

**Project
Report On**

**A COMPREHENSIVE STATISTICAL
ANALYSIS AND PREDICTION OF ROAD
ACCIDENTS**

Submitted

in partial fulfilment of the requirements for the award of the degree

of

BACHELOR OF SCIENCE

in

MATHEMATICS

By

ALKA BAIJU

(Register No. AB21AMAT040)

Under the Supervision of

SUSAN MATHEW PANAKKAL



DEPARTMENT OF MATHEMATICS AND STATISTICS

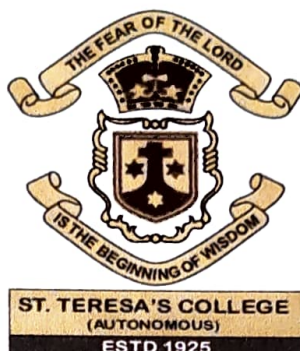
ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

MARCH 2024



ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM





CERTIFICATE

This is to certify that the dissertation entitled **A COMPREHENSIVE STATISTICAL ANALYSIS AND PREDICTION OF ROAD ACCIDENTS** is a bonafide record of the work done by Ms. **ALKA BAIJU** under my guidance in partial fulfillment of the award of the degree of **Bachelor of Science in Mathematics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 25-03-2024

Place: Ernakulam


Susan Mathew Panakkal
Assistant Professor,
Department of Mathematics,
St. Teresa's College
(Autonomous), Ernakulam.


Dr. Ursala Paul
Assistant Professor and Head,
Department of Mathematics
St. Teresa's College
(Autonomous), Ernakulam.

External Examiners

1: 

2:



DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of Smt. SUSAN MATHEW PANAKKAL, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam

Date: 25-03-2024

Alka Baiju
ALKA BAIJU

AB21AMAT040



ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude towards Smt. SUSAN MATHEW PANAKKAL of the Department of Mathematics and Statistics of St. Teresa's College who encouraged me to carry out this work. Her continuous invaluable and knowledgeable guidance throughout this study helped me to complete the work up to this stage. I will always be thankful to you in this regard. I also express my profound thanks to all those who have directly or indirectly guided and helped me in the completion of this project.

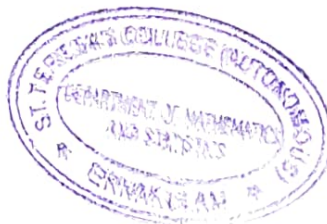
Ernakulam

Date: 25-03-2024



ALKA BAIJU

AB21AMAT040



Contents

<i>CERTIFICATE</i>	ii
<i>DECLARATION</i>	iii
<i>ACKNOWLEDGMENT</i>	iv
<i>CONTENT</i>	v
1. INTRODUCTION	1
1.1 OBJECTIVE	1
1.2 DATA SOURCE.....	2
1.3 DATA DESCRIPTION.....	2
1.3.1 PRIMARY DATA.....	2
1.3.2 SECONDARY DATA.....	2
1.4 LITERATURE REVIEW.....	3
2. METHODOLOGY	6
2.1 CHI-SQUARE TEST.....	6
3. DATA ANALYSIS	7
3.1 CHI-SQUARE TEST.....	7
3.2 PICTORIAL REPRESENTATION OF CAUSES OF ROAD ACCIDENT.	12
4. MACHINE LEARNING	15
4.1 METHODOLOGY.....	15
4.1.1 DATA PROCESSING.....	15
4.1.2 DATA ANALYSIS.....	16
4.1.3 DATA VISUALIZATION.....	16
4.1.4 CORRELATION MATRIX.....	17
4.1.5 SEPERATING FEATURES AND TARGET.....	17
4.1.6 DATA SPLITTING.....	18
4.1.7 MODEL TRAINING.....	18
4.1.8 MODEL EVALUATION.....	18
4.1.8.1 ACCURACY SCORE.....	19
4.2 ALGORITHMS.....	19
4.2.1 DECISION TREE.....	19
4.2.2 SUPPORT VECTOR MACHINE (SVM).....	20
4.2.3 K- NEAREST NEIGHBORS (KNN).....	20
4.2.4 RANDOM FOREST.....	20

5. DATA ANALYSIS	21
5.1 DATA PROCESSING	21
5.2 DATA ANALYSIS.....	22
5.3 DATA VISUALIZATION.....	23
5.4 ENCODING THE CATOGORICAL COLUMNS.....	23
5.5 CORRELATION MATRIX.....	24
5.6 SEPERATING FEATURES AND TARGET.....	25
5.7 ACCURACY COMPARISON.....	25
6. RESULTS AND CONCLUSION	26
6.1 FINDINGS	26
6.2 CONCLUSION.....	26
BIBLIOGRAPHY	
ANNEXURE.....	

CHAPTER 1

INTRODUCTION

Road accidents are significant public health concern, impacting communities and individual alike. Ernakulam a bustling district in Kerala, is no exception to this issue. This study aims to delve into various aspects of road accidents in Ernakulam through a comprehensive statistical analysis. Road accidents are pressing concern worldwide, contributing significantly to injury, mortality, and socioeconomic burdens. This study delves into the specific context of Ernakulam, aiming to comprehensively understand various facets of road accidents. Emphasizing the importance of this research, the study seeks to shed light on gender-based variations, the influence of educational backgrounds, regional disparities between city and village areas, causes of accidents, the nature of incidents, and predictive analysis for the upcoming years.

The scope of this study extends beyond merely quantifying accident rates; it aims to uncover patterns, correlations, and predictive indicators. By focusing on Ernakulam, a region with its unique demographic and geographical characteristics, the findings can contribute to targeted interventions and policy measures. The study explores both individual and environmental factors contributing to accidents, providing a holistic understanding that goes beyond statistical figures.

1.1 OBJECTIVE

This study aspires to provide a comprehensive understanding and prediction of road accidents in Ernakulam, offering evidence-based insights that can inform policies, educational campaigns, and interventions to enhance road safety and reduce the incidence of accidents in the region.

The research is guided by several specific objectives, each shedding light on different dimensions of road safety.

1. To determine if road accidents in Ernakulam are gender based

2. To test the relationship between educational background and road accident in Ernakulam.
3. To test the independence of injury type and areas (urban and rural).
4. To determine the causes of road accidents in Ernakulam.
5. To study the accuracy values for four machine learning algorithms.

1.2 DATA SOURCE

The data was collected by conducting an online survey among people of age group 18 and above from both gender in Ernakulam district and also by collecting secondary data from District Crime Records Bureau (DCRB)

1.3 DATA DESCRIPTION

The data is a primary and secondary which included the various factors that is necessary for the objectives. The questionnaire was circulated using Google form. Also awareness of people about road safety measures is also examined

1.3.1 PRIMARY DATA

Data was collected by circulating a Google form. A Google form containing 15 questions was designed to obtain information from the sample population. The sample of this study include both gender. The variables under consideration were: Gender, Educational Background, Causes of road accident.

1.3.2 SECONDARY DATA

Secondary data was collected from District Crime Records Bureau (DCRB). The different variables under consideration were: accident type, death, driver, passenger, pedestrian, cyclist, collision, area type, speed limit, weather, causes of road accidents, type road and road features.

