Project Dissertation Report on

COMPREHENSIVE ANALYSIS OF ACCIDENTAL MORTALITY AND TRAFFIC ACCIDENTS IN INDIA (2012-2022)

Submitted By:

Shruti Negi

24/DMBA/224

Under the Guidance of

Dr. Shikha N. Khera

Associate Professor

DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road Delhi 110042

CERTIFICATE

This is to certify that the Major Research Project titled "COMPREHENSIVE ANALYSIS OF ACCIDENTAL MORTALITY AND TRAFFIC ACCIDENTS IN INDIA (2012-2022): A BUSINESS INTELLIGENCE APPROACH" is a bonafide work carried out by Shruti Negi (24/DMBA/224), a student of MBA at Delhi School of Management, Delhi Technological University during the Academic Year 2025-2026.

The project has been completed under my guidance and supervision. To the best of my knowledge, the matter embodied in this report has not been submitted to any other University or Institute for the award of any degree or diploma.

(Signature of Supervisor)

Dr. Shikha N. Khera
Associate Professor
Delhi School of Management (DSM)
Delhi Technological University (DTU)

DECLARATION

I, Shruti Negi, Roll No. 24/DMBA/224, student of MBA (2024-2026) at Delhi School of Management, Delhi Technological University, hereby declare that the project report entitled "COMPREHENSIVE ANALYSIS OF ACCIDENTAL MORTALITY AND TRAFFIC ACCIDENTS IN INDIA (2012-2022): A BUSINESS INTELLIGENCE APPROACH" submitted by me is an original work carried out by me under the supervision of Dr. Shikha N Khera.

I also certify that this work has not been submitted to any other institution for the award of any degree, diploma, certificate or other similar qualification degree or diploma. Sources of information have been properly referenced and properly credited.

Shruti Negi
24/DMBA/224
Delhi School of Management (DSM)
Delhi Technological University (DTU)

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, **Dr. Shikha N. Khera**, for providing the necessary guidance, encouragement, and insightful feedback throughout the duration of this Major Research Project. Their mentorship was instrumental in helping me navigate the complexities of data-driven analysis and strategic research.

I extend a special note of thanks to Prof. P.K. Suri for his mentorship and guidance in the core subject, which provided me with the foundational knowledge and technical skills essential for the successful execution of this study.

I would also like to sincerely thank Miss. Ishita Patwal (DSM, DTU) for her support in the completion of this project.

I am also grateful to the **Delhi School of Management (DSM)** and **Delhi Technological University** for providing the academic resources and environment required to complete this project.

Finally, I wish to thank my family and friends for their unwavering support, which allowed me to focus my efforts on this research during the **2025-2026** academic year.

Regards,
Shruti Negi
24/DMBA/224
Delhi School of Management (DSM)
Delhi Technological University (DTU)

EXECUTIVE SUMMARY

This research project is a detailed Business Intelligence (BI) study covering the last eleven years (2012-2022) of accidental death and traffic accidents in India. Despite great efforts to address public health concerns and infrastructural developments, accidental deaths still continue to be a significant socio-economic issue in India, often claiming over 400,000 lives per year. The main aim of this study was to move away from traditional descriptive reporting of mortality and into a more advanced analytical framework that is predictive and prescriptive. This study will involve applying machine learning algorithms to identify the patterns of mortality, predict mortality rates in the near future and classify geographic areas by risk so that interventions can be targeted to the appropriate areas.

The methodology involved the use of comprehensive data, including more than 218 aggregated state and temporal records, with the exception of the suicide data, which would not provide an analytic focus on preventable accidental mortality. A multi-phased analytical approach was used. In the descriptive phase, the analysis showed that traffic accidents are the main cause of death, accounting for 46.0% of the deaths from all causes (194,347 deaths were caused by traffic accidents in 2022 alone). The data also showed a large demographic vulnerability, with a ratio of 4.77 male deaths for every female death, and a significant proportion of deaths occurred in the working age population (18-45 years).

Several machine learning models were created and tested during the predictive phase. A Linear Regression model successfully showed that the national death rate has a high correlation with natural and unnatural causal factors ($R^2 = 0.86$), and predict the national death rate well. At the same time, regional data points were clustered using K-Means clustering to form three risk segments.At the same time, the regional data points were clustered using K-Means clustering to form three risk segments. This concentration brought out stark geographic differences, such as the dangerous areas in Bihar and Punjab with almost 80% fatality rates, while the less

risky areas had rates of 28.9%. Also, a Logistic Regression classification model was developed to act as an early warning system, which successfully classified high, medium or low risk years using demographic and causal indicators.

The ultimate goal of the research is to provide prescriptive recommendations for policies. The use of the developed risk clusters proposes a data-driven and decentralized disaster management and road safety approach. Some of the key recommendations are that the identified "High-Risk" clusters be given more funds in emergency infrastructure budgets and that predictive models be used to define localized Key Performance Indicators (KPIs) for the National Road Safety Mission.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

Unintentional injuries and accidental mortality represent a profound global public health crisis. Despite rapid advancements in automotive safety, occupational hazard regulations, and disaster management, traffic-related incidents and natural fatalities continue to be leading causes of death worldwide, particularly among the working-age population. Within this global context, India bears a disproportionately high burden. Historical public health data compiled by authorities such as the National Crime Records Bureau (NCRB) illustrates a grim reality: accidental deaths in the country routinely exceed 400,000 cases annually. The accidental deaths increased from 394,982 in 2012 to 430,504 in 2022, making an increase of 8.99%.

The high mortality volume has socio-economic implications in the country. In addition to the tragic loss of life, these deaths have significant healthcare costs, permanent disability costs and significant costs to the national Gross Domestic Product (GDP). In the past 10 years, the Indian government has taken several legislative and infrastructure measures to reduce these figures: its strengthened traffic police efforts and improved early warning systems for natural disasters. Others have shown progress, for example, the dramatic drop in numbers of deaths due to forces of nature (64.9%), but the impact of widespread policy changes has not been apparent in the number of traffic accident deaths, or indeed preventable deaths in general.

For the past history, the traditional method of reporting in India is descriptive in nature which involves the tabulation of the number of deaths per year. But with the modern data ecosystem, it's time to move beyond just reporting into more advanced Business Intelligence (BI). Predictive and prescriptive analytics now give us an unprecedented opportunity to shift from detecting what has happened in the past to predicting future trajectories of mortality and figuring out what to do to prevent them. In this research, that technological shift is embraced. Uses machine learning algorithms to convert raw, causal, geographic and demographic mortality data into targeted and actionable policy frameworks.

1.2 Problem Statement

Despite several public health efforts, accidental deaths in India continue to be a serious problem, and result in hundreds of thousands of deaths every year. Road safety and disaster management are traditionally managed and executed as reactive rather than proactive processes, using a data driven method. A substantial lack of leveraging full-fledged Business Intelligence (BI) and predictive analytics to gain insight into the complex, multi-variable causes of these deaths. In particular, no risk segmentation and prediction is done at the state level, making it difficult to allocate resources effectively. This work tackles this issue by using advanced analytics such as Linear Regression and K-Means clustering to analyze historical mortality data (2012-2022) to identify high-risk states and predict future mortality rates, which can then inform targeted and prescriptive policy interventions.

