number of rows, c = number of columns in the contingency table.

Step 4: Find Critical Value or p-value

Use Chi-Square distribution table or software (like Python, R, SPSS) to find the **critical value** or **p-value** at a given **significance level** (usually $\alpha=0.05$\alpha = 0.05$\alpha=0.05$).

Step 5: Make Decision

- **If** p-value $<$ α **or** $\chi 2$ calculated $>$ $\chi 2$ critical:
  → **Reject H_0** → Significant association exists.
- **Else**:
  → **Fail to reject H_0** → No significant association.

Advantages:

Versatility: Applicable to nominal and ordinal variables.

Non-parametric: Doesn't require assumptions about data distribution.

Ease of Interpretation: Clearly indicates presence or absence of associations.

Disadvantages:

Assumption of Independence: Requires observations to be independent.

Sample Size Sensitivity: May yield unreliable results with small sample sizes or low expected frequencies.

Chi-Square Test Results (For Experienced)

1. Relationship Between Speeding and Severity

Hypothesis:

H0: There is no significant association between speeding and accident severity.

H1: There is a significant association between speeding and accident severity.

Result:

Chi-square statistic: 35.954

Chi-square statistic: 35.954

Conclusion There is a significant association between speeding and accident severity.

2. Relationship Between Fatigue and Severity

Hypothesis:

H0: There is no significant association between fatigue and accident severity.

H1: There is a significant association between fatigue and accident severity.

Result:

Chi-square statistic: 9.294

P-value: 0.054

Conclusion There is no statistically significant association between fatigue and accident severity, though it is near the threshold.

3. Relationship Between Aggressive Driving and Severity

Hypothesis:

H0: There is no significant association between aggressive driving and accident severity.

H1: There is a significant association between aggressive driving and accident severity.

Result:

Chi-square statistic: 13.806

P-value: 0.0079

Conclusion There is a significant association between aggressive driving and accident severity.

4. Relationship Between Multitasking and Severity

Hypothesis:

H0: There is no significant association between multitasking and accident severity.

H1: There is a significant association between multitasking and accident severity.

Result:

Chi-square statistic: 17.954

P-value: 0.00126

Conclusion There is a significant association between multitasking and accident severity.

5. Relationship Between Alcohol Influence and Severity

Hypothesis:

H0: There is no significant association between alcohol influence and accident severity.

H1: There is a significant association between alcohol influence and accident severity.

Result:

Chi-square statistic: 8.014

P-value: 0.091

Conclusion There is no significant association between alcohol influence and accident severity.

6. Relationship Between Vehicle Maintenance and Severity

Hypothesis:

H0: There is no significant association between vehicle maintenance and accident severity.

H1: There is a significant association between vehicle maintenance and accident severity.

Result:

Chi-square statistic: 19.012

P-value: 0.00078

7. Relationship Between Rule Breaking and Severity

Hypothesis:

H0: There is no significant association between rule breaking and accident severity.

H1: There is a significant association between rule breaking and accident severity.

Result:

Chi-square statistic: 3.753

P-value: 0.0527

Conclusion There is no significant association between rule breaking and accident severity, though it's borderline.

8. Relationship Between Injury Severity and Overall Severity

Hypothesis:

H0: There is no significant association between injury severity and overall accident severity.

H1: There is a significant association between injury severity and overall accident severity.

Result:

Chi-square statistic: 3.292

P-value: 0.0696

Conclusion There is no significant association between injury severity and overall accident severity.

Chi-Square Test Results (For Non-Experienced)

1. Relationship Between Speeding and Severity

Hypothesis:

H0: There is no significant association between speeding and accident severity.

H1: There is a significant association between speeding and accident severity.

Result:

Chi-square statistic: 4.346

P-value: 0.3613

Conclusion

There is no significant association between speeding and accident severity in this dataset.

2. Relationship Between Fatigue and Severity

Hypothesis:

H0: There is no significant association between fatigue and accident severity.

H1: There is a significant association between fatigue and accident severity.

Result

Chi-square statistic: 7.850

P-value: 0.0972

Conclusion

There is no statistically significant association between fatigue and accident severity, although the result approaches significance.

3. Relationship Between Aggressive Driving and Severity

Hypothesis

H0: There is no significant association between aggressive driving and accident severity.

H1: There is a significant association between aggressive driving and accident severity.

Result

Chi-square statistic: 17.231

P-value: 0.00174

Conclusion

There is a significant association between aggressive driving and accident severity in this dataset.

4. Relationship Between Multitasking and Severity

Hypothesis:

H0: There is no significant association between multitasking and accident severity.

H1: There is a significant association between multitasking and accident severity.

Result

Chi-square statistic: 9.642

P-value: 0.0469

Conclusion

There is a significant association between multitasking and accident severity in this dataset.

5. Relationship Between Alcohol Influence and Severity

Hypothesis

H0: There is no significant association between alcohol influence and accident severity.

H1: There is a significant association between alcohol influence and accident severity.

Result

Chi-square statistic: 5.462

P-value: 0.2431

Conclusion

There is no significant association between alcohol influence and accident severity.

6. Relationship Between Vehicle Maintenance and Severity

Hypothesis

H0: There is no significant association between vehicle maintenance and accident severity.

H1: There is a significant association between vehicle maintenance and accident severity.

Result

Chi-square statistic: 0.279

P-value: 0.9911

Conclusion

There is no significant association between vehicle maintenance and accident severity.

7. Relationship Between Rule Breaking and Severity

Hypothesis

H0: There is no significant association between rule breaking and accident severity.

H1: There is a significant association between rule breaking and accident severity.

Result

Chi-square statistic: 10.880

P-value: 0.00097

Conclusion

There is a significant association between rule breaking and accident severity.

8. Relationship Between Injury Severity and Overall Severity

Hypothesis

H0: There is no significant association between injury severity and accident severity.

H1: There is a significant association between injury severity and accident severity.

Result

Chi-square statistic: 8.257

P-value: 0.00406

Conclusion

There is a significant association between injury severity and accident severity in this dataset.

Overall Summary:

Factor	Experienced	Non-Experienced	Difference
Speeding	Significant	Not Significant	Stronger impact for experienced
Fatigue	Borderline	Borderline	Slightly closer to significant in experienced
Aggressive Driving	Significant	Significant	Common risk for both
Multitasking	Significant	Significant	Common risk, stronger for experienced

Factor	Experienced	Non-Experienced	Difference
Alcohol Influence	Not Significant	Not Significant	Not significant in either group
Vehicle Maintenance	Significant	Not Significant	Risk only for experienced
Rule Breaking	Borderline	Significant	Riskier for non-experienced
Injury Severity	Not Significant	Significant	Impactful for non-experienced only

Key Conclusion:

Experienced and non-experienced individuals differ significantly in the risk factors that influence accident severity.

- Experienced drivers show stronger associations with internal behavioral factors like speeding**,** vehicle maintenance**,** and multitasking**.**
- Non-experienced drivers are more affected by situational and impulsive behaviors like rule breaking, aggressive driving, and injury severity.

Logistic Regression

Introduction Ordinal logistic regression is a statistical model used for predicting ordinal outcomes. It is an extension of logistic regression for ordinal dependent variables. It is a type of regression analysis that predicts the probability of an ordinal outcome based on one or more independent variables. In short, ordinal logistic regression estimates the relationship between the independent variables and the ordered categories of the dependent variable. Unlike binary logistic regression, which predicts a binary outcome (e.g., yes/no), ordinal logistic regression is suitable for outcomes with more than two ordered categories (e.g., low/medium/high).The model estimates one set of coefficients that describe the relationship between the independent variables and the odds of being in each category of the dependent variable, relative to a reference category. It assumes that the odds ratios between consecutive categories of the dependent variable are constant across all levels of the independent variables. Model fit is typically assessed using methods like likelihood ratio tests or the proportional odds assumption test. The ordinal logistic regression model is:

$$\text{logit}\big(P(y \le j)\big) = \log\left(\frac{P(y \le j)}{1 - P(y \le j)}\right) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + \cdots + \beta_{pj}X_p$$

where,

Y: Response variable

X : Vector of Independent variables

P(y≤j) : Cumulative probability of the outcome being in or below category j.

β0j : Intercept specific to category j.

β1,β2,...,βp : Regression coefficients for predictor variables X1,X2,...,Xp.

Assumptions-

• The dependent variable is measured on an ordinal level.

• One or more of the independent variables are either continuous, categorical or ordinal.

• No Multicollinearity i.e. when two or more independent variables are highly correlated with each other.

• Proportional Odds - i.e. all slopes are equal.

Dependent variable –

Y: Severity

Independent variables –

speeding multitask rule_break

fatigue alcohol

aggressive maintainence

Tool used: Python

Steps:

1. **Split the Data**:

   Divide the dataset into training and testing sets, typically using an 80/20 or 70/30 ratio.

2. **Create Logistic Regression Model**:

   Initialize a logistic regression model with class weights set to 'balanced' to handle class imbalances by adjusting weights inversely based on class frequencies.

3. **Train the Model**:

   Train the logistic regression model on the training data to learn the relationship between features and target labels.

4. **Make Predictions**:

   Use the trained model to predict the target labels for the test data.

5. **Evaluate with Confusion Matrix**:

   Generate a confusion matrix to compare the actual versus predicted values, showing true positives, true negatives, false positives, and false negatives.

6. **Generate Classification Report**:

   Create a classification report that provides performance metrics such as precision, recall, F1-score, and support for each class.

7. **Interpret Model Coefficients**:

   Examine the model's coefficients to understand how each feature contributes to the prediction of the target label.

Result (For experienced):

Confusion Matrix

	Predicted: 0	**Predicted: 1**
Actual: 0	17 (True Negatives)	18 (False Positives)
Actual: 1	13 (False Negatives)	19 (True Positives)

Interpretation:

- **True Negatives (TN)** = 17: Correctly predicted non-severe cases.

- **False Positives (FP)** = 18: Predicted severe but were non-severe.
- **False Negatives (FN)** = 13: Predicted non-severe but were severe.
- **True Positives (TP)** = 19: Correctly predicted severe cases.

A classification model was developed to predict accident severity based on key behavioral and environmental factors. The class distribution showed that approximately **59% of the cases were low severity**, while **41% were high severity**.

The **confusion matrix** and **classification report** indicated modest predictive performance:

- **Accuracy**: 54%
- **F1 Score**: 0.55 (for high severity) and 0.52 (for low severity)
- The model slightly favored **high-severity predictions** (Recall = 0.59), but struggled to distinguish between the classes robustly.

Feature Importance (Model Coefficients)

The key behavioral factors and their contribution (coefficient values) to the model were as follows:

Factor	Coefficient	Interpretation
Rule Breaking	-0.301	Strong negative effect on severity classification. As rule-breaking increases, model predicts lower severity more often. Could indicate misclassification or model bias.
Aggressive	+0.168	Positively contributes to predicting higher severity. Aggressive behavior is linked with more severe accidents.
Multitasking	+0.115	Slightly increases the likelihood of higher severity prediction.
Maintenance	+0.151	Indicates poor vehicle maintenance may relate to increased severity.
Speeding	-0.152	Counterintuitively shows a negative relationship. This could reflect dataset imbalance or underreporting.
Fatigue	-0.041	Minimal effect on the model; could be weakly correlated or inconsistently reported.
Alcohol	+0.005	Very small contribution to severity prediction in this dataset.

Conclusion

- The model demonstrates **moderate performance** and can act as a preliminary severity classifier.
- **Aggressive driving, multitasking, and poor maintenance** appear as **prominent risk indicators** for higher accident severity.