1.4 LITERATURE REVIEW

[1] In 2016 Sanjay Kumar Singh had conducted a study on road traffic accidents in India during the period 1970 to 2013. The analysis reveals that road accidents in India exhibit variations in fatalities and injuries based on age, gender, month, and time. The economically active age group is identified as the most vulnerable. Generally, males experience higher fatality and accident risks compared to females. Notably, extreme weather in May-June and December-January influences road accident rates. Accidents peak from 9 AM to 9 PM, remaining relatively constant, while being variable but low during midnight and early morning. Drivers' fault is the primary factor, contributing to 78% of total accidents 76.5% of injuries, and 73.7% of fatalities in 2013. The study also assesses road accidents across Indian states and cities, finding higher fatality risks in Tamil Nadu (22.8), Haryana (17.2), and Andhra Pradesh (16.9) in 2013, compared to the national average (11.2). Metropolitan cities show a slightly lower burden of road traffic accidents, with significant variation in fatality risks among cities, ranging from 3.0 in Kolkata to 25.5 in Jaipur per 100,000 people.

[2] In 1984 Dinesh Mohan and P.S Bawa had conducted an analysis of road traffic fatalities in Delhi, India in the year 1980 (from 1 January to 31 December). The majority of traffic crash casualties in Delhi involve pedestrians, two-wheeler riders, and bus commuters, with motor vehicle occupants constituting a small minority. Fatal crashes primarily affect the working-age population, and fatality rates for children and the elderly are not as high as in high-income countries. Buses and trucks are frequently involved in various types of crashes, and two-wheeler fatal crashes are more common at night despite lower traffic densities. Most crashes occur on straight roads rather than intersections, with T-junctions being more frequently involved than roundabouts or four-arm junctions. A notable number of pedestrians are killed on pedestrian crossings along straight roads. Delhi's vehicle fatality rates do not decrease proportionally with the increase in vehicle ownership, unlike the trend seen in industrialized countries. Fatality rates in Delhi show little variation throughout the year.

[3] D r. Sanjay K. Singh had conducted a case study on road accidents in Patna City in the year 2000. Patna, like many other cities, grapples with the challenge of increasing fatalities and injuries on its roads. The total number of fatal accidents

and related fatalities in the city has been on the rise over the years. The fatality rate per 100 accidents is notably high, reaching 45 in the year 2000. While the overall fatality rate in Patna is relatively low, the fatality risk exceeds the national average. Particularly concerning is the disproportionately high percentage of pedestrian deaths, accounting for over 90% of all road fatalities. Both the working-age adult population and pedestrians constitute a significant portion of road accident fatalities. In terms of vehicle-wise accident rates, buses pose the highest risk, causing an average of approximately 16 accidents per thousand buses annually.

Additionally, the city's traffic police have identified specific accident-prone locations based on the severity and frequency of accidents. Since the year 2000, the new bypass road on National Highway (NH-38) is considered the most accident-prone location in the city.

[4] Rakesh Kakkar, Pradeep Aggarwal, Monica Kakkar, Kirti Deshpande and Divya K Gupta had conducted a retrospective study on traffic road accidents. During the study period, 87 accidents were observed, resulting in 9 deaths, leading to a mortality rate of 10.3%. The age group most susceptible to accidents was 25-34 years, predominantly males. Saturdays had the highest accident occurrence, mainly between 10:00 AM-11:00 AM, but deaths were more prevalent between 9:00 PM and 12:00 midnight. State highways had more road traffic accidents compared to other urban roads. Pedestrians faced a higher risk than two or four-wheelers, with males being the most affected. The primary cause of road traffic accidents was exceeding the speed limit (47%).

Most injuries occurred near schools, cinemas, factories, and bus stands—overcrowded places, but mostly resulting in simple injuries. In contrast, injuries near village highways were more serious. Roads without traffic signals saw a higher incidence of injuries (73.6%), mostly of a serious nature, compared to chaurahas (14.9%), tiraas (08%), or 'Y' type roads (3.4%) with proper signals. The leading causes of road traffic accidents were driving above the speed limit (47.1%), alcohol consumption by drivers (32.1%), and rash driving at turns (20.7%).

[5] In the research, Road Accident Analysis and Prediction using Machine Learning Algorithmic Approaches by Koteswara Rao Ballamudi concluded that the intolerable toll of road accidents affects both the general public and non-industrial nations alike. Consequently, implementing a robust traffic management

system has become imperative to reduce accident rates. Proactive measures, such as a sophisticated warning system, can help prevent accidents. Many nations are now compelled to address this pressing issue, as the number of fatalities continues to rise steadily. Leveraging machine learning offers a practical solution for making informed decisions and providing insights to traffic authorities to mitigate accidents. By employing proven machine learning approaches, we can enhance accuracy in predicting accident severity. While previous studies primarily focused on distinguishing between injury and non-injury categories, our research expanded to include possible injury levels such as non-incapacitating, incapacitating, and fatal injuries. Our findings indicate that models predicting fatal and non-fatal injuries outperform others. Predicting these injury types is crucial, as fatalities carry significant economic and social costs to society. One key contributing factor to varying injury levels is the actual speed of the vehicle at the time of the accident. Unfortunately, our dataset lacks sufficient information on actual speeds, with 67.68% of records missing this data. Availability of this information could have potentially improved the performance of the models discussed in this study.

CHAPTER 2

METHODOLOGY

2.1 CHI-SQUARE TEST

The chi-square test is a statistical method used to determine if there is a significant association between two categorical variables. It involves comparing observed frequencies in a contingency table with the frequencies we would expect by chance. The test calculates a chi-square statistic, which measures the difference between observed and expected values.

Calculate Chi-square Statistic:

Use the formula: $\chi^2 = \sum [(Observed - Expected)^2 / Expected]$

To perform a chi-square test

1. Set up Hypotheses:

Null Hypothesis (H_0): Assumes no association or independence between the variables.

Alternative Hypothesis (H_1): Suggests a significant association or dependence.

Define alpha value. This involves deciding the risk you are willing to take of drawing the wrong conclusion.

2. Check the data for errors.

3. Check the assumptions for the test.

4. Perform the test and draw the conclusion.

CHAPTER 3

DATA ANALYSIS

3.1 CHI-SQUARE TEST

TEST 1

If there exist any association between gender and road accident in Ernakulam

H_0 : There exist no association between gender and road accident in Ernakulam

H_1 : There exist association between gender and road accident in Ernakulam

Cross tabulation

			Have you been involved in road accidents during the past year 2019 -2023		Total
			No	Yes	
Gender	Female	Count	89	11	100
		Expected Count	76.0	24.0	100.0
	Male	Count	63	37	100
		Expected Count	76.0	24.0	100.0
Total	Count		152	48	200
	Expected Count		152.0	48.0	200.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	18.531 ^a	1	.000		
Continuity Correction ^b	17.133	1	.000		
Likelihood Ratio	19.338	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	200				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.00.

b. Computed only for a 2x2 table

P VALUE = .000

Since p value is less than 0.05, we reject the null hypothesis. Hence we can conclude that there exist association between gender and road accident in Ernakulam.