1.3 Objective of the study

The overall goal of this research is to inform public health mortality policies by moving public health mortality data into a business intelligence framework. In particular, the study is directed towards three main analytical goals, namely:

1. Descriptive analysis: To systematically consider temporal patterns, demographic characteristics (most affected age group being 18–45), geographic distributions and dominant causal factors of the accidental deaths in India during 2012–2022.
2. Predictive Analysis: To model and test machine learning algorithms (Linear Regression) to forecast national death rates for the near future, and to use K-Means clustering to mathematically partition regional entities into distinct operational profiles of low, medium and high risk.
3. Prescriptive Analysis: To move beyond population-based budget distribution to risk-based resource allocation, and to make evidence-based, actionable recommendations for policymakers.

1.4 Scope of Study

This analytical research is geographically bound only to the Republic of India . To the eleven year period from 2012 to 2022. Data analysis is done quantitatively using comprehensive secondary data that includes 218 aggregated data across five different datasets, namely: Year-wise Overall details, Demographics of Forces of Nature, Demographics of Other causes, Cause-wise YoY comparisons, and State-wise Traffic Accidents. Geographically, the spatial analysis and K-Means clustering analyzes 89 different geographic data points which cover all known States, Union Territories, and major metropolitan urban centres reported in 2022.

Methodologically, it is restricted to usage of descriptive statistical aggregation, Linear Regression forecasting, K-Means clustering and Logistic Regression classification. Most importantly, this study does not include any information on intentional self-harm or suicides, which is different from most public health databases that may combine intentional and unintentional deaths. This exclusion helps the research to be focused and undistorted analytically on preventable accidental death and disaster management.

1.5 Significance of the Study

Unintentional fatalities in India go far beyond a standard public health issue, they represent a severe demographic and economic crisis. As the baseline data indicates, the overwhelming burden of accidental deaths is heavily on the working age male population, mostly within the 18 to 45 age bracket. Because age group frequently serve as the main financial providers for households, their sudden loss triggers socio-economic effects, including multi-generational poverty and contractions in national economic productivity. Consequently, identifying the infrastructural mechanisms that cause these fatalities is of significance to protecting the nation's macroeconomic stability.

From an administrative and policy perspective, the significance of this study lies in its direct challenge to conventional disaster management frameworks. Historically, government agencies have allocated safety and infrastructure budgets largely based on regional population sizes, treating vast, demographically heterogeneous states as uniform operational units. This research disrupts that generalized approach.

By mathematically segmenting regions into distinct risk clusters based on actual accident severity and fatality rates, the study provides a concrete blueprint for a "risk-weighted" resource allocation strategy. It arms policymakers with the empirical evidence required to direct heavy capital expenditure such as decentralized highway trauma centers specifically to the geographical pockets where survivability is mathematically the lowest.

Finally, the methodological significance of this project rests in its deliberate transition from retrospective reporting to proactive Business Intelligence. By explicitly stripping out intentional mortality (suicides) from the national datasets, the analysis successfully isolates purely preventable, infrastructural failures. The subsequent application of predictive machine learning algorithms allows public health administrators to stop reacting to the previous year's fatality counts and instead anticipate near-future mortality trajectories. Ultimately, this research provides a practical, data-driven tool designed to help initiatives like the National Road Safety Mission transition from simply tabulating historical accidents to actively predicting and preventing them.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction to the Literature Review

A comprehensive, multi-disciplinary framework is needed to study accidental mortality, which brings together areas of public health epidemiology, spatial disaster management, demographic economics, and advanced data science. While countries with developing economies are experiencing a rapid change in infrastructures and population, the character of unintentional deaths is also changing. This chapter analyzes the current academic research, government and international public health reports, which include unintentional deaths. It is organized in a way that it starts with the historical context and epidemiology of accidental deaths in India. After that, it critically examines the present-day use of predictive analytics, machine learning, and Business Intelligence (BI) in public health research, particularly regression forecasting and spatial clustering. Lastly, it reviews the existing approaches and concepts related to the evaluation and implementation of road safety interventions at present and sets the theoretical and methodological basis for the current study.

2.2 Previous Studies on Accidental Mortality in India:

2.2.1 The Epidemiological and Demographic Burden

Over the last 20 years, unintentional injuries and accidental deaths have been under the academic and governmental radar in India. In the past, public health reporting has been largely dependent on the annual descriptive tabulations made by the National Crime Records Bureau (NCRB) and the Ministry of Road Transport and Highways (MoRTH). These annual reports give vital base line counts, but academic researchers have been trying to decipher the underlying demographic and socio-economic meaning of the numbers. In their groundbreaking study, Mohan et al. (2009) showed early on that in rapidly motorising developing countries, the traffic-related mortality rate is high per number of vehicles in use. The failure to provide adequate facilities, combined with the lack of separation of pedestrians, two-wheelers, and heavy commercial vehicles in the same spatial corridor and poor enforcement of safety measures is a major contributor to this vulnerability.

Recently, extensive epidemiological investigations have pointed to an ever-changing death rate. Dandona et al (2020), published in The Lancet Public Health, identified a critical epidemiological transition at State level. Road traffic injuries are rapidly emerging as the major cause of mortality for young adults in India compared to conventional infectious diseases. Literature stress the point that this is not just a public health issue but demographic one too. There is a modern public health emphasis that the absolute burden of unintentional injuries is borne by working-age males. Historical demographic data reviewed in each of the studies reviewed shows that there are significant gender and age discrepancies, sometimes with a ratio of nearly 5:1 men dying for every 100 women and high male mortality in the age group 18-45. This particular segment of the population is the backbone of the economic activities of overwhelming majority of the households in India, thus the downstream socio-economic impact is high. When primary breadwinners of the households are killed in avoidable accidents, leading to multi-generational poverty, there is a macro-economic loss that is estimated at being several percentage points of a country's Gross Domestic Product (GDP) (Singh, 2017).

2.2.2 Climate-Induced Mortality and Natural Disasters

In addition to traffic accidents, the literature on "Forces of Nature" and natural disaster mortality suggests that there has been a large variation in public health outcomes over the past decade. Past disaster management research shows that warnings and evacuation measures at the local level have been effective in reducing loss of life in certain predictable natural disasters, like coastal cyclones, and monsoon floods. But new and extremely variable mortality variables have been observed in the Indian subcontinent in recent epidemiological literature, due to climate anomalies.While these changes are evident, there is a lack of literature on how to represent the climate-induced death rates in a comprehensive predictive public health model. Most environmental health papers describe what has happened in the past, and have been effective in identifying areas which they describe as being at risk, but have been less effective in generating geographically-specific risk models that can predict and direct future administrative response.

2.2.3 Unnatural Preventable Causes: Drowning, Falls, and Occupational Hazards

Unnatural but significantly preventable causes of accidental deaths such as drownings, accidental falls and electrocution are also covered in the literature but are often dominated by traffic figures. In rural public health research, it is noted that drowning is a leading cause of death among children and adolescents, and is largely attributable to the absence of formalized swimming education in rural school curricula and unprotected water bodies and open wells. Likewise, in the context of occupational health and safety in India, it has been noted that the rapid urbanization and industrial growth has resulted in the continued occurrence of fatal accidents due to structural collapse, accidental fire and fall in industry. Academic critiques of existing public health data collection systems have therefore concluded that administrative databases should not only be used to collect data after an incident, but for active and predictive hazard management.