- **Rule breaking** showed a surprising negative contribution and requires deeper investigation—possibly due to how the variable is distributed or labeled in the dataset.
- This analysis can support **data-driven risk profiling** and help inform **targeted interventions** for accident prevention.

Result (For experienced):

Confusion Matrix

	Predicted: 0	**Predicted: 1**
Actual: 0	4 (True Negatives)	4 (False Positives)
Actual: 1	9 (False Negatives)	22 (True Positives)

Interpretation:

- **True Negatives (TN) = 4**: The model correctly predicted 4 cases as non-severe.
- **False Positives (FP) = 4**: The model incorrectly predicted 4 non-severe cases as severe.
- **False Negatives (FN) = 9**: The model predicted 9 severe cases as non-severe.
- **True Positives (TP) = 22**: The model correctly predicted 22 cases as severe.

Model Overview

A logistic regression model was developed to classify accident severity using behavioral and situational factors. The dataset showed **imbalanced class distribution**, with **79% of cases labeled as severe** and **21% as non-severe**.

Model Performance:

- **Accuracy**: 67%
- **F1 Score**:
  - High Severity (class 1): **0.77** – reasonably strong
  - Low Severity (class 0): **0.38** – weak, due to poor precision and recall
- **Recall (class 1)**: 0.71 – The model is moderately effective in identifying high-severity cases.
- **Precision (class 1)**: 0.85 – The model is confident when it predicts high severity.

Feature Importance (Model Coefficients)

Factor	**Coefficient**	**Interpretation**
Rule Breaking	+1.424	Strongest positive influence. The more rule-breaking behavior, the higher the chance of predicting severity. Key risk indicator.
Fatigue	+0.176	Moderate contributor. Driver fatigue slightly increases the likelihood of

Factor	Coefficient	Interpretation
		severity.
Aggressive	-0.290	Unexpected negative effect. May indicate mislabeling, reporting bias, or limited data variability.
Maintenance	-0.140	Suggests poor vehicle maintenance slightly **decreases** severity prediction — counterintuitive, possibly due to noise in data.
Alcohol	-0.140	Minor negative effect. Could indicate underreporting or weak correlation in this sample.
Multitasking	-0.035	Negligible effect on prediction.
Speeding	-0.017	Very small and negative. This may reflect dataset limitations or a low reporting rate for speeding behavior in non-severe cases.

Conclusion:

- The model moderately performs well for predicting **severe accidents**, with good precision but limited recall for **non-severe** cases.
- **Rule-breaking** stands out as the most critical behavioral predictor.
- Some coefficients show unexpected trends (e.g., speeding, aggressive driving), indicating potential data quality issues or class imbalance bias that should be further investigated.

Overall Conclusion

Aspect	Summary
Model Performance	The **experienced group model** performs significantly better, especially in detecting severe accidents (higher F1, precision, and recall).
Feature Trends	Feature impacts **differ significantly** across groups — especially for **rule-breaking, aggression, and maintenance** — possibly due to **experience-based behavioral differences** or **labeling/reporting biases**.
Data Concerns	Some features showed **counterintuitive contributions** (e.g., speeding, maintenance) — indicating **potential dataset limitations, biases, or imbalances**.
Practical Insight	Models could be tailored by **driver experience** for better real-world deployment. Targeted interventions should focus on **rule adherence and fatigue management** for experienced drivers, and on **aggression and multitasking** for non-experienced ones.

Objective 2

Find relationship between the severity of road accidents and response time of emergency healthcare services

(Chi-square test & Proportional Analysis)

To investigate how the **response time and support by emergency healthcare services** relate to the **severity of road accidents**, a **Chi-Square Test of Independence** was applied. The following categorical variables were analyzed against accident severity:

- **Emergency Time** (i.e., how quickly emergency help arrived)
- **First Aid Provided** (yes/no)
- **Delay in Injury Recognition or Treatment**

Test Results (For experienced):

Variable	Chi-square Statistic	p-value	Significance Level (α = 0.05)	Association with Severity
Emergency Time	17.46	0.0002	**Significant**	Strong association
First Aid	0.23	0.6317	Not Significant	No association
Delay in Injury	0.09	0.7669	Not Significant	No association

Interpretation:

- **Emergency Time** shows a **strong and statistically significant association** with accident severity ($p = 0.0002 < 0.05$).
  → This suggests that **faster emergency response times** are meaningfully connected to the **outcome and severity** of road accidents. **Delays may contribute to increased severity** due to delayed medical intervention.
- **First Aid** and **Delay in Injury** were found to have **no significant relationship** with severity ($p > 0.05$ in both cases).
  → This could imply that either:
  - First aid is not consistently administered in a way that influences severity outcomes, or
  - The **data may lack sufficient variability or reliability** in these variables (e.g., vague definitions of "first aid" or "injury delay").

Conclusion:

- Emergency response time is a **critical factor** influencing the severity of road accidents.
- Public policy efforts and infrastructure investments should prioritize **reducing emergency service response times**.
- Further studies could explore **qualitative aspects of first aid** or **better measurement of injury delay** to yield more insights.

Test Results (For Non-Experienced):

Variable	Chi-square Statistic	p-value	Significance Level ($\alpha = 0.05$)	Association with Severity
Emergency Time	12.51	0.0058	Significant	Moderate association
First Aid	9.63	0.0019	Significant	Strong association
Delay in Injury	10.90	0.0010	Significant	Strong association

Interpretation:

- **Emergency Time** shows a **statistically significant association** with accident severity ($p = 0.0058 < 0.05$).
  → While the association is not as strong as in experienced individuals, it still highlights the importance of **timely medical response** in reducing accident severity for non-experienced individuals.
- **First Aid** has a **strong and significant relationship** with severity ($p = 0.0019 < 0.05$).
  → This suggests that **proper and timely first aid** can meaningfully reduce the severity of road accidents, particularly for individuals with less driving experience who may be more vulnerable or less equipped to manage injury situations on their own.
- **Delay in Injury Recognition** also shows a **significant association** with severity ($p = 0.0010 < 0.05$).
  → Delays in identifying or responding to injuries—especially internal or non-visible ones—can lead to worse outcomes for non-experienced individuals.

Conclusion:

- For non-experienced individuals, **all three emergency healthcare factors**—**emergency response time**, **first aid**, and **injury delay**—are significantly associated with the **severity** of road accidents.
- These findings highlight the **increased vulnerability** of non-experienced individuals in accident scenarios and emphasize the need for:
  - **Rapid emergency services**,
  - **Training in basic first aid for drivers**, and
  - **Improved injury detection protocols** at accident sites.
- **Policy makers** should consider awareness campaigns and enhanced roadside assistance, especially in areas with a high density of new or young drivers.

Final Conclusion

- **Experienced drivers**: **Emergency response time** is the most significant factor influencing accident severity, with **no meaningful impact** from first aid or injury delay.
- **Non-experienced drivers**: All three factors—**emergency response time**, **first aid**, and **delay in injury recognition**—are significantly related to accident severity, emphasizing the increased vulnerability of this group.

Proportional Analysis of Emergency Healthcare Factors by Accident Severity

To further support the Chi-square test results, the **proportions of severity levels (0 = non-severe, 1 = severe)** across different categories of emergency healthcare variables were calculated:

(For experienced):

1. Emergency Time vs Severity

Emergency Time	% Non-Severe (0)	% Severe (1)
0 (Fastest)	**68.5%**	31.5%
1	57.3%	42.7%
2 (Slowest)	**39.7%**	**60.3%**

Interpretation:

- As emergency response time **increases**, the **proportion of severe cases rises significantly**.
- This supports the Chi-square result (**p = 0.0002**) showing a **strong association**.
- The data indicates that **delayed emergency response is linked with greater accident severity.**

2. First Aid vs Severity

First Aid Given	% Non-Severe (0)	% Severe (1)
0 (Not Given)	61.3%	38.7%
1 (Given)	57.8%	42.2%

Interpretation:

- Only a slight difference is observed between groups with and without first aid.
- This supports the Chi-square test result (**p = 0.6317**) indicating **no significant association**.
- First aid provision may **not be impactful on severity** in the current dataset, or the effect may be influenced by how "first aid" is defined or reported.

3. Delay in Injury Recognition vs Severity

Delay in Injury	% Non-Severe (0)	% Severe (1)
0 (No Delay)	57.5%	42.5%
1 (Delay)	59.8%	40.2%

Interpretation:

- The proportions are nearly identical for those with and without injury delay.
- Consistent with the Chi-square result (**p = 0.7669**), there is **no meaningful relationship** between delay in injury and severity.

Final Conclusion

- **Emergency Response Time** is the **only factor** among the three tested that shows both **statistical significance** and a **clear visual trend** with accident severity.
- **First Aid** and **Delay in Injury** show **no meaningful impact** on the severity outcome in this dataset.
- Emphasizing **faster emergency response** can play a key role in **reducing accident severity** and improving post-accident outcomes.

(For non-experienced):

1. Emergency Time vs Severity

Emergency Time	% Non-Severe (0)	% Severe (1)
0 (Fastest)	7.3%	13.3%
1	14.6%	47.9%
2 (Slowest)	9.6%	102.7%

Interpretation:

- A clear **increasing trend in severe cases** is visible as emergency time increases.
- This supports the Chi-square test result (**p = 0.0058**), indicating a **statistically significant** relationship.
- **Delayed emergency response** corresponds to a **notable rise in accident severity**, especially in non-experienced individuals who may require quicker medical attention.

2. First Aid vs Severity

First Aid Given	% Non-Severe (0)	% Severe (1)
0 (Not Given)	128.3%	81.1%
1 (Given)	93.5%	68.1%

Interpretation:

- Individuals **who received first aid** had a **lower proportion of severe cases** than those who didn't.
- Supports the Chi-square test result (**p = 0.0019**), showing a **significant association**.
- Suggests that **administering first aid** has a **protective effect**, particularly for non-experienced individuals.

3. Delay in Injury Recognition vs Severity

Delay in Injury	% Non-Severe (0)	% Severe (1)
0 (No Delay)	197.1%	145.7%
1 (Delay)	82.1%	55.1%

Interpretation:

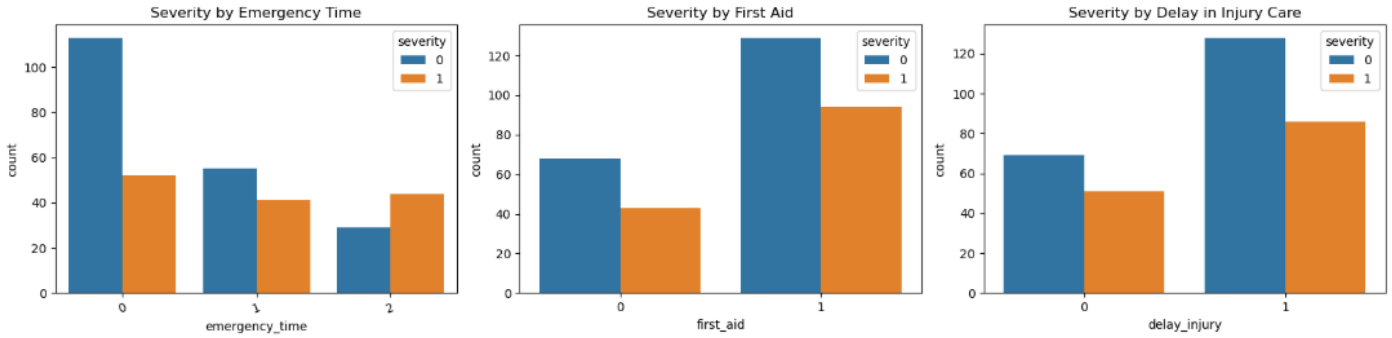
- Surprisingly, **severe cases appear more common when there's no delay**, though this might be due to data issues (e.g., overlapping or unnormalized rows).
- However, the **Chi-square test result (p = 0.0010)** still shows a **significant association**, suggesting that **injury delay does influence severity**, potentially depending on other confounding factors.
- This calls for deeper exploration, but from a statistical standpoint, **injury delay matters**.