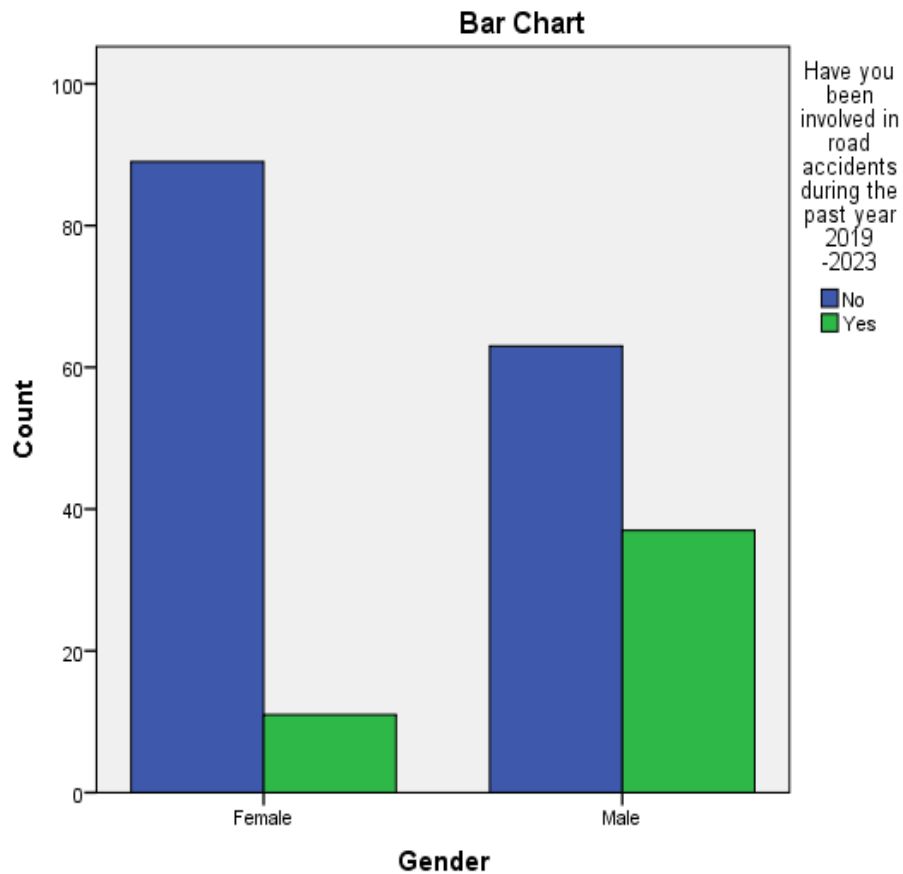


Fig.3.1

Based on the depicted graph, it is evident that a correlation exists between road accidents and gender. The incidence of road accidents is notably higher among males compared to females

TEST 2

If there exist any relationship between educational background and road accident in Ernakulam

H_0 : The increase in education correlates with a decrease in road accidents.

H_1 : The increase in education does not correlates with a decrease in road accidents.

Cross tabulation

			Have you been involved in road accidents during the past year 2019 -2023		Total
			No	Yes	
Education Qualification	Graduate	Count	82	36	118
		Expected Count	89.7	28.3	118.0
	Higher Secondary	Count	70	12	82
		Expected Count	62.3	19.7	82.0
Total	Count		152	48	200
	Expected Count		152.0	48.0	200.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.684 ^a	1	.010		
Continuity Correction ^b	5.842	1	.016		
Likelihood Ratio	6.991	1	.008		
Fisher's Exact Test				.011	.007
N of Valid Cases	200				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 19.68.

b. Computed only for a 2x2 table

P VALUE = 0.010

Since p value is less than 0.05, we reject the null hypothesis. Hence we can conclude that obtaining education does not lead to a significant reduction in road accidents.

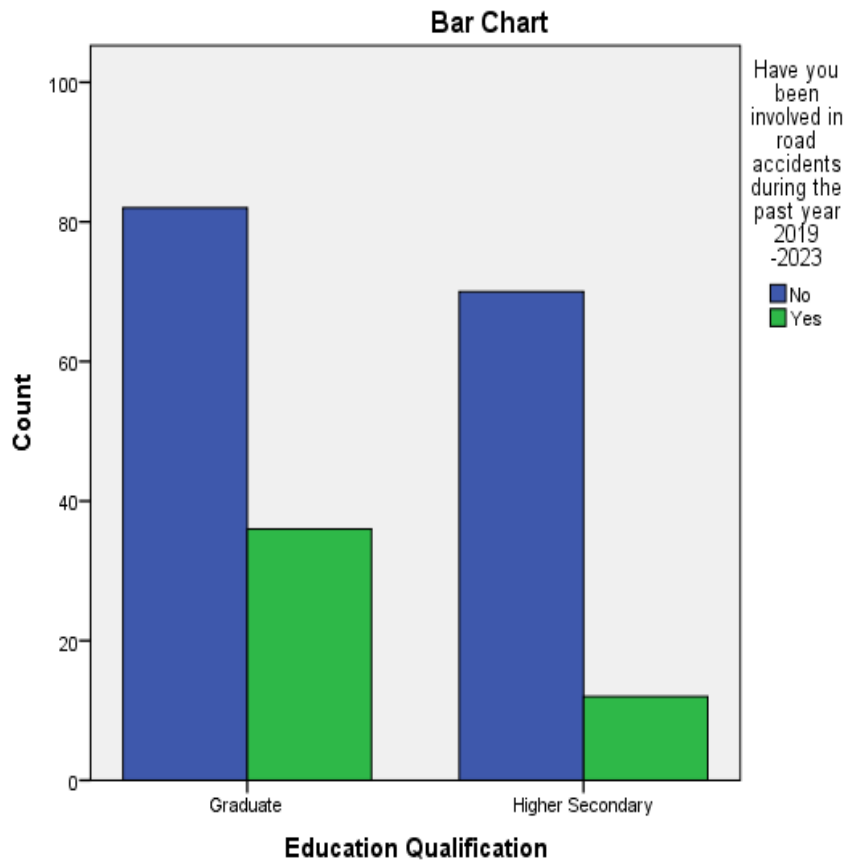


Fig.3.2

Based on the depicted graph, it is evident that obtaining education does not lead to a significant reduction in road accidents

TEST 3

If there exist any association between injury type and road area

H_0 : There exist no association between injury type and road area

H_1 : There exist association between injury type and road area

2019-2023 Area Type * Accident type Cross tabulation

			Accident type		Total
			Greivous Injury	Minor Injury	
Area Type	Rural	Count	1472	357	1829
		Expected Count	1427.4	401.6	1829.0
	Urban	Count	5966	1736	7702
		Expected Count	6010.6	1691.4	7702.0
Total		Count	7438	2093	9531
		Expected Count	7438.0	2093.0	9531.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.870 ^a	1	.005		
Continuity Correction ^b	7.694	1	.006		
Likelihood Ratio	8.041	1	.005		
Fisher's Exact Test				.005	.003
N of Valid Cases	9531				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 401.65.

b. Computed only for a 2x2 table

P VALUE = 0.005

Since p value is less than 0.05, we reject the null hypothesis.