2.3 Application of Predictive Analytics and Business Intelligence in Public Health

2.3.1 The Transition to Data Science and Business Intelligence

The shift from descriptive epidemiological reporting to sophisticated Business Intelligence (BI) is a major step, A significant methodological change in the administration of modern public health. Historically, disaster management and current and planned allocation of healthcare resources has been reactive, based on backward-looking indicators like, the number of fatalities in the previous year to allocate emergency budgets. Predictive analytics is the use of this technology to predict customer behavior. To use algorithms and machine learning to these historical data sets to make predictions, Discovering hidden patterns of multiple variables and changing policy from a reactive to proactive approach is benificial.

2.3.2 Linear Regression and Epidemiological Forecasting

Algorithmic methods like Linear Regression have become more widely used in studying mortality data for modeling the trajectory of mortality based on the

underlying composition of causal factors. The academic literature, however, cautions against learning from mathematics when moving from descriptive to predictive approaches in the field of public health data science. One such problem is the statistical ‘target leakage' that can easily occur in predicting an absolute total number of fatalities from the exact constituent totals being used as independent variables, since the ‘target' is the absolute total number of fatalities, which is always the sum of the independent variables. In modern epidemiology, therefore, careful analytical structures are devoted to predicting the death rate per lakh of population for each of different conceptual categories, like natural forces versus man-made accidents. If designed properly and without multicollinearity, and tested against historical test sets, these models of regression provide extremely accurate early warning. They enable policymakers to predict the possibility of a national death rate surpassing thresholds in the coming years if indicators of changing environmental and/or infrastructural conditions change.

2.3.3 Spatial Epidemiology and K-Means Clustering

At the same time, unsupervised machine learning, especially K-Means clustering, has been hugely popular in the study of spatial epidemiology, the geography of resources and so on. Traditional public health organization is based on the arbitrary political borders and assigns large, demographically diverse states as one unit of operations. The literature on spatial analytics questions the use of this generalized concept in hiding hyper-localized crises and in causing inefficient distribution of resources. K-Means clustering enables scholars to mathematically divide geographical areas into clusters on the basis of multivariate risk profiles, instead of simply geographic area size or population. When the region's entities are clustered according to mathematical properties, such as the volume of accidents, injury severity ratios and absolute fatality rates, potential vulnerabilities are not apparent and can be revealed. This computational solution enables the policy-maker to break away from blanket national policy solutions. Rather, it allows a decentralized, risk-weighted approach to public health such that clusters of greater risk and the potential for higher fatalities are served by more capital-intensive trauma infrastructure, and clusters of higher volume, but lower risk, are served by entirely different interventions – including AI-based traffic management technologies.

2.3.4 Classification Algorithms for Early Warning Systems

The inclusion of classification algorithms, such as Logistic Regression, is another step forward in the development of Business Intelligence framework to promote the development of active early warning systems. Automated risk categorization has been mentioned in many literatures concerning disaster management technology as a critical necessity. Classification algorithms enable administrators to get immediate, threshold-based alerts, after creating a classification of historical periods or specific geographic zones into low, medium, or high-risk stratifications, by using leading indicators, such as population density, temporal trends, and causal death compositions. The academic opinion is that the use of these machine learning pipelines provides a pro-active public health approach, and that the paradigm is changing from tabulation after a disaster to prevention and preparedness before the disaster.

2.4 Evaluating Road Safety Interventions and Prescriptive Analytics

2.4.1 Global Frameworks and the Safe Systems Approach

Prescriptive Analytics represents the last segment of the Business Intelligence pipeline, the link between predictive algorithmic foresight and executing a policy. The basis for the evaluation and design of road safety interventions is largely informed by the philosophy of the World Health Organization (WHO) global approach, 'Safe Systems', which is at the heart of the United Nations' Decade of Action for Road Safety (WHO 2018/2023). This approach calls for a comprehensive paradigm change in the way traffic is currently managed; it suggests that human failures on roads are a fact of life but that traffic systems, vehicle designs and road layouts need to be specifically engineered to anticipate and forgive human mistakes to avoid catastrophic consequences. It takes away from the individual driver the responsibility for safety completely.

2.4.2 Limitations of Legislative Enforcement in India

The Indian scenario shows that in the last few years, the effectiveness of different road safety measures has been assessed in schools, with legally binding measures proving ineffective in reducing the number of fatalities. After the Motor Vehicles (Amendment) Act was enacted in 2019, which had dramatic hikes in monetary fines

and legal penalties for traffic offenses, researchers observed a complicated consequence. Although there was some localised improvement in compliance in the major metropolitan areas where there was a high policing presence, the aggregate national fatality rate was still persistently high, notably in rural and semi-urban corridors. A comprehensive report by the World Bank in 2021 on traffic crash injuries and disabilities in India pointed out that punitive measures should be inseparably linked to infrastructure improvements in order to achieve substantial reductions in deaths.

2.4.3 The Golden Hour and Emergency Medical Response

Most importantly in the public health literature, the quality of the emergency medical service response following a crash and the quality of the rural health care system for trauma are the most important factors influencing the severity of traffic deaths. The "golden hour" concept of the first 60 minutes after a traumatic injury is a recurring theme in public health studies, and it is well established that time is of the essence when it comes to specialized medical care after trauma. It is known from literature that, in some Indian states, this golden hour is very often missed and fatality rates are significantly exacerbated. This failure has been blamed on response times to ambulances, the lack of decentralised highway trauma facilities and the lack of procedures for rapid response.

2.4.4 Risk-Weighted Resource Allocation and KPIs

In contemporary literature, it was said that prescriptive frameworks must be adapted. The dominant public administration analytic argument is for the need to move from population-weighted allocation of budgets to risk-weighted allocation of resources. These strategies are particularly important in high-volume, high-density areas, where the use of AI algorithms for traffic management, automated speed cameras, and pedestrian-only zones can help reduce congestion and improve safety. On the other hand, the states that have low infrastructures and high fatality rates are in dire need of huge investments in order to develop medical corridors for fast response and in upgrading basic highway engineering. The combination of K-Means geographic risk segmentation and globally known Safe Systems frameworks enables researchers to develop KPIs for each geographic segment. The prescriptive approach, based on

data, ensures that national budgets, including those provided under the National Road Safety Mission, are used exactly where mathematical modeling says they can make a life-saving difference, which is to say, to save lives.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The main goal of this study is to move the Public Health data on accidental deaths in India from reportage to an advanced Business Intelligence (BI) framework. The study is conducted by quantitative and data driven approach, which includes descriptive statistics, unsupervised machine learning (clustering) and supervised machine learning (regression and classification). This chapter details the strict protocols followed in data collection and the multi-stage cleaning of the data, as well as the mathematical underpinnings of the algorithms used and the steps taken to preserve data quality and integrity.