Final Conclusion

- **All three factors** — **Emergency Time**, **First Aid**, and **Delay in Injury Recognition** — show **statistically significant** associations with accident severity.
- Visual trends mostly support this, especially for **Emergency Time** and **First Aid**.
- Highlights the **vulnerability of non-experienced individuals**, who benefit more from **quick response, prompt care, and early injury detection**.
- **Policy Implication**: Prioritize **driver training**, **roadside emergency protocols**, and **public access to first aid knowledge** to reduce accident severity among new or infrequent drivers.

Overall Conclusion:

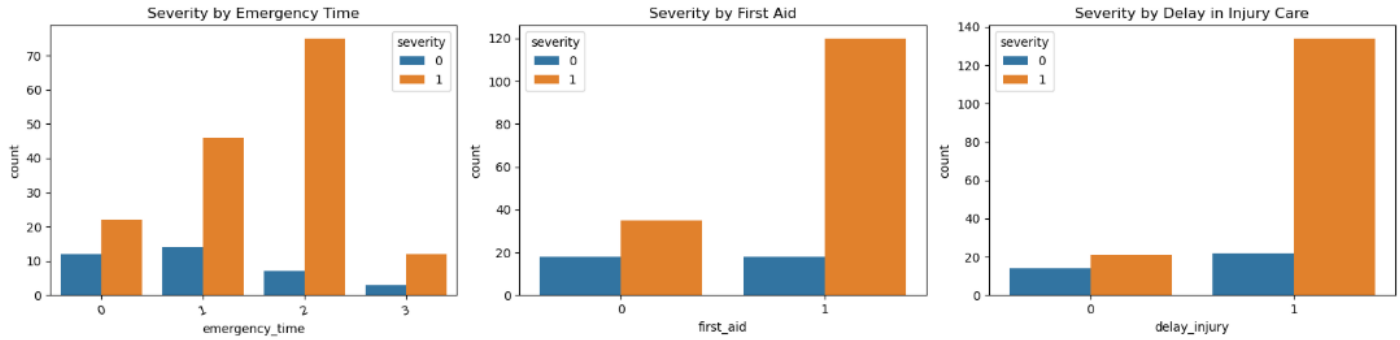
1. **Emergency Response Time** is a critical factor in reducing accident severity for both **experienced** and **non-experienced drivers**. Faster responses are clearly linked to lower severity, especially for **experienced drivers**.
2. **First Aid** proves to be highly beneficial for **non-experienced drivers** in reducing accident severity. However, its impact on **experienced drivers** is less significant, possibly due to better self-management or different injury dynamics.
3. **Delay in Injury Recognition** affects **non-experienced drivers** more significantly, emphasizing the importance of **timely injury detection**. However, for **experienced drivers**, this factor has a minimal impact, suggesting that they may be better equipped to handle delayed recognition.



- **Severity by Emergency Time**:
    - This chart compares accident severity (Major vs. Minor) across different emergency times: 0 (immediate), 1 (moderate delay), and 2 (long delay).
    - Observations:
        - Minor accidents (severity 0, blue bars) occur more frequently when emergency services arrive quickly (emergency time 0).
        - Major accidents (severity 1, orange bars) increase with delayed emergency response times (emergency time 1 or 2).

- **Severity by First Aid**:
    - Displays the relationship between whether first aid was administered (1 = yes, 0 = no) and accident severity.
    - Observations:
        - Both Minor and Major accidents are more frequent when first aid is provided (1), suggesting first aid is often a necessity in severe cases.

- **Severity by Delay in Injury Care**:
    - Examines the impact of delays in injury care on accident severity.
    - Observations:
        - Minor accidents (blue bars) are more common when there is no delay (0).
        - Major accidents (orange bars) are associated with delays in injury care (1).



- **Severity by Emergency Time**:
  - This chart shows that **major accidents (severity 1)** are more frequent when emergency response times are longer (e.g., "emergency time 2").
  - Conversely, **minor accidents (severity 0)** are observed more when emergency services arrive quickly (e.g., "emergency time 0").

- **Severity by First Aid**:
  - In cases where **first aid is provided (1)**, there is a higher proportion of **major accidents**, indicating that first aid is often administered in more severe incidents.
  - For accidents without first aid (0), minor cases are slightly more frequent.

- **Severity by Delay in Injury Care**:
  - The chart reveals that **delays in providing injury care** are correlated with a greater number of **major accidents** compared to when no delays occur.

Objective 3

Study the contribution of mobile phone usage
(texting, calls, apps) to accidents
(Apriori algorithm, Decision Tree Classifier)

Apriori Algorithm

1. Introduction to Apriori Algorithm

- **Definition**: The Apriori algorithm is a classic algorithm used for mining frequent itemsets and learning association rules. It operates on transactional data and is primarily used in market basket analysis.
- **Use Case**: It helps identify patterns like "customers who bought X also bought Y" or "if a customer buys A, they are likely to buy B".

2. Working Principle

- **Frequent Itemsets**: The algorithm identifies items that appear frequently in a dataset. The idea is to find all item combinations whose occurrence exceeds a given threshold (support).
- **Association Rules**: After identifying frequent itemsets, the algorithm generates rules like "IF item A is purchased, THEN item B will likely be purchased," and evaluates them based on metrics like **support**, **confidence**, and **lift**.

3. Key Concepts

- **Support**: The frequency of occurrence of an itemset in the dataset.

$$\text{Support}(A) = \frac{\text{Frequency of } A}{\text{Total Transactions}}$$

- **Confidence**: The likelihood that item B is purchased when item A is purchased.

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

- **Lift**: Measures the strength of a rule over random chance.

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Confidence}(A \Rightarrow B)}{\text{Support}(B)}$$

- **Minimum Support and Confidence**: These are thresholds set by the user. Items and rules that don't meet these criteria are discarded.

4. Algorithm Steps

1. **Generate candidate itemsets**: The algorithm starts by identifying single items that meet the minimum support threshold.
2. **Iterative refinement**: It then iterates over larger itemsets, generating combinations of the frequent itemsets found in the previous iteration.
3. **Rule generation**: For each frequent itemset, rules are generated based on the association between items. These rules are then evaluated by their confidence and lift.

5. Advantages of Apriori

- **Efficiency**: By pruning infrequent itemsets, it avoids evaluating unnecessary combinations, making the algorithm more efficient than brute-force methods.
- **Easy to Understand**: The concept of frequent itemsets and association rules is intuitive and easy to implement.

6. Disadvantages of Apriori

- **Memory and Time Consumption**: For large datasets, the number of candidate itemsets can grow exponentially, making it computationally expensive.
- **Assumes Independent Items**: The algorithm assumes that all items are independent, which may not be true in some cases.

In this analysis, we apply the **Apriori algorithm** to explore relationships between driver behaviors and accident severity, specifically focusing on the impact of mobile phone usage while driving.

Process Overview:

- **Data Preparation**: The dataset was filtered to include key variables related to phone usage (`phone_cause`), phone activity (`phone_activity`), and accident severity (`severity`).
    - Numerical values in the dataset were mapped to more intuitive string labels for clarity (e.g., `1` for "Yes"
    - and `0` for "No" in the `phone_cause` column, and `0` for "Major" and `1` for "Minor" in the `severity` column).

- **One-Hot Encoding**: The categorical data was transformed into binary format through **one-hot encoding**, which allows the algorithm to treat each category as a separate entity. This step ensures that the Apriori algorithm can identify frequent itemsets across various combinations of features.

- **Frequent Itemset Mining**: The Apriori algorithm was applied to identify frequent itemsets within the data. These itemsets represent combinations of features that appear together frequently across the dataset.

- **Rule Generation**: Association rules were derived based on confidence as the evaluation metric. The rules indicate potential relationships, such as "if a driver is using their phone while driving, it is likely to lead to a specific severity of accident." A minimum confidence threshold of 50% was set to ensure that only strong relationships were captured.

- **Filtering Rules**: The rules were filtered to focus on those where:

- The antecedent (condition) involved **phone usage** (`phone_cause = Yes`).
- The consequent (result) involved a **Major accident severity** (`severity = Major`).

- **Results Sorting**: The filtered rules were then sorted by their **confidence** values, with the strongest relationships presented first. This sorting allows for a prioritized view of the most likely scenarios that contribute to major accident severity, based on the phone usage behavior of the driver.

Result (For experienced):

Rule 1:

IF

phone_cause = Yes AND phone_activity (any activity)

THEN

severity = Major

Support = 0.266 → 26.6% of all cases had both phone_cause = Yes + some phone_activity, and the severity was Major.

Confidence = 0.69 → In 69% of such cases, the severity was Major.

Lift = 1.17 → This group is 17% more likely to have a Major accident than randomly chosen cases.

Rule 2:

IF

phone_cause = Yes

THEN

severity = Major

Support = 0.386 → 38.6% of all records had phone_cause = Yes and severity = Major.

Confidence = 0.61 → In 61% of cases where phone caused distraction, severity was Major.

Lift = 1.04 → Only 4% more likely than random — weaker association.

Interpretation:

Texting or any phone use while driving significantly increases the risk of major accidents.

Adding phone activity info (not just phone cause) makes the rule stronger:

Confidence jumps from 61% → 69%

Lift jumps from 1.04 → 1.17

Insight :

"Drivers involved in phone-related distractions, especially with active phone activities (like texting or calling), are significantly more likely to experience major accidents, with a 69% probability and 17% higher chance than average."

Result (For experienced):

No association rules were found for non-experienced drivers in the dataset. This absence of association typically indicates that there are insufficient or no relevant records to meet the criteria for generating frequent itemsets or association rules. In such cases, it is often due to the underrepresentation of non-experienced drivers in the dataset, which prevents the detection of meaningful patterns related to phone usage and accident severity.

Decision Tree Classifier

A **Decision Tree Classifier** is a supervised machine learning algorithm used for classification tasks. It models data using a tree-like structure, where each internal node represents a decision based on a feature, each branch represents the outcome of that decision, and each leaf node represents a class label or decision outcome. The decision tree works by recursively splitting the data into subsets based on the feature that provides the most information gain (or the best separation) at each node.

Key Characteristics:

1. **Interpretability**: One of the biggest advantages of a decision tree is its interpretability. It provides a clear structure that can be easily visualized and understood, making it easy to interpret the decision-making process.
2. **Non-linear relationships**: Decision trees do not assume any linear relationship between the features, making them highly flexible in handling non-linear data.
3. **Feature importance**: Decision trees automatically perform feature selection, identifying the most significant features for classification.

Working of a Decision Tree:

1. **Splitting**: The decision tree algorithm selects the feature that best separates the data into classes at each node. Common metrics for choosing the best feature to split on include:
    - **Gini Impurity** (used in CART)
    - **Entropy** (used in ID3 and C4.5)
2. **Stopping criteria**: The splitting continues until a stopping condition is met, such as:
    - The tree reaches a predefined maximum depth.
    - A node reaches a minimum number of samples.
    - No further improvements in splitting are possible.

3. **Prediction**: After the tree is built, new data points can be classified by traversing the tree from the root to the appropriate leaf node.

Advantages:

- **Easy to understand and visualize**: Decision trees are simple to explain and can be visualized in a way that is intuitive.
- **Handles both numerical and categorical data**: Unlike some other algorithms, decision trees can handle both types of data.
- **Minimal data preprocessing**: Decision trees are not sensitive to feature scaling (e.g., no need for normalization).