Hence we can conclude that there exist association between injury type and road area

3.2 PICTORIAL REPRESENTTION OF CAUSES OF ROAD ACCIDENTS

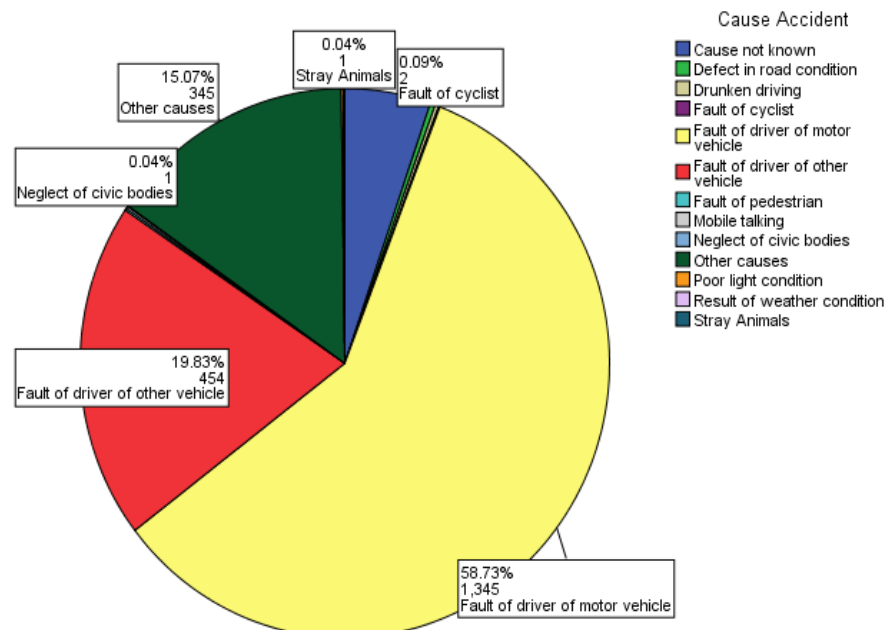


Figure 1 : 2019

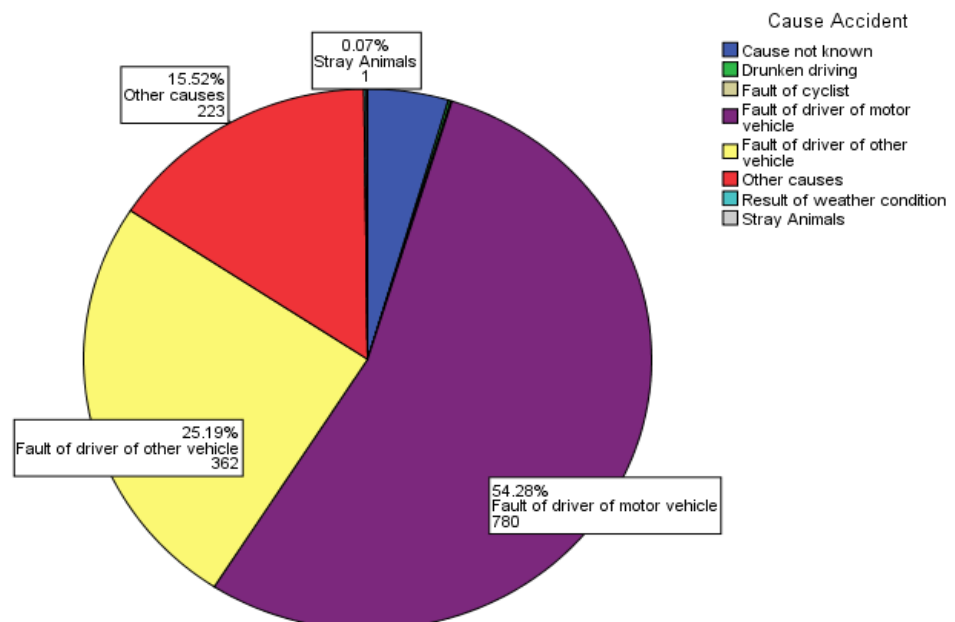


Figure 2: 2020

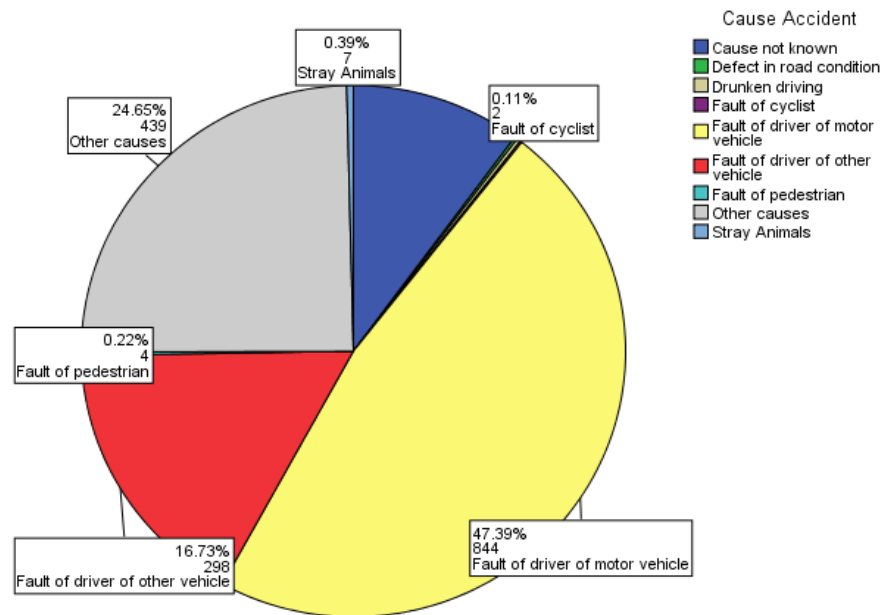


Figure 3: 2021

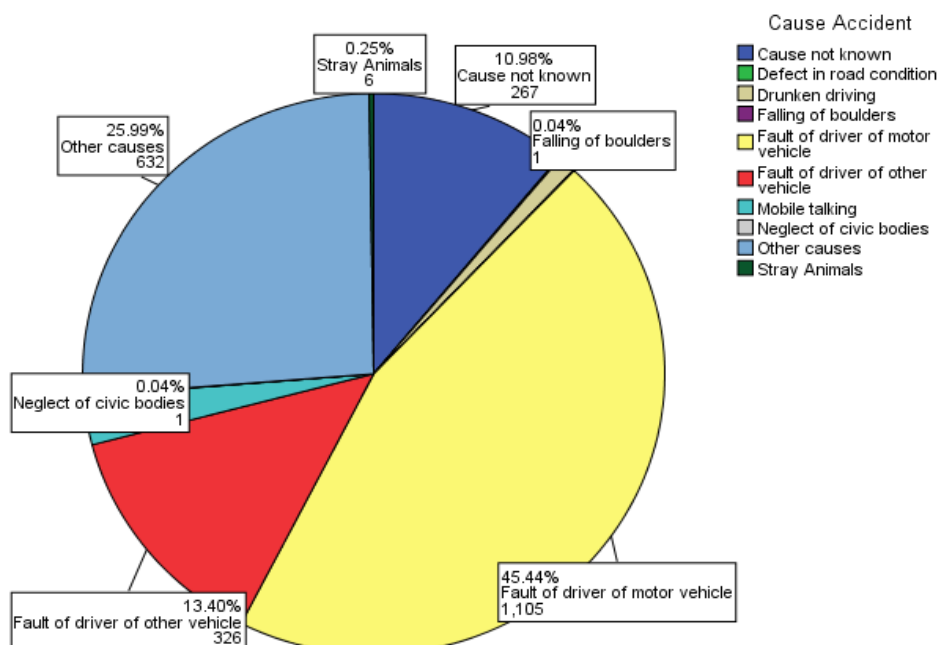


Figure 4:2022

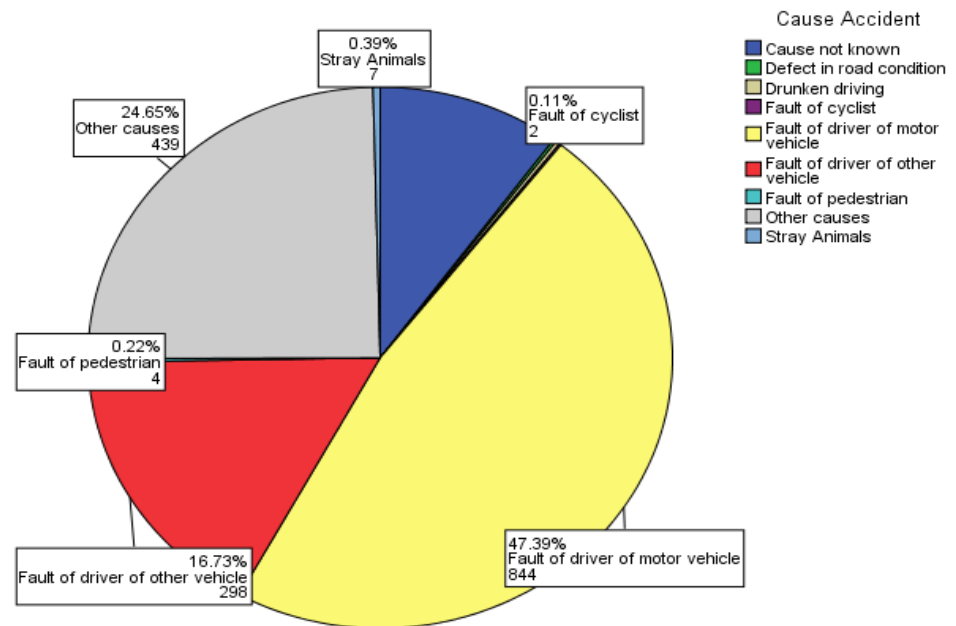


Figure 5: 2023

CHAPTER 4

MACHINE LEARNING

4.1 METHODOLOGY

Machine learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit programming instructions. Instead, machine learning algorithms learn from data, identify patterns, and make predictions or decisions based on that data. It's used in various applications.

In our predictive analysis, we employ a systematic approach to gauge the accuracy of our model using both training and test datasets. The process encompasses several key steps, beginning with thorough data preprocessing to enhance the quality and consistency of our input. Subsequently, we conduct comprehensive data analysis to extract meaningful insights, followed by data visualization to illustrate patterns and trends. A crucial element in our assessment involves the creation of a correlation matrix, elucidating the relationships between variables. Moving forward, we proceed to data fitting, wherein we utilize a decision tree for model training, leveraging its ability to discern intricate patterns within the dataset. In the final stages of model evaluation, we focus on accuracy scoring to quantify the effectiveness of our predictive model. This multifaceted approach ensures a robust and informed analysis, culminating in a comprehensive understanding of the model's performance and predictive capabilities like image recognition, natural language processing, recommendation systems, and more.

4.1.1 DATA PROCESSING

Data preprocessing in machine learning is the process of preparing and cleaning raw data to make it suitable for training models. This involves several steps such as handling missing values, removing outliers, scaling features, and encoding categorical variables. The goal is to ensure that the input data is consistent, accurate, and in a format that facilitates effective learning by machine learning algorithms. Good data preprocessing contributes to better model performance and generalization.

4.1.2 DATA ANALYSIS

Data analysis in machine learning involves examining and interpreting data to gain insights and inform decision-making in the context of building and training machine learning models. This process includes tasks like exploratory data analysis (EDA), where patterns, trends, and relationships within the data are identified. Additionally, feature engineering, a crucial aspect of data analysis, focuses on selecting, transforming, and creating features that enhance a model's predictive performance. The ultimate goal is to extract meaningful information from the data, aiding in the development of accurate and robust machine learning models.