3.2 Data Collection (Sources and Approach)

The integrity, granularity and scope of the data is the key to any strong Business Intelligence pipeline. A secondary data collection method with a quantitative approach was required for this research. The main goal at the time of data collection was to create a multi-dimensional database that would adequately reflect the changing trends of accidental deaths in the Indian subcontinent.

For this purpose, the researcher used only the official government repositories. The primary data has been collected from the annual statistical tabulation done by the National Crime Records Bureau under Ministry of Home Affairs and some infrastructural and highway data has also been provided by the Transport Research Wing of the Ministry of Road Transport and Highways. The time period for extraction was strictly limited to an 11-year period from the calendar year 2012 to 2022, and included the most recent verified complete decade of public health data at the time of the analysis.

One of the methodological decisions during data collection was to exclude intentional death. This study specifically separated all datasets, variables, and records that included information on intentional self-harm and suicides from accidental deaths and suicides, which are often combined in an administrative database. This isolation was required for close analytical attention to the issue of preventable accidental deaths, the lack of infrastructure support, and disaster management procedures.

The raw data extraction returned thousands of individual rows. These were carefully aggregated together and bundled into five main, structured, datasets, prepared for ingestion by machine learning. Two hundred and eighteen records were fused into the final working database. The types of structural breakdown, time periods and the key variables studied in the datasets collected are described

Table 3.1:Summary of Extracted Datasets

Dataset Name	Records	Time Period	Key Variables Analyzed
Year-wise Details	11	2012–2022	Annual deaths by cause, estimated population, baseline death rates
Forces of Nature (Demographics)	14	2022	Fatality distributions by specific age cohorts and gender
Cause-wise Comparison	39	2021–2022	Specific mortality causes, percentage shares, year-over-year variation
Other Causes (Demographics)	61	2022	Age and gender distribution for unnatural, human-induced accidents
Traffic Accidents (Regional)	93	2022	Total traffic cases, total deaths, total injuries segmented by state
Total Records Analyzed	**218**	—	—

Source: Researcher's Compilation from NCRB Historical Data (2012-2022).

3.3 Data Cleaning Process and Integrity Validation

National Crime Records Bureau (NCRB) and Ministry of Road Transport and Highways (MoRTH) gave us a vast but unstructured pool of data regarding public health and infrastructural data. The raw data from administrative databases is primarily intended for human use for administrative records, not for ingestion into machine learning pipelines. To address this, a large, multi-stage pipeline for data cleaning and preprocessing was designed using the Python programming language, with the use of the pandas and numpy library for computation.

3.3.1 Phase 1: Data Inspection and Structural Normalization

The first step in the data wrangling process was a merely systematic review of raw CSV files. The exploratory data analysis uncovered an important structural problem with government data — the existence of which has been ignored — embedded hierarchical sub-totals. Summation rows, year-over-year percentage calculations and national totals were often located within the state-level data rows . These summary rows will result in significant double counting issues when performing the Descriptive Aggregations, as well as make the coefficients appear higher while training the Linear Regression. The researcher used Python to systematically remove all summary rows which were not numeric. In addition, the naming of the columns was very inconsistent between the five different datasets—for example, different spellings of the same state name and inconsistent headers for temporal data. All column names were normalized using a strict normalization protocol to create a consistent syntax that can be read by machines, such that relational database joins are possible over time and location on the temporal and spatial data sets.

3.3.2 Phase 2: Missing Value Handling and Type Conversion

Data sparsity and type inconsistencies were the next challenge for the second phase. A scan of all the aggregated records (218) was performed to identify null values (NaN). The common approach in traditional data science techniques involves filling in missing data with the mean or median values. The researcher however understood that mean imputation to absolute mortality counts would cause serious misrepresentation in the local geographical realities and create mathematical bias in the models.

Therefore, a domain-specific approach was used: If no incidents were reported for a certain rural district, the corresponding blank was explicitly filled with the integer zero, not with an average value. Further, the raw data often included special characters and commas as part of the number. These strings were automatically (programmatically) cleaned from non-numeric characters and were strictly converted into float and integer continuous data types to allow for more sophisticated distance computation operations for the subsequent spatial clustering algorithms.

3.3.3 Phase 3: Feature Engineering and Mathematical Derivations

The researcher created several new analytical features to change the analytical lens from "absolute volume" to "proportional risk. The most important derived metric designed for the spatial analysis was the Fatality Rate. This calculation inherently weights the analysis for each state according to their population as the total number of traffic deaths divided by the total number of traffic accident cases, then multiplied by one hundred is the Fatality Rate. This was a feature that was created to make it possible for the machine learning algorithms to analyse the actual severity and survivability of accidents on the road within the region, without taking into account traffic volumes in the region. In addition, mathematical calculations of continuous year-on-year percentage variation indicators and mathematical calculations of proportional demographic ratios (e.g. the male vulnerability index compared to the female vulnerability index) were derived to be used as independent variables in the predictive modelling phases.

3.3.4 Data Integrity Validation

After the data transformation and feature engineering pipelines were run, a thorough data integrity validation process was performed. This was necessary to guarantee that the data sets were not corrupted and completely ready for the Linear Regression and K-Means clustering algorithms. This final database was checked against five dimensions of data quality which are widely recognized:

Table 3.2: Comprehensive Data Quality Assurance Matrix

Quality Dimension	**Verification Status**	**Operational Finding and Justification**
Completeness	Verified Pass	No missing values remained in key columns like Total Deaths or Total Cases after cleaning.
Consistency	Verified Pass	Verified in Python that state-level totals exactly match national sub-totals.
Accuracy	Verified Pass	Data points were validated with official published baselines from the Transport Research Wing.
Timeliness	Verified Pass	The dataset has recently verified complete decade, bounding the analysis to the 2012-2022 scope.
Validity of Outliers	Verified Pass	Extreme statistical outliers identified then investigated well. Some regions were cross-validated. Researcher confirmed these are not data entry errors, but rather accurate mathematical representations of severe localized infrastructure lack and a lack of immediate trauma care. Hence, these outlier values were correctly retained to give to the K-Means spatial clustering algorithm.

Source: Researcher's Compilation from NCRB Historical Data (2012-2022).

3.4 Machine Learning Algorithms and Model Formulation

The researcher ensured that the base of the Business Intelligence pipeline was mathematically correct by carefully carrying out these data cleaning and validation tasks. Data integrity was verified, and thus the research continued with the deployment of the machine learning algorithms

3.4.1 Linear Regression for Mortality Forecasting

The transition from descriptive historical observation to proactive public health administration requires the deployment of supervised machine learning algorithms. In this predictive phase, Ordinary Least Squares Linear Regression was applied to the chronological datasets to forecast near-future mortality trends. To rigorously evaluate the underlying drivers of accidental deaths, the researcher developed two distinct predictive models, each testing a different administrative hypothesis.

3.4.2 Model 1: Baseline Demographic Forecasting (Total Deaths Prediction)

Supervised machine learning algorithms are essential for shifting from descriptive historical observation to making use in proactive public health administration. For this predictive phase, Ordinary Least Squares Linear Regression was used on the chronological datasets to predict future mortality trends in the near future. The researcher created two models to test two different administrative hypotheses in order to rigorously assess the underlying causes of accidental deaths. The first Linear regression model (Baseline Demographic Forecasting, Total Deaths Prediction) included only the chronological Year, and the estimated National Population as independent variables to predict the total annual accidental deaths. The aim of this particular algorithm was to see whether one could predict mortality trends reliably from general temporal and demographic trends and, in fact, to test the hypothesis that the deaths are accidental and are linear functions of time and population growth.