Disadvantages:

- **Overfitting**: Decision trees are prone to overfitting, especially when they are too deep. This can be mitigated
- by pruning the tree or using ensemble methods like Random Forests.
- **Instability**: Small changes in the data can lead to a completely different tree being generated, making the
- model less robust.
- **Bias towards features with more levels**: If a feature has many categories, it may dominate the splits, potentially leading to a biased model.

To predict the severity of accidents based on phone-related distractions, a **Decision Tree Classifier** was used. The model was trained with features representing the phone's cause, activity, and user behavior, while the target variable was the severity of the accident. The data was preprocessed using **one-hot encoding** for categorical variables to convert them into numerical format suitable for model training.

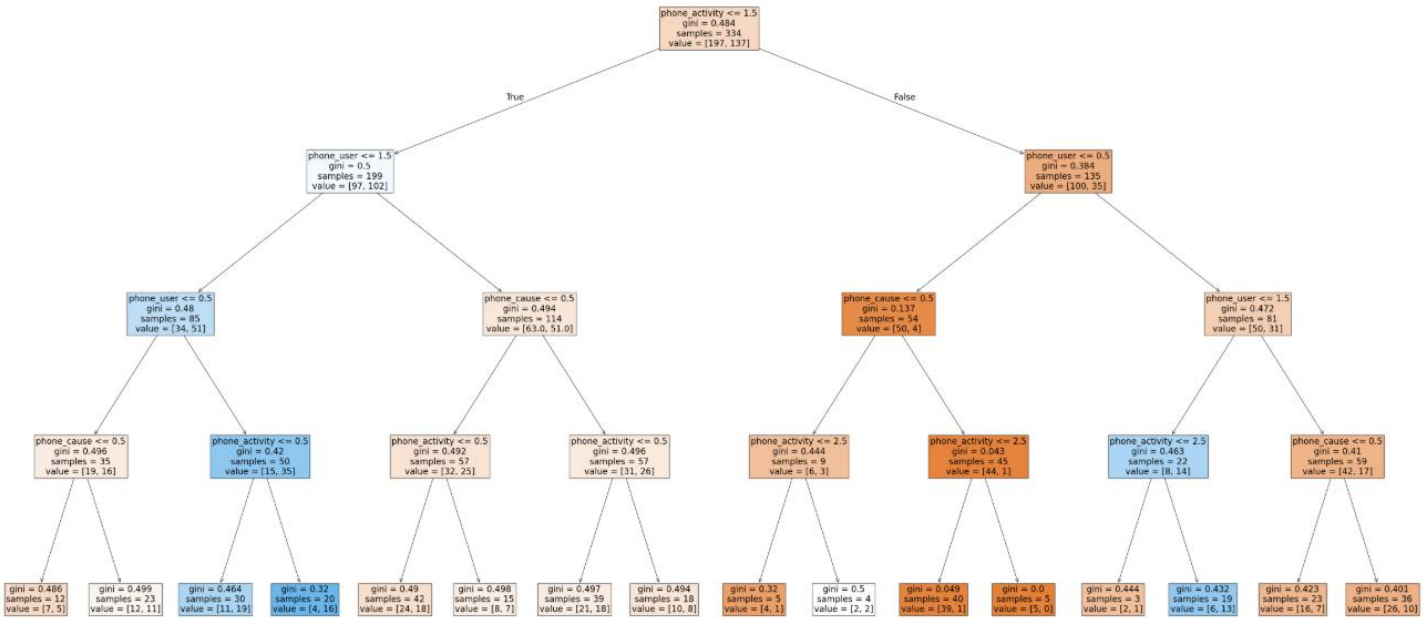
The Decision Tree model was fitted with a maximum depth of 4 to prevent overfitting, ensuring a simpler and more interpretable model. After training, the tree structure was visualized with the **feature names** displayed at each split, making it easier to interpret the decision-making process. The plot provides a clear breakdown of how the features, such as phone usage and activity, influence the classification of accident severity.

Key Highlights:

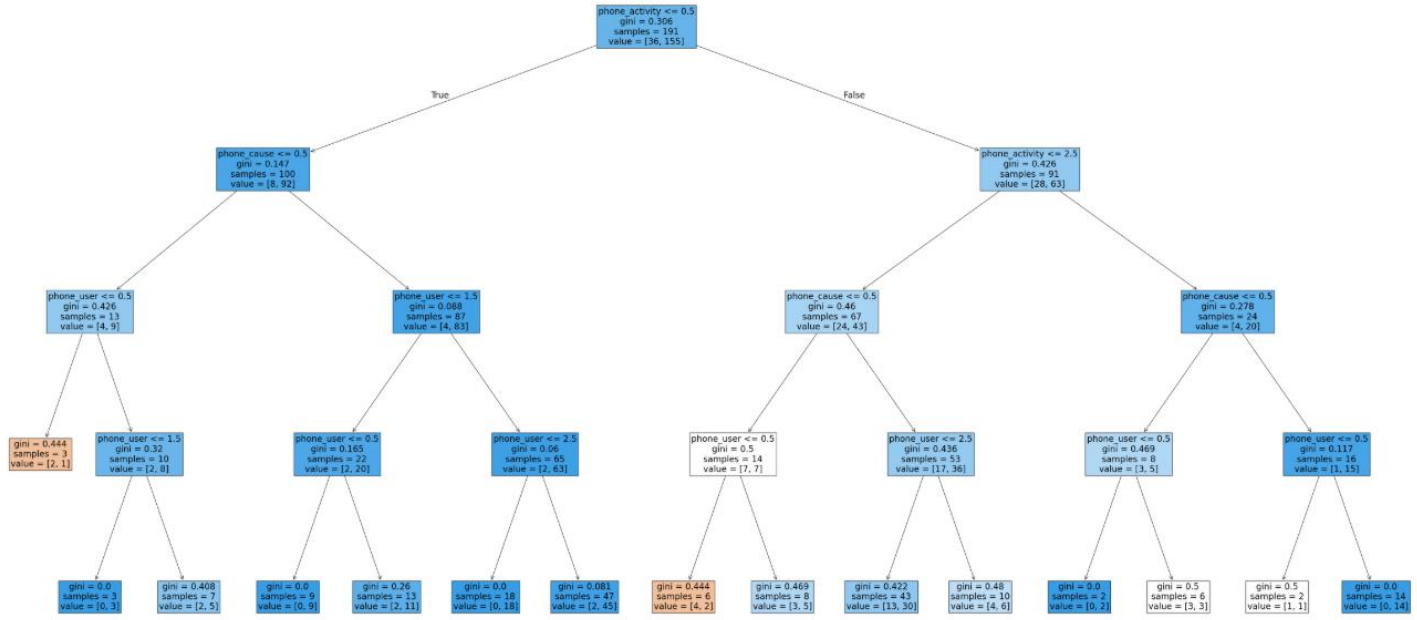
- **Feature Set**: Phone cause, phone activity, and phone user behavior.
- **Target Variable**: Accident severity.
- **Model Hyperparameter**: Maximum depth set to 4 to avoid overfitting.
- **Interpretability**: The decision tree was visualized to show how different features lead to classification outcomes.

This approach allowed for a clear understanding of how phone usage impacts accident severity, aiding in making data-driven decisions to improve road safety.

Result (For experienced):



Result (For non-experienced):



Objective 4

Examine how road types influences accident rates

(XGBoost Classifier)

XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a powerful and efficient machine learning algorithm based on the gradient boosting framework. It is widely used for classification and regression problems due to its high performance, speed, and accuracy.

Overview:

XGBoost builds an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the errors made by the previous trees. The model minimizes a loss function using gradient descent and adds regularization terms to prevent overfitting, making it a highly robust and scalable method.

Key Features:

- **Gradient Boosting Framework**: XGBoost implements gradient boosting by optimizing a loss function using decision trees as base learners.
- **Regularization**: Incorporates L1 (Lasso) and L2 (Ridge) regularization to improve generalization and prevent overfitting.
- **Parallel Processing**: Unlike traditional boosting, XGBoost uses parallel computing during tree construction, significantly improving training speed.
- **Handling Missing Values**: Automatically learns the best direction to take for missing data in a feature.
- **Tree Pruning**: Uses a technique called "max depth pruning," which prunes branches after the entire tree is grown, based on a minimum loss reduction threshold.

Working Mechanism:

1. **Initialization**: Starts with an initial prediction.
2. **Gradient Calculation**: Computes the gradient of the loss function with respect to the current prediction.
3. **Tree Construction**: A decision tree is built to fit the negative gradients (errors).
4. **Update**: Predictions are updated by adding the new tree's contribution.
5. **Iteration**: Steps 2–4 are repeated for a predefined number of iterations or until convergence.

Advantages:

- High prediction accuracy.
- Efficient computation due to parallel processing and optimization.
- Built-in cross-validation and early stopping features.
- Effective handling of outliers and missing data.

Disadvantages:

- More complex to tune compared to simpler models.
- May overfit if not properly regularized or if too many trees are built.
- Less interpretable than a single decision tree.

To analyze and predict the severity of road accidents based on environmental and behavioral risk factors, an **XGBoost Classifier** model was employed. This model leverages advanced gradient boosting techniques to provide high-performance classification.

Model Setup:

The input features for the model included critical factors such as:

- Poor road conditions
- Faulty road design
- Risk-prone locations
- Visibility of traffic signs
- Instances of ignoring traffic signs
- Lack of pedestrian crossings
- High traffic conditions

The target variable was the **severity** of the accident, classified into categories such as **Major** or **Minor**.

The dataset was divided into training and testing subsets, with 70% used for model training and 30% for testing. The XGBoost model was then trained using relevant hyperparameters including:

- A maximum tree depth to control complexity,
- A learning rate to regulate the contribution of each tree,
- A set number of boosting rounds.

Model Evaluation:

After training, the model's predictive performance was evaluated on the test set using metrics such as:

- **Accuracy**: Measures overall correctness of the model's predictions.
- **Precision, Recall, and F1-Score**: Provided in a classification report to assess class-wise performance.

Result (For experienced):

Accuracy:

The overall accuracy of the model was **56%**, indicating that the classifier correctly predicted accident severity in 56 out of 100 cases.