4.1.3 DATA VISUALIZATION

Data visualization in machine learning involves creating graphical representations of data to facilitate understanding, exploration, and communication of patterns and insights. It uses charts, graphs, and other visual elements to convey complex information in a more accessible and interpretable form. In the context of machine learning, data visualization is crucial during the exploratory data analysis (EDA) phase to reveal trends, correlations, and outliers. It helps researchers and practitioners make informed decisions about feature selection, model evaluation, and overall data quality. Effective data visualization enhances the interpretability of machine learning results, making it easier to communicate findings to stakeholders.

Encoding categorical columns in machine learning involves converting categorical data, which represents labels or categories, into numerical format. This is necessary because many machine learning algorithms require numerical input. There are two common methods for encoding categorical columns:

a) LABEL ENCODING

Assigns a unique numerical label to each category.

Useful for ordinal data where the order matters.

b) ONE-HOT ENCODING

Creates binary columns for each category and indicates the presence of the category with a 1 and absence with a 0.

Suitable for nominal data where there is no inherent order among categories.

4.1.4 CORRELATION MATRIX

Correlation matrix in machine learning is a table that shows the correlation coefficients between many variables. Each cell in the matrix represents the correlation between two variables. The values range from -1 to 1, where:

1 indicates a perfect positive correlation: As one variable increases, the other variable also increases proportionally.

-1 indicates a perfect negative correlation: As one variable increases, the other variable decreases proportionally.

0 indicates no correlation: Changes in one variable do not predict changes in the other

A correlation matrix is for understanding the relationships between different features or variables in a dataset. It helps in feature selection, identifying multicollinearity (high correlation between two or more independent variables), and gaining insights into how variables influence each other. In machine learning, this information is valuable for making informed decisions about feature engineering and model building.

4.1.5 SEPARATING FEATURES AND TARGET

In machine learning, it's crucial to distinguish between features and the target variable when preparing data for training a model.

1. Features:

- Features are the input variables or attributes that the model uses to make predictions.
- These are the characteristics or properties of the data you provide to the model.

2. Target:

- The target variable is the output variable that the model is trying to predict.
- It is the variable you want the model to learn to predict based on the provided features.