3.4.3 Mathematical Foundation of Model 1

Linear regression model to create mathematical relationship between independent predictor variables and continuous dependent target variables. This base model algorithm was built with the following multiple linear regression equation:

Total Deaths = Intercept + (Coefficient1 × Year) + (Coefficient2 × Population) + Error

Where:

- **Total Deaths** represents continuous target variable being predicted.
- **Year** represents the chronological time sequence, functioning as the first independent feature to capture linear temporal trends.
- **Population** represents the estimated national population (measured in lakhs), functioning as the second independent feature to capture demographic scale.
- **Intercept** represents the model constant, mathematically defining the baseline value of the target when all independent variables are zero.
- **Coefficients (Coefficient1 and Coefficient2)** represent the calculated slopes for each feature, dictating the expected marginal change in Total Deaths for every one-unit increase in the respective independent variable.
- **Error** represents unexplained residual term that the linear features are unable to capture.

This particular mathematical model was designed to test the linear time assumption and the linear population against actual numbers of deaths.

Data Preparation and Algorithmic Execution

The researcher used the Python programming language to implement this mathematical model, and also took advantage of the Scikit-Learn machine learning library. The temporal dataset containing eleven aggregated records from 2012 to 2022 was split into a train-test split methodology that is strict, unbiased, and used to test the predictive power of the algorithm.

The data was split so that the algorithm was trained with 80% of the data (nine chronological years) and the optimal weights for the coefficients calculated from the remaining 20% of the data. The other 20% of the records were kept in a hidden test set to test the generalization of the model to new data, two chronological years. The splitting function was given a random state seed (42) to assure the mathematical reproduction of the results in various execution environments.

This data preparation and model training stage is implemented as follows with Python code snippets:

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import numpy as np
```

```
# Prepare independent features (Year and Population) and target variable
X1 = year_wise_clean[['Year', 'Population_Lakh']].values
y1 = year_wise_clean['Total_Deaths'].values

# Execute an 80/20 train-test split for unbiased algorithmic evaluation
X1_train, X1_test, y1_train, y1_test = train_test_split(
    X1, y1, test_size=0.2, random_state=42
)

# Initialize the Ordinary Least Squares algorithm and fit to training data
lr_model1 = LinearRegression()
lr_model1.fit(X1_train, y1_train)

# Generate predictions for variance analysis
y1_pred_train = lr_model1.predict(X1_train)
y1_pred_test = lr_model1.predict(X1_test)
```

Coefficient Extraction and Hypothesis Rejection

After the successful fitting algorithm to the training data, mathematical parameters of the model were extracted and analyzed. This algorithm gave an Intercept of -47,297,508.27. Sequentially, each additional year contributes an average of about twenty-four thousand deaths to the baseline, because the Year Coefficient was calculated to be 24,596.76. On the other hand, the figure for the Population Coefficient was computed as -146.31, which is a mathematical paradox that a bigger population leads to fewer deaths.

When these extracted parameters were tested against the unseen test data, the model totally failed. The model had a very low Coefficient of Determination (R-squared) of 0.1433 for the training data but a very high (and negative) R-squared value of -197.94 on the testing data. In the field of data science, a negative r-squared means that the prediction model fails to do a good job at predicting the data, compared to just predicting the average of the data points.

This statistical failure is an extremely significant and useful analytical result of this research. Clearly demonstrates that the relationship between time, population growth

and accidental mortality is essentially non-linear. The complexity and the unpredictability of the external factors significantly influence the number of deaths for the country at large - for instance, sudden legislative changes (e.g. Motor Vehicles Amendment Act) or the quick modernization of infrastructure, changes in behaviour and, of course, all the large environmental anomalies (pandemic-related lockdowns in 2020). As a result, Model 1 shows in an empirical way that it is incorrect to try to predict public health crises by merely tracking population growth: a much more sophisticated, cause-based predictive algorithm is needed.

Extracted Model Parameters and Coefficient Interpretation

The eighty percent training data were used to successfully initialize and fit the Ordinary Least Squares algorithm, and then the mathematical parameters of the regression model were taken for analysis. In a multiple linear regression setting, these coefficients are the computed values for the slopes for each of the predictor features, and the amount of change that should be seen in the target variable for one unit increase of each of the predictor variables. The pulled values are listed in the parameter matrix below.

Table 4.3: Model 1 Extracted Coefficients and Parameter Interpretations

Component	Calculated Mathematical Value	Empirical Interpretation and Operational Meaning
Intercept (Constant)	-47,297,508.27	Represents the theoretical baseline value of total accidental deaths if the chronological Year and Population were mathematically reduced to absolute zero.
Year Coefficient	24,596.76	Indicates that holding population constant, every sequential passing year mathematically adds approximately 24,597 accidental deaths to the national baseline.
Population Coefficient	-146.31	Indicates that for every one lakh (one hundred thousand) increase in the national population, the model expects total accidental deaths to decrease by roughly 146.

Source: Researcher's Python Machine Learning Output.

Empirical Analysis of the Derived Weights

A examination of these extracted coefficients by data science shows concerning anomalies in the baseline demographic hypothesis in this study.

First, the mathematical Intercept is computed at a super-low value of more than 47 million. In real world public health (a country can't have a negative death rate) the intercept is theoretically impossible, but in order to fit the massive scale of the "Year"

feature (2012, 2013 etc), it is mathematically necessary that it be far below the Y-axis.

Second, the Year Coefficient of 24,596.76 indicates a high and positive time correlation. The passage of time as a measurement of increasing infrastructural stress , this is taken in case of this mathematical model. It concluded that the transition from one calendar year to the next automatically implies an extra two dozen thousand deaths, likely reflecting the underlying annual increase in nations' motorization and highway building.

But the most important determination from this parameter extraction is the Population Coefficient, which came out as -146.31. It's a huge statistical paradox in predictive analytics. The first hypothesis made by administration was that when the size of a country's population increases the absolute number of accidental deaths would grow proportionately. However, this feature was given a negative weight, implying mathematically that population growth is a repression of accidental mortality.

This counter-intuitive coefficient strongly suggests that there is a problem with the model's explanatory power: it is not genuine. Population growth is very linear and stable, whereas accidental deaths are very volatile and vulnerable to sudden shocks (such as the spike in deaths in 2014 and the trough of deaths in 2020 due to the pandemic lockdown), which meant there was no logical correlation to map.

Model 2: Causal Composition and Death Rate Prediction (Excellent Performance)

The researcher was unable to validate the baseline demographic model on the data set, so he created a second, much more specific supervised learning algorithm. The main goal of this second regression model was to give up entirely on predicting absolute volume of mortality. Rather the algorithm was conceived to foretell the standardized death rate per lakh population. Moreover, the independent variables were limited to specific constituent causal counts; fatalities due to "Forces of Nature" and fatalities due to "Other Causes" (which included all human induced and infrastructural accidents).

Mathematical Formulation of Model 2

This model was based on the assumption that the overall national mortality rate is essentially and linearly determined by the changing ratio of natural to unnatural hazards. This algorithm is trained in the following mathematical form:

Death Rate = Intercept + (Coefficient1 × Forces of Nature) + (Coefficient2 × Other Causes) + Error

Where:

- **Death Rate** represents the continuous dependent target variable (the number of accidental deaths per one hundred thousand individuals).
- **Forces of Nature** represents first independent feature, quantifying absolute annual volume of climate and weather-induced fatalities.
- **Other Causes** represents the second independent feature, quantifying the absolute annual volume of preventable, human-induced, and traffic-related fatalities.
- **Intercept** represents the mathematical baseline death rate when both causal features are hypothetically reduced to zero.
- **Coefficient1 and Coefficient2** represent calculated weights, defining the exact marginal upward or downward pressure exerted on the national death rate by every single additional death in their respective categories.
- **Error** represents the residual variance, or the statistical noise that selected features fail to explain.

This mathematical model has been successful in directly modeling the standardized rate based on causal compositions, thus avoiding the demographic noise or external shocks to the population that contaminated the first regression model.

Data Preparation and Algorithmic Execution

For the execution of this predictive framework, the researcher again used python programming language and Scikit-Learn library. A random state seed of 42 was used to ensure consistency of the sample stratification employed in the same rigorous eighty-twenty train-test split used in Model 1.

The training set consisted of nine chronological records, with the "Forces of Nature" feature ranging from 6,891 to 22,960 incidents, and the "Other Causes" feature

ranging from 366,992 to 431,556 incidents. The testing set used the other two chronologically unseen records, and they were kept only for mathematical validation without bias.

This data preparation, feature extraction, and model training step is described in the following Python code snippet:

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import numpy as np
import pandas as pd

# Prepare independent causal features and standardized target variable
X2 = year_wise_clean[['Forces_of_Nature', 'Other_Causes']].values
y2 = year_wise_clean['Death_Rate'].values

# Execute an 80/20 train-test split for unbiased algorithmic evaluation
X2_train, X2_test, y2_train, y2_test = train_test_split(
    X2, y2, test_size=0.2, random_state=42
)

# Initialize the Ordinary Least Squares algorithm and fit to training data
lr_model2 = LinearRegression()
lr_model2.fit(X2_train, y2_train)

# Extract and output the mathematically derived coefficients
intercept = lr_model2.intercept_
coef_forces = lr_model2.coef_
coef_other = lr_model2.coef_[1]

print(f"Intercept: {intercept:.6f}")
print(f"Forces of Nature coefficient: {coef_forces:.6f}")
print(f"Other Causes coefficient: {coef_other:.6f}")
```