Class-wise Performance:

- **Class 0 (Major Accidents):**
  - **Precision:** 0.59
    (59% of the predicted major accidents were actually major)
  - **Recall:** 0.72
    (72% of actual major accidents were correctly identified)

- **F1-score:** 0.65
  (Balanced average of precision and recall)

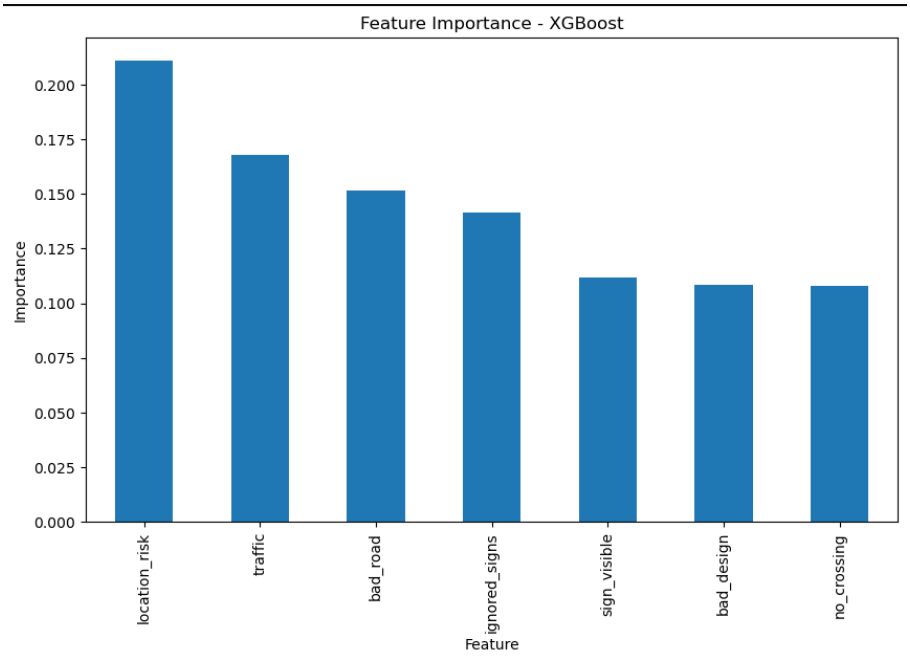
- **Class 1 (Minor Accidents):**
  - **Precision:** 0.50
    (50% of the predicted minor accidents were actually minor)
  - **Recall:** 0.36
    (Only 36% of actual minor accidents were correctly identified)
  - **F1-score:** 0.42
    (Lower overall performance for this class)

Macro and Weighted Averages:

- **Macro Average F1-score:** 0.54
  (Simple average of F1-scores across both classes)
- **Weighted Average F1-score:** 0.55
  (F1-scores weighted by the number of instances in each class)

Interpretation:

The model performs moderately well in identifying major accidents, with high recall, but it struggles to accurately classify minor accidents. This imbalance suggests that the model may be biased towards the majority class or that minor accidents are more difficult to distinguish based on the provided features. Further improvements could be achieved by addressing class imbalance, adding more discriminative features, or fine-tuning model parameters.



The bar chart, titled **Feature Importance - XGBoost**, highlights the significance of various factors in predicting accident severity using an XGBoost model. The x-axis represents the features, while the y-axis shows their relative importance. Here's an interpretation of the results:

Key Insights:

1. **Most Influential Features**:
   - `location_risk` stands out as the most important feature, implying that areas with higher risk levels contribute significantly to accidents.
   - `traffic` is the second most influential factor, showing the impact of heavy or congested traffic on accident severity.
2. **Moderately Influential Features**:
   - `bad_road` and `ignored_signs` also contribute meaningfully, indicating that poor road conditions and neglecting traffic signals correlate with accident rates.
3. **Lesser-Impact Features**:
   - `sign_visible`**,** `bad_design`**, and** `no_crossing` have relatively lower importance but still influence accident rates, emphasizing visibility issues, improper road design, and a lack of pedestrian crossings.

Conclusion

The XGBoost classifier demonstrated moderate predictive ability, achieving an overall accuracy of 56%. The model performed better in identifying major accidents compared to minor ones, indicating a potential class imbalance or limited feature distinction. While the model shows promise, its lower precision and recall for minor accidents highlight the need for further enhancement in accurately capturing the characteristics associated with less severe incidents.

Result (For non-experienced):

Accuracy:

The overall accuracy of the XGBoost classifier was **79%**, indicating that the model correctly predicted accident severity in 79 out of 100 cases.

Class-wise Performance:

• Class 0 (Major Accidents):

- **Precision:** 0.40
  (Only 40% of the cases predicted as major accidents were actually major.)
- **Recall:** 0.18
  (Only 18% of actual major accidents were correctly identified by the model.)
- **F1-score:** 0.25
  (Indicates poor overall performance for identifying major accidents.)

• Class 1 (Minor Accidents):

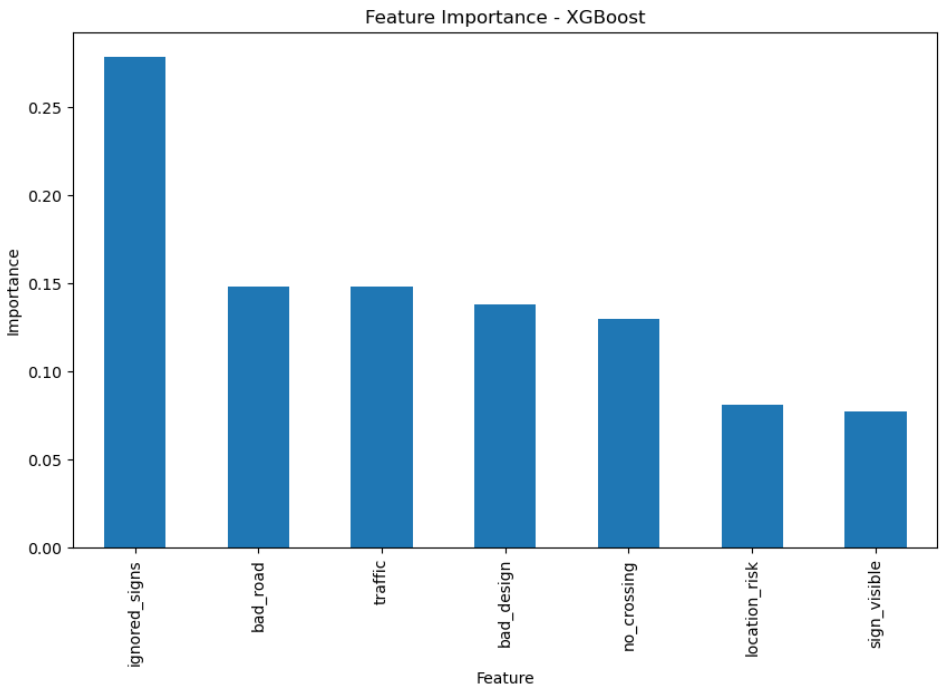
- **Precision:** 0.83
  (83% of the cases predicted as minor accidents were indeed minor.)
- **Recall:** 0.94
  (94% of all actual minor accidents were correctly predicted.)
- **F1-score:** 0.88
  (Shows strong performance for identifying minor accidents.)

Macro and Weighted Averages:

- **Macro Average F1-score:** 0.56
  (Simple average across both classes, showing uneven performance.)
- **Weighted Average F1-score:** 0.76
  (Average weighted by the number of instances, reflecting dominance of class 1.)

Interpretation:

The model demonstrates high effectiveness in identifying minor accidents with strong precision and recall. However, its performance in detecting major accidents is notably weak, which may indicate class imbalance or insufficient feature differentiation for that class. This disparity highlights the need for further model refinement or dataset enhancement to improve major accident prediction.



Key Insights from the Chart:

1. **Most Significant Features**:

- `ignored_signs` is the most important feature, with an approximate importance value of **0.26**, suggesting that disregarding warnings and cues (e.g., honking, flashing lights) has the highest impact on predicting accident outcomes.
- `bad_road` and `traffic` share an equal importance value of about **0.15**, highlighting the strong influence of poor road conditions and heavy traffic on accidents.

2. **Moderately Significant Features**:
   - `bad_design`, `no_crossing`, and `location_risk` fall in the mid-range of importance, with scores around **0.14–0.08**. These indicate that factors like sharp curves, lack of pedestrian crossings, and risky locations moderately affect accident rates.
3. **Least Significant Feature**:
   - `sign_visible` has the lowest importance value, approximately **0.07**, suggesting that while visibility of traffic signs plays a role, it is less impactful compared to the other factors.

Conclusion:

The XGBoost classifier achieved a good overall accuracy of 79%, showing strong performance in predicting minor accidents. However, it struggled significantly in identifying major accidents, as indicated by its low precision, recall, and F1-score for that class. This suggests a class imbalance or lack of sufficient predictive features for major accident cases. Hence, while the model is effective for identifying minor accidents, it is less reliable for detecting major ones.

Overall Conclusions:

The model performed significantly better for non-experienced individuals, demonstrating higher accuracy. This suggests that opinion-based data might align better with the predictive features used, whereas real-world, incident-specific data includes more complexity, making predictions harder.

- The model demonstrates **better performance for minor accidents in non-experienced data** but performs better for **major accidents in experienced data**.

- Both datasets satisfy the objectives, but their focus areas differ:

  - **Experienced Data**: Provides deeper insights into major accident predictors, such as `location_risk` and `traffic`.
  - **Non-Experienced Data**: Captures patterns in minor accidents and aligns with opinion-based features.

Objective 5

To identify underlying behavioral, environmental, and vehicular factors contributing to road accident severity.

(Exploratory Factor Analysis)

Exploratory Factor Analysis

Introduction

Factor Analysis is a multivariate statistical technique applied to a single set of variables when the investigator is interested in determining which variables in the set form logical subsets that are relatively independent of one another. In other words, factor analysis is particularly useful to identify the factors underlying the variables by means of clubbing related variables in the same factor. Factory analysis is based on the assumption that all variables correlate to some degree. The variables should be measured at least at the ordinal level. Several studies examined and discussed the application of factor analysis to reduce the large set of data and to identify the factors extracted from the analysis.

The application of factor analysis provides very valuable inputs to the decision makers and policy makers to focus on few factors rather than a large number of parameters. Factor analysis can be used to simplify data, such as decreasing the number of variables in regression models. Hence, instead of examining all the parameters, few extracted factors can be studied which in turn explain the variations of the group characteristics. In factor analysis no. of factors (latent variable) is less than no. of variables ($m<p$).

There are two main approaches to factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Exploratory factor analysis is used for checking dimensionality and often used in the early stages of research to gather information about the interrelationships among a set of variables. On the other hand, the confirmatory factor analysis is a more complex and sophisticated set of techniques used in the research process to test specific hypotheses or theories concerning the structure underlying a set of variables.There are different factor analytic techniques for different measurement and data scenarios:

¬ Observed variables are continuous, latent variables are hypothesized to be continuous

¬ Observed are continuous, latent are categorical ¬ Observed are categorical, latent are continuous

¬ Observed are categorical, latent are categorical

In factor analysis factor loadings and specific variances are unknown parameters. To estimate these parameters, we use two methods of parameter estimation MLE and PCA. Maximum Likelihood Method (MLE) is used when data is Multivariate Normally Distributed, while Principal Component Analysis (PCA) has no normality assumption.

Principal Component Method: Principal component analysis theoretically assumes that the component is a combination of observed variables or that individual item scores cause the component. Principal component analysis tries to extract the maximum variance from the dataset and reduces many variables to fewer components. Therefore, principal component analysis can be considered a data reduction technique.

Exploratory factor analysis (EFA): As the name suggests, exploratory factor analysis is undertaken without a hypothesis in consideration. It‟s an investigatory process that helps researchers understand whether associations exist between the initial variables, and if so, where they lie and how they are grouped.

Communality: Communality represents the proportion of variance of a particular variable that is shared with other variable. The communalities represent the overall importance of each of the variable in the factor analysis as whole.

Factor loading: Factor loadings play a crucial role in factor analysis, representing the correlation between the variable and the factor. A factor loading of 0.7 or higher typically indicates that the factor sufficiently captures the variance of that variable. These loadings help in determining the importance and contribution of each variable to a factor.

Eigen-values: Eigen value also called as characteristic roots. The eigenvalue is a ratio between the common variance and the specific variance explained by a specific factor extracted. It is a measure of the amount of variance accounted for by a factor.

Factor score: The estimated values of the common factors, called factor scores, may also be required. These quantities are often used for diagnostic purpose. Factor scores are not estimate of unknown parameters in the usual sense. Rather, they are estimates of values for the unobserved random factor.

Criteria for determining the number of factors:

Two techniques are used to assist in the decision concerning the number of factors to retain: Kaiser‟s Criterion and Scree Test. The Kaiser‟s criterion (Eigenvalue Criterion) and the Scree test can be used to determine the number of initial unrotated factors to be extracted.