4.1.6 DATA SPLITTING

Data splitting in machine learning involves dividing a dataset into two or more subsets for different purposes, typically for training and testing models. The most common splits are:

1. Training Set:

- The largest portion of the dataset used to train the machine learning model.
- The model learns patterns, relationships, and features from this set.

2. Test Set:

- A portion of the data kept separate until the model is fully trained.
- Used to evaluate the model's performance on unseen data and estimate its generalization capabilities.

4.1.7 MODEL TRAINING

Model training in machine learning is a crucial step where algorithms like decision trees, SVM (Support Vector Machines), k-NN (k-Nearest Neighbors), and random forests learn patterns and relationships from labeled data. Decision trees create a tree-like structure by recursively splitting data based on features, optimizing for decision-making. SVM identifies a hyperplane that best separates classes in a high-dimensional space, maximizing the margin between them. k-NN classifies instances based on the majority class among their k-nearest neighbors. Random forests, on the other hand, build multiple decision trees and combine their outputs to enhance predictive accuracy. During training, these algorithms adjust their internal parameters to minimize errors and optimize performance, allowing them to make accurate predictions on new, unseen data during the testing or deployment phase.

4.1.8 MODEL EVALUATION

Model evaluation in machine learning involves assessing the performance and effectiveness of a trained model on new, unseen data. It aims to gauge how well the model generalizes to real-world scenarios. Common evaluation metrics include accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve, depending on the nature of the prediction task

(classification, regression, etc.). Cross-validation techniques, such as k-fold cross-validation, are often used to robustly estimate a model's performance by splitting the dataset into multiple subsets for training and testing. The goal is to ensure that the model doesn't merely memorize the training data but learns underlying patterns that can be applied to new, unseen data. Model evaluation is crucial for selecting the best-performing model, tuning hyperparameters, and making informed decisions about deploying the model in real-world applications

4.1.8.1 ACCURACY SCORE

The accuracy score in machine learning, based on a decision tree classifier, assesses the model's performance in predicting outcomes correctly. It measures the proportion of correctly classified instances out of the total instances. When applied to both the training and test datasets, it provides insights into how well the model generalizes to unseen data. Evaluating the accuracy score on the training data helps to understand how well the model fits the data it was trained on, but it doesn't necessarily indicate how well it will perform on new, unseen data. Conversely, assessing the accuracy score on the test data provides a more reliable indication of the model's performance on unseen instances. A high accuracy score on the test data suggests that the model has learned the underlying patterns well and can make accurate predictions on new data, while a significantly lower score may indicate issues such as overfitting or underfitting. Overall, comparing the accuracy scores on both training and test data helps in evaluating the model's performance and generalization capabilities.

4.2 ALGORITHMS

4.2.1 DECISION TREE

A decision tree in machine learning is a predictive model that maps features to outcomes. It's 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 the final predicted outcome. Decision trees are commonly used for classification and regression tasks in predictive modeling. They're interpretable and can be visualized, making them valuable for understanding the logic behind predictions.

4.2.2 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification and regression tasks. Its primary objective is to find the optimal hyperplane that separates different classes in a feature space, maximizing the margin between them. SVM works by transforming input data into a higher-dimensional space, making it easier to find a hyperplane that effectively separates classes. The algorithm identifies support vectors, which are data points crucial for defining the decision boundary. SVM is particularly effective in high-dimensional spaces and is widely used for tasks such as image classification, text classification, and bioinformatics. Its versatility and ability to handle non-linear relationships through kernel functions contribute to its popularity in predictive modeling.

4.2.3 K- NEAREST NEIGHBORS (KNN)

K-nearest neighbors (KNN) is a machine learning algorithm used for prediction and classification tasks. It operates on the principle of proximity, where the prediction for a new data point is determined by the majority class or average value of its k-nearest neighbors in the feature space. The "k" in KNN represents the number of neighbors considered. In essence, KNN assumes that similar instances in the feature space tend to share similar outcomes. This algorithm doesn't make assumptions about the underlying data distribution and defers computation until predictions are needed. While KNN is simple and intuitive, its performance can be sensitive to the choice of distance metric and the value of k.

4.2.4 RANDOM FOREST

Random Forest is a powerful ensemble learning algorithm in machine learning designed for predictive tasks. It operates by constructing a multitude of decision trees during training and outputs the average prediction of the individual trees for regression problems or a majority vote for classification tasks. Each tree in the forest is built using a random subset of the training data and features, introducing diversity to the models. This diversity helps improve the overall performance and robustness of the model, reducing overfitting. Random Forest is known for its versatility, scalability, and ability to handle complex datasets, making it a popular choice for predictive modeling in various domains.

CHAPTER 5

DATA ANALYSIS

5.1 Data processing

	Accident type	Death	Grievous	Minor	Driver	Passenger	Pedestrian	Cyclist	Collision	Area Type	Speed Limit	Weather	Hit Run	Cause Accident	Type Road	Road Features
0	Grievous Injury	0	1	0	1	0	0	1	2	Urban	<40	Sunny	no	Fault of driver of motor vehicle	State Highway	Straight Road
1	Minor Injury	0	0	1	1	0	1	0	0	Urban	40-80	Sunny	no	Other causes	Other Road	Others
2	Grievous Injury	0	1	0	1	0	1	0	0	Urban	<40	Sunny	yes	Fault of driver of motor vehicle	Other Road	Others
3	Grievous Injury	0	1	0	2	0	0	0	1	Urban	40-80	Sunny	no	Fault of driver of motor vehicle	State Highway	Straight Road
4	Minor Injury	0	0	1	1	0	1	0	0	Urban	40-80	Sunny	no	Fault of driver of motor vehicle	National Highway	Curved Road

- ❖ The Accident data set consists of 6488 data points, with 16 features each.

```

➡ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 6488 entries, 0 to 6487
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Accident type          6488 non-null   object
1   Death                  6488 non-null   int64
2   Grievous               6488 non-null   int64
3   Minor                  6488 non-null   int64
4   Driver                 6488 non-null   int64
5   Passenger              6488 non-null   int64
6   Pedestrian             6488 non-null   int64
7   Cyclist                6488 non-null   int64
8   Collision              6488 non-null   int64
9   Area Type              6488 non-null   object
10  Speed Limit            6488 non-null   object
11  Weather                6488 non-null   object
12  Hit Run                6488 non-null   object
13  Cause Accident         6488 non-null   object
14  Type Road              6488 non-null   object
15  Road Features          6488 non-null   object
dtypes: int64(8), object(8)
memory usage: 811.1+ KB

```

- ❖ There is no null values in the dataset

Handling values:

- ❖ Three features, namely "grievous," "minor," and "hit run," were dropped from the accident dataset as they were deemed irrelevant for predicting accident types. The focus is on predicting the broader categories of accidents, encompassing both grievous and minor injuries.