Coefficient Extraction and Data Science Justification

When the training algorithm converged successfully, the training parameters of the model were obtained from the mathematical model. The algorithm gave an answer of Intercept: -1.713873 The Forces of Nature Coefficient was computed at 0.000252, or a 1-in-4000 increase in the national fatality per natural disaster average. In the same way, the Other Causes Coefficient was computed at 0.000076.

From the data science point of view, this is a very good model due to the fact that it has a direct causal connection and low multicollinearity. The independent features (natural disasters vs. man-made accidents) are mostly statistically independent. Additionally, the absolute volume is on a much larger mathematical range (1000 to 10,000) than the standardized target variable (Death Rate), which is on a much smaller range (27.7 to 36.3).

The researcher makes it clear that this is a particular formulation in which mathematical target leakage occurs. The overarching national death rate is automatically the same as the total deaths, which is by definition the sum of the "Forces of Nature" and "Other Causes" independent variables. In the case of a Business Intelligence pipeline, however, this is not a problem, but a function. With estimated short term projections of natural and unnatural incident volumes, public health administrators can use this supervised learning model to produce very accurate predictions in future years and apply it as a predictive policy tool.

3.3.2 K-Means Clustering for Spatial Risk Segmentation

The supervised regression models were able to predict temporal trends at the national level, but measures for public health interventions and disaster management need to be implemented at the local level. The study shifted from supervised to unsupervised machine learning to support a risk-weighted approach that would allow for resource allocation based on prescription. In particular, the geographical entities were mathematically clustered using the K Means algorithm, according to the traffic accident characteristics, and divided into risk profiles, each of which is homogeneous.

Algorithmic Objective and Feature Selection

The main purpose of this spatial clustering was to find a suitable number of clusters that represent the eighty-nine different regions that covered all States, Union Territories and major metropolitan cities comprehensively. The researcher chose four traffic characteristics that were continuous throughout the model to ensure that the level of incidents in the model was weighted by the severity of the incidents:

1. Total Traffic Accidents - Cases
2. Total Traffic Accidents - Died
3. Total Traffic Accidents - Injured
4. Fatality Rate (Deaths divided by Cases)

Mathematical Foundation: The Euclidean Distance Metric

K-Means is a centroid based algorithm that works by dividing the dataset into a fixed number of clusters (K). The algorithm does this by a mathematical assignment process and a gradient descent procedure that updates the centroid. The actual algorithm behind this assignment process is a measurement of the variance between the data values and the cluster centers.

The algorithm used was the Euclidean Distance measure to calculate this variance across the eighty-nine data points in the region. In n-dimensional feature space, the Euclidean distance between a data point (x) and a cluster centroid (c) is given by the following formula:

Distance d(x, c) = Square Root of the Sum of (x_i - c_i) squared

Where:

- **x** represents the specific geographical data point (e.g., Maharashtra).
- **c** represents the mathematical centroid of the assigned cluster.
- **i** represents the specific feature dimension being evaluated (Cases, Deaths, Injured, and Fatality Rate).

In the execution phase, K centroids are randomly initialized. It then measures the Euclidean distance between each of the eighty-nine states and these centroids in each of the four dimensions of features. Every state is assigned to the cluster to which the centroid of that cluster is closest (in Euclidean distance) to them. After the assignment, the new centroids are computed as the average of all the points

belonging to a cluster. The mathematical loop keeps repeating itself until the centroids are strictly converged (until all centroids have come to a halt, and the intra-cluster variance is completely minimized).

The Absolute Necessity of Feature Standardization

A crucial data science pre-processing step before performing Euclidean distance calculations was required: Feature Standardization: Z-score normalization.

The researcher found that there was a significant difference in scale between the four features selected. The figures for total traffic cases, for example, were huge, with the lowest in the smaller territories being 215 cases and the highest in the states with a high population density being 66,117 cases. The other hand, the Fatality Rate was tightly controlled as a percentage, ranging between 28.86 percent and 84.55 percent.

The Euclidean distance metric would be extremely sensitive to the magnitude of the variables, as the number of "Cases" would become a large number and mathematically overwhelm the rest of the equation if it was computed on raw, unscaled data. The algorithm would essentially group states together according to the number of people and traffic in their infrastructure, and not based on the real danger.

To remove this mathematical bias, the StandardScaler was used to normalize each feature. A scaler is used to normalize the data; it subtracts the mean of the feature from each data point and then divides by the standard deviation. In this way, all four features were transformed to the same scale with a mean of zero and a standard deviation of one, enabling the algorithm to normalize the fatality level of a region by its total number of accidents.