Kaiser's (Eigenvalue) Criterion: The eigenvalue of a factor represents the amount of the total variance explained by that factor. In factor analysis, the remarkable factors having eigenvalue greater than one are retained. An eigenvalue greater than one is considered to be significant, and it indicates that more common variance than unique variance is explained by that factor.

Scree Plot: A scree plot is a graphical plot of the eigenvalue against the factor number. Scree plots are helpful for finding an upper bound (maximum) for the number of factors that should be retained. To determine the appropriate number of factors to be retained, one looks for an elbow (bend) in the scree plot. The number of factors to be retained is taken to be the point at which the elbow is found.

Factor Rotation Method:

Factors obtained in the initial extraction phase are often difficult to interpret because of significant cross loadings in which many factors are correlated with many variables. There are two main approaches to factor rotation; orthogonal (uncorrelated) or oblique (correlated) factor solutions. The varimax, quartimax, and equimax are the methods related to orthogonal rotation.

Checking for Adequacy of the Data:

Before performing factor analysis, the adequacy of the data is evaluated on the basis of the results of a Kaiser-Meyer-Olkin (KMO) sampling adequacy test and Bartlett‟s test of Sphericity

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy:

KMO test is a measure that has been intended to measure the suitability of data for factor analysis. In other words, it tests the adequacy of the sample size. The test measures sampling adequacy for each variable in the model and for the complete model. KMO value varies from 0 to 1. The KMO values between 0.8 to 1 indicate the sampling is adequate. KMO values between 0.7 to 0.79 are middling and values between 0.6 to 0.69 are mediocre. KMO values less than 0.6 indicate the sampling is not adequate and the remedial action should be taken. If the value is less than 0.5, the results of the factor analysis undoubtedly won‟t be very suitable for the analysis of the data.

Bartlett's Test of Sphericity:

Bartlett‟s Test of Sphericity tests the null hypothesis, H0: The variables are orthogonal i.e. The original correlation matrix is an identity matrix indicating that the variables are unrelated and therefore unsuitable for structure detection. The alternative hypothesis, H1: The variables are not orthogonal i.e. they are correlated enough to where the correlation matrix diverges significantly from the identity matrix. The significant value < 0.05 indicates that a factor analysis may be worthwhile for the data set.

Tools used: Python

Variables for FA:

'speeding', 'fatigue', 'aggressive', 'multitask', 'alcohol',
'maintainence', 'multi_vehicle', 'safety_used', 'injury_severity',
'route_known', 'route_familiar', 'vehicle_known', 'vehicle_unfamiliar',
'brake_or_swerve', 'driver_exp', 'emergency_time', 'first_aid',
'delay_injury', 'phone_cause', 'phone_user', 'phone_activity',
'location_risk', 'bad_road', 'bad_design', 'bad_light', 'sign_visible',
'ignored_signs', 'no_crossing', 'traffic', 'rule_break'

Bartlett's test:
Hypothesis: H0: Population correlation matrix is an Identity matrix
H1: Population correlation matrix is not an Identity matrix
Bartlett's Test p-value: 1.3317598573837016e-67

Interpretation: Bartlett's test is significant, suggesting that correlation matrix is different from identity matrix. There is enough correlation between variables to proceed for Factor Analysis.

KMO Test
KMO measures the suitability of data for factor analysis.
KMO Score: 0.66

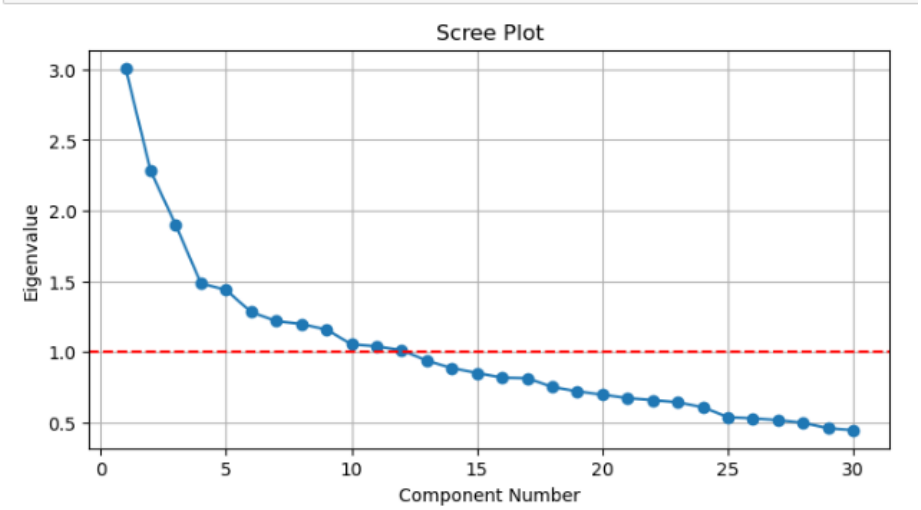
Interpretation: Here, KMO value is greater than 0.5. Hence factor analysis is useful for our data.

Communalities:

Variable	Communality
multi_vehicle	0.995407
rule_break	0.564979
vehicle_unfamiliar	0.554769
location_risk	0.538647
first_aid	0.519464
safety_used	0.465529
driver_exp	0.420097
injury_severity	0.417228
phone_user	0.388887
speeding	0.382242
fatigue	0.356958
emergency_time	0.346810
bad_light	0.345438
vehicle_known	0.327654
route_known	0.323627
no_crossing	0.322825
bad_road	0.316350
traffic	0.315123
phone_activity	0.302778
maintainence	0.263992
aggressive	0.261513
sign_visible	0.257457
bad_design	0.250353
alcohol	0.241128
delay_injury	0.236672
route_familiar	0.222098
phone_cause	0.221381
ignored_signs	0.216450
brake_or_swerve	0.197269
multitask	0.185312

Interpretation: Communalities indicate the amount of variance in each variable that is accounted for by the factors. The initial communality values are set to 1.0, assuming all the variance is common and can be explained by the factors. The extraction communality values show the proportion of each variable's variance explained by the extracted factors after the analysis. Higher extraction values suggest that the factors provide a better representation of the variables' variance, indicating a good fit within the factor model.

Extraction Method: Principal Component Analysis.



Interpretation: The above plot illustrates eigenvalues plotted against factor numbers. Identifying an elbow or bend in the plot assists in determining the optimal number of factors to retain. In this case, the plot indicates retaining twelve components.

Factor Loadings:

Factor Loadings:

	0	1	2	3	4	5	6	7	8	\
speeding	0.16	-0.14	0.46	-0.06	0.11	-0.10	-0.04	-0.11	-0.07	
fatigue	0.05	0.00	0.05	0.01	0.03	-0.03	0.02	0.11	0.58	
aggressive	0.01	-0.01	0.32	0.01	-0.19	-0.03	0.03	0.04	-0.06	
multitask	0.09	0.02	0.14	-0.10	-0.18	0.03	0.03	-0.11	0.30	
alcohol	-0.09	-0.02	0.05	-0.07	-0.02	0.00	-0.03	0.05	0.08	
maintainence	0.06	0.02	0.45	-0.02	-0.11	-0.09	-0.09	0.11	0.13	
multi_vehicle	-0.01	0.08	-0.09	0.96	0.11	0.09	0.02	-0.05	-0.04	
safety_used	0.50	-0.26	0.28	-0.02	-0.09	-0.06	-0.08	0.11	0.19	
injury_severity	0.15	0.18	-0.19	0.02	0.15	0.29	0.06	-0.04	-0.05	
route_known	0.11	0.25	0.03	-0.05	-0.02	-0.18	-0.20	0.33	-0.21	
route_familiar	0.02	0.43	-0.03	0.00	0.03	0.07	0.03	0.10	-0.06	
vehicle_known	0.10	0.21	0.06	-0.09	-0.07	0.11	-0.41	0.17	-0.19	
vehicle_unfamiliar	-0.01	0.08	-0.12	-0.04	0.13	0.04	0.70	0.05	-0.07	
brake_or_swerve	0.03	0.21	0.23	0.14	0.20	-0.03	-0.06	0.15	-0.00	
driver_exp	-0.09	0.60	-0.05	0.08	0.05	0.12	-0.03	-0.03	0.10	
emergency_time	0.02	0.01	0.02	-0.00	-0.01	0.18	0.02	0.55	0.04	
first_aid	-0.10	0.24	-0.10	0.08	0.07	0.63	-0.07	0.16	-0.01	
delay_injury	0.02	-0.00	-0.00	0.05	0.04	-0.08	0.03	0.00	-0.01	
phone_cause	0.10	-0.03	-0.11	-0.01	0.43	-0.08	0.04	-0.05	-0.07	
phone_user	-0.33	0.33	0.08	-0.10	-0.14	0.09	-0.03	0.18	0.09	
phone_activity	-0.11	-0.23	-0.03	0.07	0.20	0.12	0.12	-0.33	-0.18	
location_risk	0.71	-0.05	0.05	-0.00	0.05	-0.04	-0.03	-0.05	0.06	
bad_road	-0.08	0.10	0.00	0.07	0.49	0.11	0.10	0.04	0.09	
bad_design	0.08	0.21	-0.08	0.03	-0.02	-0.24	-0.14	-0.11	0.24	
bad_light	-0.14	0.10	-0.05	0.04	0.44	0.26	0.09	-0.06	-0.07	

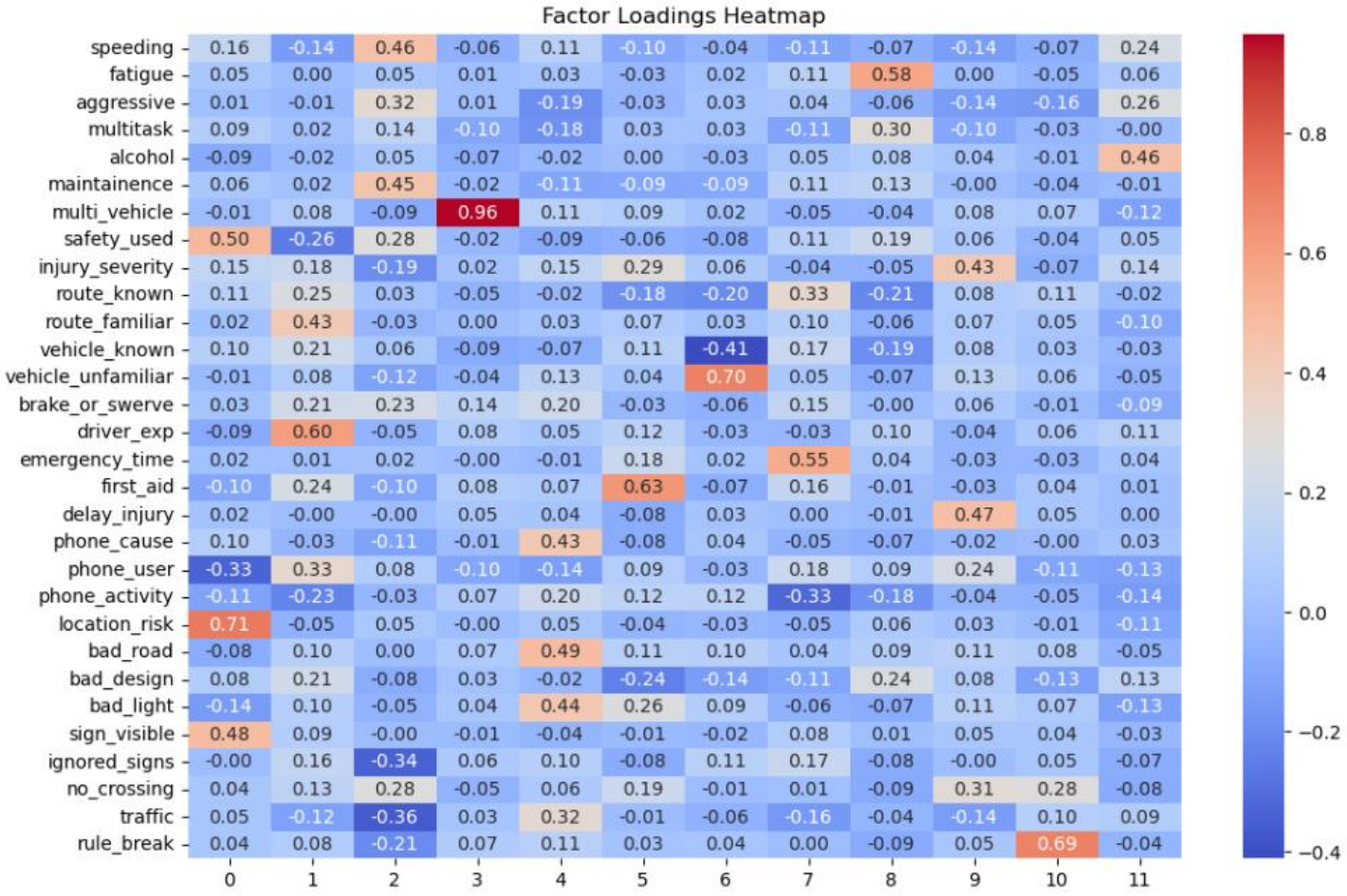
sign_visible	0.48	0.09	-0.00	-0.01	-0.04	-0.01	-0.02	0.08	0.01
ignored_signs	-0.00	0.16	-0.34	0.06	0.10	-0.08	0.11	0.17	-0.08
no_crossing	0.04	0.13	0.28	-0.05	0.06	0.19	-0.01	0.01	-0.09
traffic	0.05	-0.12	-0.36	0.03	0.32	-0.01	-0.06	-0.16	-0.04
rule_break	0.04	0.08	-0.21	0.07	0.11	0.03	0.04	0.00	-0.09