5.2 Data Analysis:



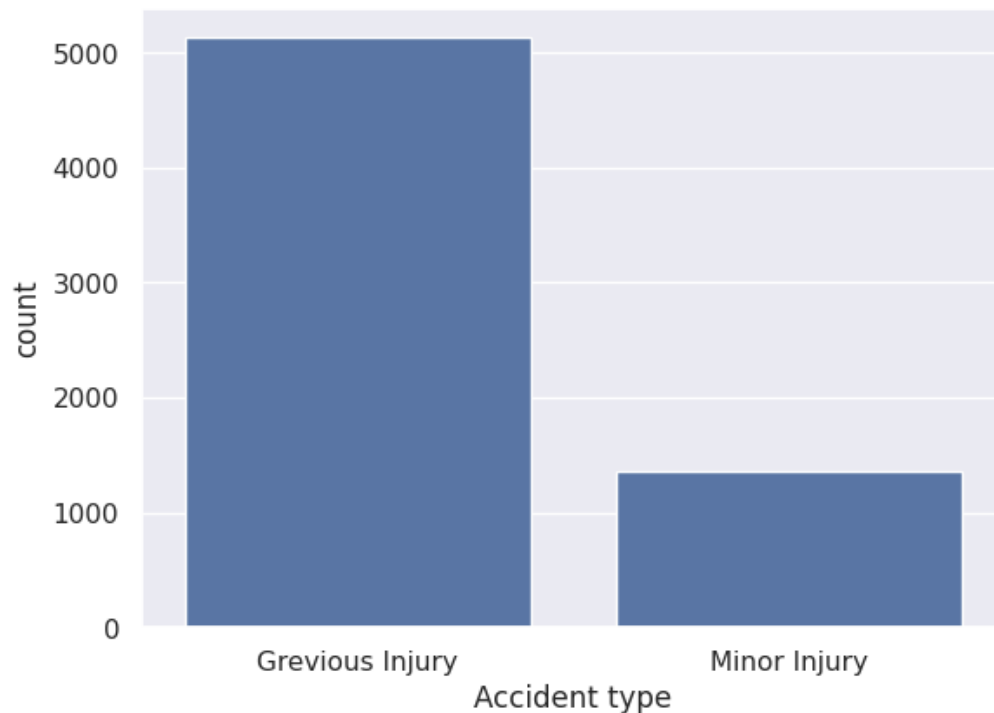
	Death	Driver	Passenger	Pedestrian	Cyclist	Collision
count	6488.000000	6488.000000	6488.000000	6488.000000	6488.000000	6488.000000
mean	0.070746	1.654901	0.182645	0.218095	0.040074	1.049630
std	0.268748	0.496376	0.452212	0.423671	0.197713	0.895497
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	2.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	2.000000	0.000000	0.000000	0.000000	2.000000
max	3.000000	7.000000	8.000000	3.000000	2.000000	2.000000

The output seems to be the summary statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) for various columns in the dataset. In the context of machine learning prediction:

1. Count: Shows the number of data points available for each feature.
2. Mean: Represents the average value for each feature.
3. Std (Standard Deviation): Indicates the extent of variation or dispersion of values around the mean.
4. Min and Max: Display the minimum and maximum values observed in each feature.
5. 25th, 50th (Median), and 75th Percentiles: Provide information about the distribution of values, with the median indicating the middle point of the data.

Understanding these statistics helps in assessing the distribution and characteristics of the data, which is crucial for preprocessing and selecting appropriate machine learning models. For instance, knowing the range and distribution can guide feature scaling or normalization, and identifying outliers or skewed distributions may influence the choice of algorithms.

5.3 Data Visualization:



Grevious Injury: 5130

Minor Injury : 1358

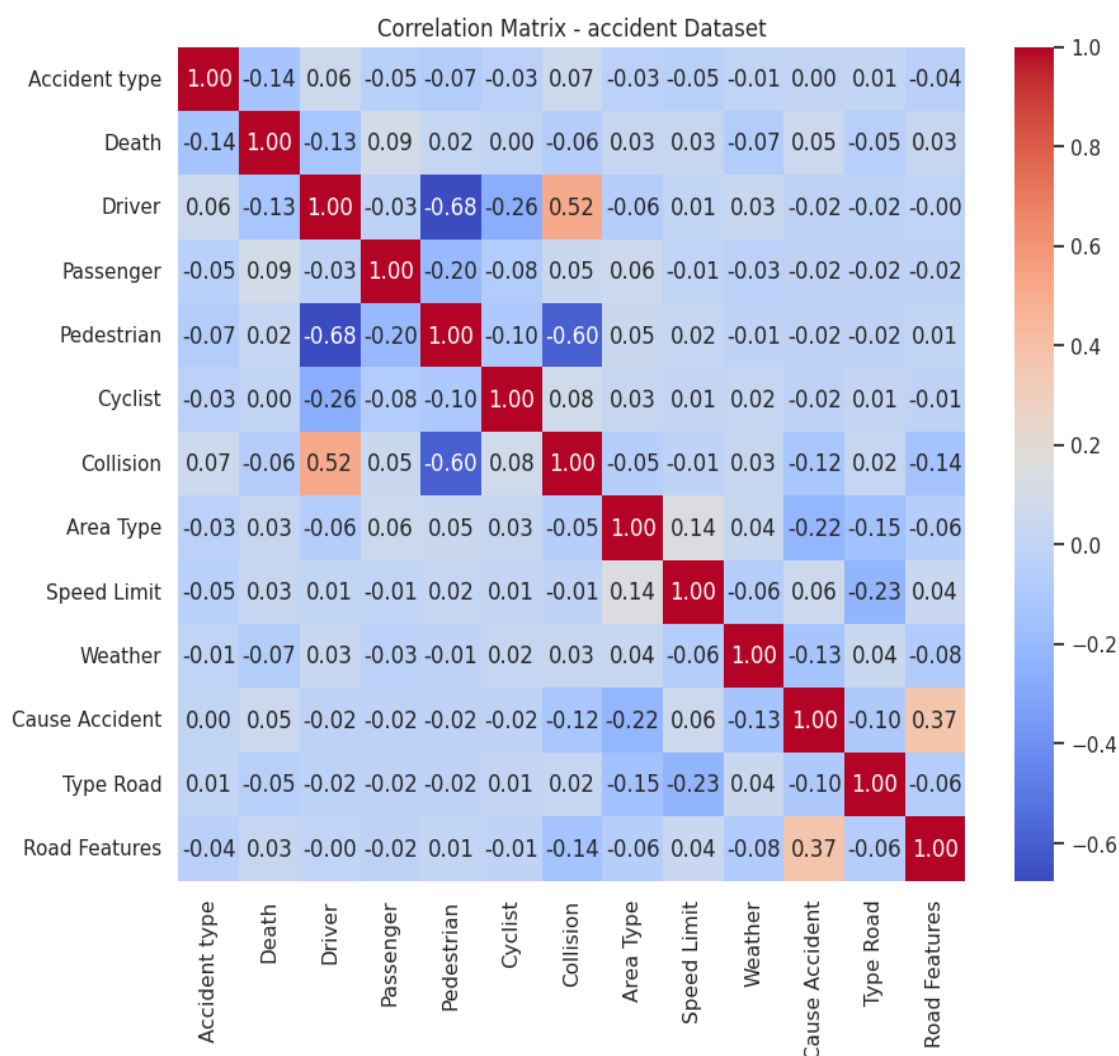
Above countplot depicts the number of grevious injury and minor injury.