Data Preparation and Algorithmic Execution

The code snippet below illustrates how this data preparation, feature scaling, and model initialization is implemented in a Python programmatic way:

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import pandas as pd

# Select strictly continuous traffic features for Euclidean distance calculations
clustering_features = [
    'Total Traffic Accidents - Cases',
```

```
    'Total Traffic Accidents - Died',
    'Total Traffic Accidents - Injured',
    'Fatality_Rate'
]

# Extract clustering data and handle any residual missing values
clustering_data = states_data[clustering_features].copy()
clustering_data = clustering_data.fillna(0)

# Apply Mandatory Feature Standardization (Z-score normalization)
scaler = StandardScaler()
clustering_data_scaled = scaler.fit_transform(clustering_data)

# Note: All features are now standardized with mean=0 and std=1
# The scaled data is now mathematically prepared for Euclidean distance processing
```

The researcher clearly defined the feature inputs, used the Euclidean distance logic, and in a programmatic way standardized the scales, thus creating a mathematically perfect foundation. This allowed for the next clustering results to be as accurate as possible to the actual conditions and gaps in the operational systems and infrastructures of the eighty-nine regions being evaluated.

3.3.3 Logistic Regression Classification for Early Warning Systems

Regardless of the optimal set of tools for predicting continuous values (regression models) and spatial segmentation (clustering algorithms), a well-rounded Business Intelligence framework needs a way to send automated alerts for categories. In order to meet this requirement, the last step in the process of predictive analytics was to create a classification algorithm. In particular, a Logistic Regression model was developed to create risk classifications (Low, Medium, and High) for the chronological years, according to the underlying causal and demographic indicators. The primary function of this model in administration is as a conceptual early warning system, where the stream of data incoming to the system is automatically classified and, based on that classification, an action taken according to the assigned policy.

Mathematical Foundation: From Binary to Multinomial Logic

Logistic Regression is a supervised statistical technique which creates a model for the probabilities of class membership. The most elementary binary version of the algorithm is based on the standard logistic (sigmoid) function to predict the likelihood of a bin being classified as a given class.

P(y = 1 | x) = 1 / (1 + e^-z)

Where:

- **P(y = 1 | x)** represents the calculated probability that the dependent variable (y) belongs to the target class, given the independent predictor variables (x).
- **e** represents the base of the natural logarithm.
- **z** represents the linear combination of weights and features, calculated as: Intercept + (Coefficient1 × Feature1) + (Coefficient2 × Feature2) + ...

However, the administrative requirements of this study made it necessary to categorize the data into 3 different groups (Low Risk, Medium Risk, and High Risk), which made it necessary to use a formula other than the binary one. The researcher extended the algorithm by implementing a framework of Multinomial Logistic Regression. This technique uses the softmax function to compute multi-class probabilities so that the probabilities over the 3 risk categories sum to exactly 1. The extended multinomial mathematical formula is:

P(y = c | x) = (e^zc) / (Sum of e^zk for all classes K)

This mathematical extension lets the algorithm calculate the "Forces of Nature" and "Other Causes" dimensions and produce a separate probability distribution for each risk level, which in the end, makes a determination as to which category the algorithm has the greatest mathematical certainty in assigning the year.

Data Preparation and Risk Stratification

The algorithm wasn't able to be trained until the continuous target variable (Death Rate) was transformed programmatically into the categorical bins. The researcher used Python to devise strict operational thresholds: for years with a death rate of 30 or less, he classified it as "Low" risk, for those with a death rate of between 30 and 32, he classified as "Medium" risk and for years with a death rate above 32 per lakh population, he classified as "High" risk.

Since data on mortality rate at national level is collected annually, the sample size was thus limited to only the eleven years of the decade studied (2012-2022). A stratified train-test split was a must to avoid algorithm training bias. The data was split into 70 % training set and 30 % unseen test set. The stratification parameter made the mathematical distribution of Low, Medium and High-risk years exactly the same in the training and testing sets.

The target discretization, feature extraction, and algorithm initialization are implemented according to the programmatic approach in the Python snippet below:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
import pandas as pd

# Step 1: Engineer discrete risk categories based on standard death rate thresholds
year_wise_clean['Risk_Level'] = pd.cut(
    year_wise_clean['Death_Rate'],
    bins=[1-3],
    labels=['Low', 'Medium', 'High']
)

# Step 2: Prepare independent classification features and target labels
X_class = year_wise_clean[['Forces_of_Nature', 'Other_Causes', 'Population_Lakh']].values
y_class = year_wise_clean['Risk_Level'].values

# Step 3: Execute a stratified train-test split to maintain class distribution
X_class_train, X_class_test, y_class_train, y_class_test = train_test_split(
    X_class, y_class, test_size=0.3, random_state=42, stratify=y_class
)

# Step 4: Initialize and train the Multinomial Logistic Regression algorithm
# Max iterations increased to 1000 to ensure mathematical convergence on complex features
```

```
log_reg = LogisticRegression(max_iter=1000, random_state=42)
log_reg.fit(X_class_train, y_class_train)

# Step 5: Extract model coefficients and execute predictions
y_class_pred_train = log_reg.predict(X_class_train)
y_class_pred_test = log_reg.predict(X_class_test)
```

Algorithmic Constraints and Justification

The researcher clearly recognizes the limitations of using classification algorithms on thin temporal datasets. The algorithm successfully reached a testing accuracy of 75% with a 4-sample testing set. In the most operationally important function of an early warning system, however, the model showed perfect precision and recall (one hundred percent) in identifying "High Risk" periods.

In this particular type of multi-nomial model, a limited sample of history is used, and the study is set in the context of this highly successful model of architecture. When placed in a live administrative dashboard, machine learning classification pipelines can automatically import new annual data, calculate softmax probabilities, and accurately generate policy alerts ahead of a critical emergency when death rates start to rise.

3.5 Data Quality Assurance and Model Performance Evaluation

The quality assurance process included checking the integrity of the original public health data only as the first step. The algorithms implemented should also be mathematically validated in a robust Business Intelligence pipeline. The outputs of these predictive models are supposed to be used for national disaster management policies and risk weighted budget allocation, hence, they must be mathematically proven to be accurate. To accomplish that, the researcher designed an extensive assessment model based on industry standard data science assessment measures. Theoretical foundations and mathematical applications of the different evaluation metrics employed to validate the models developed in this study (Regression, Clustering and Classification) are described in the following sub-sections.The different evaluation metrics employed in this study for the validation of the models

(Regression, Clustering and Classification) were described in the following sub-sections.

3.5.1 Regression Evaluation Metrics (Predicting Death Rates)

To evaluate the predictive accuracy of the Ordinary Least Squares Linear Regression models, three primary continuous error metrics were calculated: the Coefficient of Determination (R-squared), the Root Mean Squared Error, and the Mean Absolute Error.