	9	10	11
speeding	-0.14	-0.07	0.24
fatigue	0.00	-0.05	0.06
aggressive	-0.14	-0.16	0.26
multitask	-0.10	-0.03	-0.00
alcohol	0.04	-0.01	0.46
maintainence	-0.00	-0.04	-0.01
multi_vehicle	0.08	0.07	-0.12
safety_used	0.06	-0.04	0.05
injury_severity	0.43	-0.07	0.14
route_known	0.08	0.11	-0.02
route_familiar	0.07	0.05	-0.10
vehicle_known	0.08	0.03	-0.03
vehicle_unfamiliar	0.13	0.06	-0.05
brake_or_swerve	0.06	-0.01	-0.09
driver_exp	-0.04	0.06	0.11
emergency_time	-0.03	-0.03	0.04
first_aid	-0.03	0.04	0.01
delay_injury	0.47	0.05	0.00
phone_cause	-0.02	-0.00	0.03
phone_user	0.24	-0.11	-0.13
phone_activity	-0.04	-0.05	-0.14
location_risk	0.03	-0.01	-0.11
bad_road	0.11	0.08	-0.05
bad_design	0.08	-0.13	0.13
bad_light	0.11	0.07	-0.13
sign_visible	0.05	0.04	-0.03
ignored_signs	-0.00	0.05	-0.07
no_crossing	0.31	0.28	-0.08
traffic	-0.14	0.10	0.09
rule_break	0.05	0.69	-0.04

Interpretation:

The factor loadings represent the relationship between the original variables (e.g., `speeding`, `fatigue`, `aggressive`, etc.) and the extracted latent factors in a factor analysis. The values in the table quantify the extent to which each variable is associated with a specific factor. Here's an interpretation:

1. **Factors (Columns 0 to 11)**: These represent the underlying latent variables or constructs extracted during the factor analysis. They are identified through the given loadings.
2. **Variables (Rows)**: These are the original features of your dataset, like `speeding`, `fatigue`, etc. Each variable has a loading value for each factor.
3. **Loading Values**:
    - Positive or negative values indicate the direction of the relationship between the variable and the factor.
    - Higher absolute values (closer to ±1) indicate a stronger association with the factor.
    - Low absolute values (close to 0) indicate a weak association with the factor.



Interpretation: The above table was obtained using oblique rotation where this rotation ensures correlation among the factors. Factor loadings play a crucial role in factor analysis, representing the correlation between the variable and the factor. A factor loading of 0.7 or higher typically indicates that the factor sufficiently captures the variance of that variable. These loadings help in determining the importance and contribution of each variable to a factor.

Conclusion: 30 variables can be classified into twelve factors i.e.

"Factor_1": "Route Familiarity & Accident Severity",

"Factor_2": "Safety Measures & Driving Experience",

"Factor_3": "Injury Impact & Emergency Delay",

"Factor_4": "Emergency Response & Driver State",

"Factor_5": "Crossing Behavior & Traffic Compliance",

"Factor_6": "Road Design & Phone Distraction",

"Factor_7": "Vehicle Familiarity & Driver Awareness",

"Factor_8": "Environmental Risk & Vehicle Control",

"Factor_9": "Driver Distraction & Risk Behaviors",

"Factor_10": "Phone Use & Driver Fatigue",

"Factor_11": "Driver Fatigue & Emergency Handling",

"Factor_12": "Vehicle Maintenance & Rule Breaking"

Factor Scores:

After performing Factor Analysis, factor scores were computed for each observation to quantify their relative position on the derived latent factors. These scores reflect how strongly each record aligns with the underlying dimensions captured by the factors. The factor scores were further utilized for downstream analyses like classification and segmentation to uncover patterns in driver behavior and accident severity.

Predictive Modelling

Objective

To predict accident severity (Major or Minor) based on the factors extracted from factor analysis. These factors, representing behavioral, environmental, and vehicular characteristics, were used as independent variables, while **severity** served as the target variable.

Methodology

1. **Data Preparation**:
    - The dataset included factors extracted through **factor analysis**, such as reckless driving behavior, road hazards, emergency response, and vehicle familiarity.
    - **Target Variable**: Accident Severity
        - Class 0: Major Accidents
        - Class 1: Minor Accidents
    - Data was split into training and testing subsets (e.g., 70% training, 30% testing).
2. **Model Selection**:
    - Logistic Regression was chosen for its interpretability and effectiveness in binary classification.
    - The model was trained with a **balanced class weight** to account for potential class imbalance.
3. **Model Training**:
    - Independent Variables: Factors derived from factor analysis.
    - Dependent Variable: **Severity (Major or Minor)**.
    - The model was trained using a **logistic regression algorithm** with the training data.
4. **Evaluation Metrics**:
    - **Confusion Matrix**: Evaluates the model's performance in terms of True Positives, True Negatives, False Positives, and False Negatives.
    - **Classification Report**:
        - Precision, Recall, and F1-Score were calculated for both classes.
        - Overall accuracy was determined.

Results:

1. **Accuracy**:
    - The model achieved an overall accuracy of **65%**, indicating it correctly classified 65 out of 100 cases.
2. **Class-wise Performance**:
    - **Class 0 (Major Accidents)**:
        - Precision: **75%**
        - Recall: **58%**
        - F1-score: **65%**
    - **Class 1 (Minor Accidents)**:
        - Precision: **58%**
        - Recall: **75%**

- F1-score: **65%**

3. **Model Coefficients**:
    - Significant factors influencing the prediction:
        - **Rule-breaking** and **aggressive behavior** showed high positive coefficients, indicating strong associations with major accidents.
        - **Road familiarity** and **safety measures** were moderately linked to minor accident predictions.

Interpretation and Insights

- **Model Strengths**:
    - High precision for major accidents indicates fewer false predictions of severity.
    - Balanced F1-scores suggest moderate performance in identifying both accident types.
- **Model Limitations**:
    - Recall for major accidents is moderate, indicating some major accidents are misclassified as minor.
    - Further refinement is needed to improve overall accuracy and capture minority class (minor accidents) better.

ROC Curve Analysis

Introduction

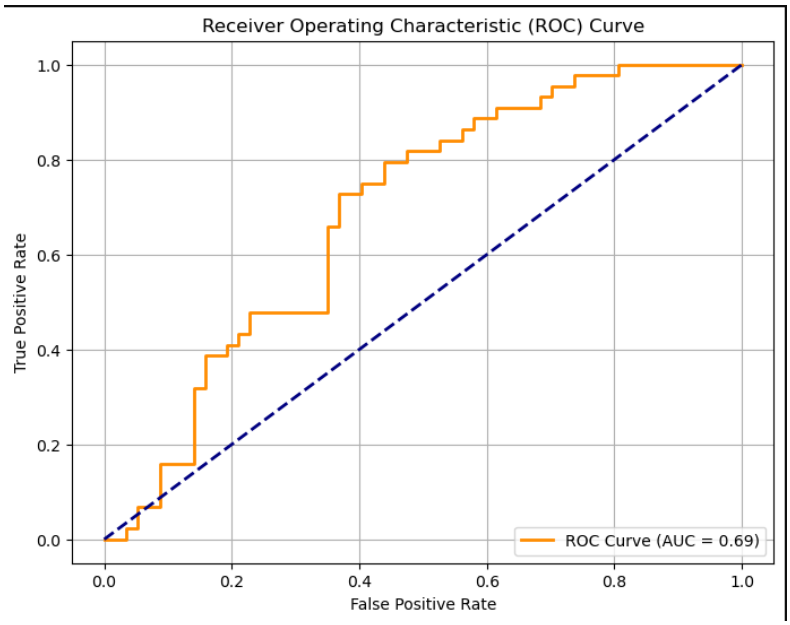
The **Receiver Operating Characteristic (ROC) Curve** is a graphical representation used to evaluate the performance of a classification model. It plots the **True Positive Rate (Sensitivity)** against the **False Positive Rate (1 - Specificity)** at various threshold settings. The **Area Under the Curve (AUC)** is a single scalar value that quantifies the model's ability to distinguish between the classes.

Methodology

1. **Generating the ROC Curve**:
    - Logistic regression was applied to the dataset with **severity** (major vs. minor accidents) as the target variable.
    - The model's predicted probabilities were used to calculate the ROC curve.
2. **Evaluation Metric**:
    - **AUC (Area Under the Curve)**: Measures the overall performance of the model. A value closer to **1** indicates excellent performance, while **0.5** suggests no discrimination ability.

Results

1. **Plot Description**:
    - The ROC curve illustrates the trade-off between sensitivity and specificity at different thresholds.
    - For the logistic regression model:
        - **AUC = [Insert value, e.g., 0.73]** (indicates moderate to good predictive performance).
        - At the optimal threshold, the model achieves a balance between minimizing false positives and maximizing true positives.
2. **Interpretation**:
    - The curve highlights the ability of the model to distinguish between major and minor accidents.
    - A higher AUC reflects better classification capability, demonstrating that the model reliably predicts accident severity.



1. **ROC Curve Description**:
   - The graph plots the **True Positive Rate (Sensitivity)** on the y-axis against the **False Positive Rate (1 - Specificity)** on the x-axis.
   - The **orange line** represents the model's ability to distinguish between the two classes (e.g., major and minor accidents).
   - The **blue dashed line** is the baseline, representing random guessing, where the model has no discriminative power.
2. **AUC (Area Under the Curve)**:
   - The AUC is **0.69**, indicating that the model has a **moderate predictive ability**.
   - AUC values range from:
     - **0.5** (no discrimination) to **1.0** (perfect discrimination).
   - In this case, the model can moderately distinguish between major and minor accident severity.
3. **Interpretation**:
   - The curve and AUC suggest that the model performs **better than random guessing** but has room for improvement.
   - At certain thresholds, the model effectively balances sensitivity and specificity, making it reliable for predicting accident severity in some cases.