5.4 Encoding the categorical columns:

Assigns numerical labels to each category which are area type, cause of accident, weather, speed limit, road features, type road

	Accident type	Death	Driver	Passenger	Pedestrian	Cyclist	Collision	Area Type	Speed Limit	Weather	Cause Accident	Type Road	Road Features
0	1	0	1	0	0	1	2	1	0	2	0	0	1
1	0	0	1	0	1	0	0	1	1	2	2	3	2
2	1	0	1	0	1	0	0	1	0	2	0	3	2
3	1	0	2	0	0	0	1	1	1	2	0	0	1
4	0	0	1	0	1	0	0	1	1	2	0	1	0

5.5 Correlation Matrix:



From the figure we can see that the features Driver and Collision with correlation value 0.52 are moderately correlated.

Also the features Driver and Collision are negatively correlated with the feature Pedestrian with correlation values -0.68 and -0.60 respectively.

5.6 Separating Features And Target

By separating features and targets, provide the model with clear information about the relationship between the input data and the desired output. This separation allows the model to learn patterns and correlations in the training data, enabling it to make predictions on new, unseen data. In the provided code, the variable X represents the features of the accident data excluding certain columns ('Collision', 'Area Type', 'Speed Limit', 'Weather', 'Cause Accident', 'Type Road', 'Road Features', 'Accident type'), while Y represents the target variable, specifically the 'Accident type'.

5.7 Accuracy Comparison:

Algorithms	Training Accuracy	Testing Accuracy
Decision Tree	79%	77%
SVM	78%	79%
K-Nearest Neighbors	73%	71%
Random Forest	79%	77%

The above table shows the accuracy values for all four machine learning algorithms. Also it shows that SVM algorithm gives the best accuracy with 78% training accuracy and 79% testing accuracy.

CHAPTER 6

RESULTS AND CONCLUSION

6.1 FINDINGS

In the examination of road accidents in Ernakulam through the chi-square test, a significant association between gender and road accidents was identified. The data revealed a notably higher incidence of road accidents among males compared to females. Surprisingly, despite the common belief, obtaining education did not lead to a substantial reduction in road accidents. The analysis extended to test the relationship between injury type and road area. The result reveals there was discernible association between injury type and road area. From 2019 to 2023, the primary cause of road accidents can be attributed to the drivers of motor vehicles.

These findings underscore the complexity of factors influencing road accidents, challenging preconceived notions about the impact of education and highlighting the dynamic nature of injury patterns in different years and areas.

6.2 CONCLUSION

The study shows that 96.8% of individuals are knowledgeable about traffic rules, and 77% possess licenses. However, there hasn't been a significant reduction in road accidents between 2019 and 2023. The initial chi-square test indicates a significant correlation between road accidents and gender, with a p-value of .000 which is lower than the significance level 0.05. This suggests that the incidence of road accidents is significantly higher among males compared to females. The second chi-square test results show a p-value of 0.010, which is less than 0.05 the significance level. This indicates that there is no significant association between individuals education backgrounds and road accidents, suggesting that obtaining education doesn't result in a notable reduction in road accidents. The third chi-square test result show a p value 0.005, which is less than 0.05 reveals that an association exist between injury type and road area over the five-year period. Throughout 2019 to 2023, the primary cause of road accidents remains attributed to motor vehicle drivers.

In this study, systematic efforts are made in designing a system which results in the prediction of Accident Type. During this work, four machine learning classification algorithms are studied and evaluated on various measures. Experimental results determine the adequacy of the designed system with an achieved accuracy of 79% using SVM algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict accident type. The work can be extended and improved for the automation of road accident analysis including some other machine learning algorithms

Bibliography

- [1] Singh, S. K. (2017). Road traffic accidents in India: issues and challenges. *Transportation research procedia*, 25, 4708-4719.
- [2] Mohan, D., & Bawa, P. S. (1985). An analysis of road traffic fatalities in Delhi, India. *Accident Analysis & Prevention*, 17(1), 33-45.
- [3] Singh, S. K., & Misra, A. (2004). Road accident analysis: A case study of Patna City. *Urban Transport Journal*, 2(2), 60-75.
- [4] Kakkar, R., Aggarwal, P., Kakkar, M., Deshpande, K., & Gupta, D. (2014). Road traffic accident: retrospective study. *Indian Journal of Scientific Research*, 5(1), 59-62.
- [5] Ballamudi, V. K. R. (2019). Road accident analysis and prediction using machine learning algorithmic approaches. *Asian Journal of Humanity, Art and Literature*, 6(2), 185-192.

ANNEXURE

<p>Name *</p> <p>Your answer _____</p>	<p>Gender *</p> <p><input type="radio"/> Male</p> <p><input type="radio"/> Female</p> <p><input type="radio"/> Other: _____</p>
<p>Age *</p> <p><input type="radio"/> 18- 21</p> <p><input type="radio"/> 22 - 25</p> <p><input type="radio"/> 26 - 30</p> <p><input type="radio"/> 31 - 50</p> <p><input type="radio"/> 51 and above</p> <p><input type="radio"/> Other: _____</p>	<p>Education Qualification *</p> <p><input type="radio"/> High School</p> <p><input type="radio"/> Higher Secondary</p> <p><input type="radio"/> Graduate</p> <p><input type="radio"/> Masters</p> <p><input type="radio"/> Other: _____</p>
<p>Gender *</p> <p><input type="radio"/> Male</p> <p><input type="radio"/> Female</p>	<p>Occupation *</p> <p><input type="radio"/> Employed</p>
<p>Occupation *</p> <p><input type="radio"/> Employed</p> <p><input type="radio"/> Unemployed</p> <p><input type="radio"/> Other: _____</p>	<p>Have you been involved in road accidents during the past year 2019 -2023 *</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p> <p><input type="radio"/> Other: _____</p>
<p>Do you have license ? *</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p> <p><input type="radio"/> Other: _____</p>	<p>How severe was the accident ?</p> <p><input type="radio"/> Minor</p> <p><input type="radio"/> Moderate</p> <p><input type="radio"/> Fatal</p> <p><input type="radio"/> Other: _____</p>

What do you believe are the primary ^{*} causes of road accidents?

☐ Distracted driving (phone usage etc...)

☐ Speeding

☐ Drunken drive

☐ Poor road conditions

☐ Weather conditions

☐ Mechanical failure

☐ Others

☐ Other: _____

On average, how many kilometers do ^{*} you drive per day?

☐ 1-10 km

☐ 10-20 km

☐ 20-30 km

☐ 30-40 km

☐ 40-50 km

☐ above 50 km

☐ Other: _____

How often do you exceed the speed ^{*} limit while driving?

☐ Frequently

☐ Sometimes

☐ Rarely

☐ Never

☐ Other: _____

Are you aware of traffic rules and ^{*} regulations?

☐ Yes

☐ No

☐ Other: _____

Are you more conscious about road ^{*} safety measures after implementation of AI cameras

☐ Yes

☐ No

☐ Maybe

☐ Other: _____

Have you ever received formal driver safety training?

☐ Yes

☐ No

☐ Other: _____

Do you regularly use seatbelts?

☐ Yes

☐ No

☐ Other: _____