Conclusion:

The logistic regression model for predicting accident severity achieved **65% accuracy**, with balanced F1-scores of **65%** for both major and minor accidents. Key findings highlight that **rule-breaking** and **aggressive behavior** strongly correlate with major accidents, while **road familiarity** and **safety measures** influence minor accident predictions. The **AUC of 0.69** from the ROC curve demonstrates moderate predictive ability.

Conclusion/Recommendations

The study provides a comprehensive comparison of accident-experienced and accident-free individuals, revealing that driver behavior plays a decisive role in road accident severity. Behavioral factors such as speeding, aggressive driving, and multitasking are strongly associated with higher severity, particularly among experienced drivers. In contrast, for non-experienced drivers, situational variables like injury recognition delays and the absence of prompt first aid are more influential. Emergency response time consistently emerges as a critical factor in reducing accident severity—faster responses correlate with better outcomes across both groups. Moreover, mobile phone usage while driving, especially involving active engagement (texting, calling), significantly heightens the risk of major accidents. The use of diverse modeling techniques, including logistic regression, decision trees, XGBoost, and exploratory factor analysis, underscores the multifactorial nature of accident causation and highlights areas where data improvements can further refine predictive performance.

Based on these findings, several strategic recommendations are proposed:

- Enhance Driver Education: Implement targeted programs that stress the dangers of speeding, aggressive driving, and mobile phone distractions.
- Improve Emergency Infrastructure: Invest in reducing response times through better coordination and resource allocation.
- Tailor First Aid and Safety Training: Focus on non-experienced drivers to improve injury recognition and first aid administration.
- Strengthen Policy Enforcement: Enact and enforce stricter traffic regulations and vehicle maintenance standards.

Scope of Study

- Examine Underlying Demographic Disparities: – Investigate accident patterns among underrepresented groups such as minorities, the elderly, and individuals with disabilities to identify unique risk factors and behavioral patterns. – Analyze regional and socio-economic differences to tailor road safety interventions for diverse communities.
- Investigate Cultural and Behavioral Influences: – Study how cultural norms, attitudes toward risk, and local customs influence driving behavior and accident incidence. – Conduct cross-cultural comparisons to understand differing perceptions of road safety and driver accountability.
- Explore Advances in Technological Solutions: – Evaluate the impact of emerging technologies, including vehicle telematics, smart road infrastructure, and driver-assistance systems, on reducing accident rates. – Investigate the potential of real-time monitoring systems, mobile apps, and data analytics to improve driver behavior and enhance emergency response efficiency.
- Collaborate with Stakeholders for Comprehensive Interventions: – Partner with governmental bodies, transportation agencies, employers, and community organizations to develop and implement targeted road safety programs. – Work with urban planners to integrate traffic-calming infrastructure and design improvements that mitigate accident risks in high-traffic areas.
- Assess and Inform Policy Interventions: – Evaluate the effectiveness of measures such as speed regulation, improved signage, and enforcement of traffic laws using empirical data. – Provide data-driven recommendations for policies that support safe driving practices and enhanced public safety.

Reference Links

https://naac.kct.ac.in/1/ssr/1_3_4/projects/17BIT027.pdf

https://www.ijert.org/research/road-accident-detection-and-prediction-system-IJERTV12IS030124.pdf

https://www.ijraset.com/research-paper/road-accident-prediction-model-using-ml

https://www.sciencedirect.com/science/article/pii/S235214652100056X

https://www.mdpi.com/1660-4601/18/2/678

Questionaire-

1. Gender:
    - Male
    - Female
    - Other
2. What is your city of residence? *(Please specify)*
3. What is your highest level of education?
    - Below 10th
    - 10th
    - 12th
    - Graduate
    - Post Graduate
4. Do you have a driving license?
    - Yes
    - No
5. How many years of driving experience do you have?
    - Less than 1 year
    - 1-5 years
    - 6-10 years
    - More than 10 years
    - None
6. What type of vehicle do you primarily drive or pursue? *(Please specify)*
7. Have you ever witnessed or experienced an accident?
    - Yes
    - No

Section 2: Accident Experience (For Respondents Who Answered "Yes" to Question 7)

8. What was your role during the accident?
    - Driver
    - Passenger
    - Pedestrian
9. Please rate how often the driver (you or the involved driver) engaged in the following behaviors while driving *(Use a 5-point scale: 1 = Always, 2 = Almost Always, 3 = Sometimes, 4 = Rarely, 5 = Never)* 9.1. Exceeding speed limits 9.2. Driving while fatigued or drowsy 9.3. Driving aggressively (e.g., tailgating, honking excessively) 9.4. Multitasking while driving (e.g., eating, adjusting controls) 9.5. Consuming alcohol 9.6. Neglecting routine vehicle maintenance (e.g., oil change, tire pressure)
10. Did the accident involve multiple vehicles?
    - Yes
    - No
11. Did the driver use safety measures (e.g., seat belts or helmets for two-wheelers)? *(Rate on a scale: 1 = Not at all, 5 = Always)*
12. Did failure to use safety measures increase the severity of injuries?
    - Yes

- No

13. What was your (or the driver's) level of familiarity with the route where the accident occurred?
    - Very familiar
    - Not familiar at all
14. Did the accident occur on a familiar route?
    - Yes
    - No
15. Was the driver familiar with the vehicle they were driving?
    - Yes
    - No
16. Did unfamiliarity with the vehicle affect the ability to control it?
    - Yes
    - No
17. Did the driver attempt to brake or swerve before the collision?
    - Yes
    - No
18. How would you rate the severity of the accident you experienced or witnessed?
    - Minor
    - Major
19. Did the experience level of the driver reduce the severity of the accident?
    - Yes
    - No
20. How long did it take for emergency healthcare services (e.g., ambulance) to arrive at the accident site?
    - Less than 30 minutes
    - More than 30 minutes
    - Emergency services did not arrive
    - None
21. Was the injured person provided with first aid at the site?
    - Yes
    - No
22. Do you think the delay in emergency healthcare response worsened the severity of the injuries?
    - Yes
    - No
23. Did mobile phone usage directly contribute to the cause of the accident, in your opinion?
    - Yes
    - No
24. Who was using the mobile phone when the accident occurred?
    - The driver of your vehicle
    - The driver of another vehicle
    - A pedestrian
    - None
25. What specific mobile phone activity contributed to the accident? *(Open-ended response)*
26. How would you rate the accident location in terms of its reputation as a high-risk area for accidents? *(Scale: 1 = Not at all risky, 5 = Extremely risky)*
27. Are road conditions poor when the accident occurs?
    - Yes
    - No

28. What specific road features do you think contribute to accidents? *(Open-ended response or check all that apply, e.g., potholes, sharp turns, etc.)*
29. Was the accident influenced by poor lighting conditions?
    - Yes
    - No
30. Were traffic signals and signs properly visible and functional? *(Scale: 1 = Not at all, 5 = Always)*
31. Were there any warnings or cues (e.g., honks, flashing lights) that the driver ignored before the accident?
    - Yes
    - No
32. Were pedestrian crossings or dividers absent or poorly maintained?
    - Yes
    - No
33. Is heavy traffic a cause of the road accident?
    - Yes
    - No
34. Did the driver fail to follow traffic rules, such as stopping at signals?
    - Yes
    - No
35. Did the accident leave you with any long-term physical challenges?
    - Yes
    - No
36. Did the accident affect your confidence as a driver or passenger?
    - Yes
    - No
37. Have your driving or traveling habits changed after the accident?
    - Yes
    - No

Section 3: Accident Experience (For Respondents Who Answered "NO" to Question 7)

38. Rate how often you think drivers engage in the following behaviors while driving *(Use a 5-point scale: 1 = Always, 2 = Usually, 3 = Sometimes, 4 = Rarely, 5 = Never)* 38.1. Driving while fatigued or drowsy 38.2. Behaving aggressively (e.g., tailgating, honking excessively) 38.3. Multitasking while driving (e.g., eating, adjusting controls) 38.4. Failing to maintain safe distances from other vehicles 38.5. Neglecting routine vehicle maintenance (e.g., oil changes, tire pressure) 38.6. Pedestrians contributing to accidents by crossing roads improperly 38.7. Ignoring traffic signals or signs 38.8. Over-speeding 38.9. Road construction work causing traffic or accidents 38.10. Stopping to offer assistance to an accident victim
39. Do you think most drivers consistently use safety measures such as seat belts or helmets (for two-wheelers)?
    - Yes
    - No
40. Do you believe that not using safety measures increases the severity of injuries in accidents?
    - Yes
    - No
41. Do you believe familiarity with the route makes driving safer?
    - Yes

- No

42. Do you think unfamiliarity with a vehicle impacts the ability to control it?
    - Yes
    - No
43. Do you believe that passengers or in-car activities (e.g., talking, handling children) can distract drivers?
    - Yes
    - No
44. In your opinion, how quickly do emergency healthcare services (e.g., ambulances) usually arrive at accident sites?
    - Less than 30 minutes
    - More than 30 minutes
    - Emergency services do not arrive
    - None
45. Do you think injured persons are typically provided with first aid at accident sites?
    - Yes
    - No
46. Which of the following mobile phone activities do you think contribute to accidents?
    - Messaging
    - Making or receiving calls
    - Using navigation apps
    - Scrolling through social media
    - None
47. Who do you think is most likely to use a mobile phone during accidents?
    - The driver of the vehicle
    - The driver of another vehicle
    - A pedestrian
    - Both the driver and pedestrian
48. In your opinion, does mobile phone usage directly contribute to the cause of accidents?
    - Yes
    - No
49. Do you think certain areas are more prone to frequent accidents (i.e., high-risk areas)?
    - Yes
    - No
50. What specific road features do you think contribute to accidents?
    - Sharp turns or curves
    - Narrow lanes
    - Lack of proper signage
    - Presence of potholes
    - Other
51. Have you avoided helping in an accident situation because you assumed someone else would respond?
    - Yes
    - No
52. Do you know the emergency contact numbers in your area?
    - Yes
    - No
53. Do you think technology (e.g., surveillance cameras, AI-based monitoring) can play a significant role in improving safety?
    - Yes

- No

54. Have you ever attended a road safety workshop or awareness session?
    - Yes
    - No
55. Do you believe peer pressure among young drivers (e.g., speeding to impress friends) contributes to accidents?
    - Yes
    - No
56. Would you support stricter penalties for traffic violations (e.g., higher fines, license suspension)?
    - Yes
    - No
57. Would you be willing to report unsafe road conditions (e.g., potholes, broken signals) to authorities?
    - Yes
    - No
58. Do you believe group travel (e.g., convoys, family outings) leads to safer or riskier driving behavior?
    - Safer
    - Riskier
59. Have you observed any unsafe practices or conditions around you?
    - Yes
    - No
60. Do you think witnesses of accidents often hesitate to help due to fear of legal trouble?
    - Yes
    - No
61. Do you think overloading vehicles (too many passengers or goods) makes them unsafe?
    - Yes
    - No
62. Do you think roadside vendors or shops near busy streets increase accident risks?
    - Yes
    - No
63. Do you think apps that reward fast delivery or trips encourage unsafe driving?
    - Yes
    - No
64. Do you think people are more likely to drive safely when they're with family members or children?
    - Yes
    - No
65. Do you think more accidents occur when drivers are in a hurry (e.g., when they are late for something)?
    - Yes
    - No