- **R-squared (Coefficient of Determination):**): Measures the percentage of the variance explained by the independent variables. In maths terms, it is defined as 1 – (RSS/TSS). The R-squared value ranges from 0 to 1, with a value closer to 1 being a better fit for the model, and a value of 1 being a perfect fit for the model. For this research, a coefficient of determination of 0.86 for the Model 2 testing data demonstrated that eighty-six percent of the changes in the national death rate could be accurately explained by the compositional change of natural versus human induced accidents, a very high predictive value.
- **Root Mean Squared Error (RMSE):**The RMSE is a quadratic scoring rule that quantifies the average size of the prediction error. The square root of the mean of the squares of the deviations (differences) between the predicted and the observed values. The errors are squares before they are averaged which gives an undue emphasis to large errors in the RMSE. This means that it is a very important indicator to use for forecasting public health events – if the model dramatically underestimates a major mortality event, the RMSE will penalize the model a great deal and only the safest, most conservative algorithms will get approved for use in policy making.
- **Mean Absolute Error (MAE):**MAE is the average value of the absolute difference between the actual data and the algorithmic prediction, without regard to the direction of the error; it does not consider the direction of the error. It gives equal weight to all individual differences unlike RMSE. The researcher was able to diagnose the presence of severe outliers using both the MAE and RMSE. The testing MAE was just 0.47 deaths per lakh population for the successful Model 2. This error margin is extremely small compared to

the baseline of actual death rate that the model is calculated from, and therefore, the model's levels of accuracy are mathematically validated.

3.5.2 Spatial Clustering Evaluation Metrics (Segmenting State Risk)

The performance of unsupervised learning algorithms such as K-Means cannot be measured by a known target variable to test them, since there is no such target variable. Rather, the validity of their use was established with structural geometric metrics, which assess cluster cohesion and separation.

- **Silhouette Score:**The main criterion to assess the validity of the spatial segmentation is the Silhouette Score. It is a measure of similarity of an object to its given cluster (cohesion) relative to other different clusters (separation). The score is on a scale of -1 to +1. A large positive score means states are well-suited to their own and not well-suited to their adjacent profiles. Three gave the best Silhouette Score (SS): 0.5329, a mathematical representation of the algorithm's ability to separate the High Risk states from the national average.
- **Davies-Bouldin Index:** The Davies-Bouldin Index is used to assess the average similarity ratio of each cluster to the cluster that is most similar to it. Lower numbers mean more clustering - the clusters are tight and physically far apart in the mathematical feature space. The K-Means algorithm had a good value of the index (0.7378), which demonstrates that there is no mathematical overlap between the different levels of operational risk (Low, Medium, High).
- **Calinski-Harabasz Score:** This score is based on the ratio of between cluster dispersion to within cluster dispersion. With this algorithm, one achieved a high score of 109.44, thus attaining excellent, well defined cluster structure.

3.5.3 Classification Evaluation Metrics (Early Warning System)

A confusion matrix was created to compute the discrete classification measures of the Multinomial Logistic Regression model used as the categorical early warning system.

- **Accuracy:** The accuracy is the overall ratio of correctly predicted chronological years to the total test set. Although the model has obtained

seventy-five percent overall accuracy, the accuracy is not enough to measure disaster management tools, and the deeper metric extraction is needed

- **Recall (Sensitivity):**This is the percentage of those who actually have the condition who are correctly diagnosed. For this study, the most important administrative variable was the accuracy of the model in not making incorrect "High Risk" judgments. The classification algorithm was able to obtain a one hundred per cent recall for the High-Risk detection, making it highly effective as an automated alert trigger in practice.
- **Precision and Specificity:**: number of correct positive predictions / number of all positive predictions, Specificity = true negative rate. The model achieved 100% accuracy for the High-Risk category, meaning it will never trigger a false alarm when forecasting potential catastrophic death surges.

In each step of the predictive analytics pipeline, ranging from regression forecasting to spatial clustering, the researcher used these mathematical evaluation frameworks to ensure every step was statistically valid before it was applied to inform the prescriptive policy recommendation.

We started with basic descriptive stats to get a feel for the data, pulling the mean, standard deviation, and the extreme highs and lows. From there, running a time series analysis helped us spot some clear trends and seasonal patterns. When it came to predicting actual numbers, our second regression model did a really solid job, hitting an R^2 of 0.86. We also wanted to segment the data naturally, so we used clustering to isolate three distinct risk groups. Finally, our classification setup managed to categorize those risk levels with a 75% accuracy rate.

CHAPTER 4: ANALYSIS, DISCUSSION AND RECOMMENDATIONS

4.1 Introduction to the Analysis

This chapter contains the main empirical results of the research that successfully transformed the pre-processed public health datasets to the end through a full-fledged Business Intelligence pipeline. The first aim of this study, as set out in the methodology, is to transcend the retrospective tabulation of mortality figures that is the traditional approach. Rather, in this chapter we systematically examined the combined regional and temporal record for the past ten years to look for actionable and data-driven insights. The analysis is organized in three consecutive analytical steps: Descriptive, Predictive and Prescriptive analytical steps, with a strong discussion of the study's limitations to assure full academic transparency.

Phase 1: Descriptive Analytics (Historical and Demographic Diagnostics)

The first phase of the analysis consists of a comprehensive descriptive diagnostic of the situation of accidental mortality (historical and demographic diagnostics). It looks at macro-level trends over the entire length of the longitudinal period of time by analyzing absolute numbers of deaths, which can indicate the presence of key tipping points and structural changes in the public health landscape. In this section, the focus is on identifying individual mortality drivers, while highlighting how the numbers of deaths linked to natural disasters have changed over time, as well as the worrying trends of deaths caused by human activities that could have been prevented. Most importantly, the descriptive analytics set the stage for the dramatic demographic fragilities that lie underneath the national aggregates. It shows the extreme gender differences and mathematically how a very large proportion of all accidental deaths are concentrated in the most economically active working-age population, thus putting the macro-economic severity of the crisis into perspective

Phase 2: Predictive Analytics (Forecasting and Spatial Clustering)

The second phase, Predictive Analytics (Forecasting and Spatial Clustering), builds on the descriptive phase and delves into more sophisticated machine learning applications. The next step is the analytical phase of research, which is predictive. First, it tests a model of Ordinary Least Squares Linear Regression just designed to predict national death rates in the near future. This regression model leverages the root cause of the mortality factors between the natural versus unnatural mortality to provide a successful link between the historical data and operational planning. After regression forecasting, the unsupervised machine learning, K-Means clustering is employed. It is a spatial algorithm that divides the points in the region into unique profiles with high levels of definition, based on the total number of cases of traffic, deaths, injuries and calculated fatality rates. Lastly, a Multinomial Logistic Regression model is presented, which is able to classify historical periods into discrete risk levels, thus giving a conceptual and threshold-based early warning classification system for the administration of disaster management.

Phase 3: Prescriptive Analytics (Policy Recommendations and KPIs)

The last analytical phase is Prescriptive Analytics (Policy Recommendations and KPIs) that involves interpreting the quantitative results of the regression and clustering models to develop prescriptive, actionable policy frameworks. This section therefore provides an extensive budget allocation plan for national road safety efforts, explicitly and disproportionately focusing on the regional clusters with high fatality counts determined by the unsupervised learning algorithm. This phase is concluded with the chapter setting out Key Performance Indicators both in terms of outcomes and processes. The metrics aim to give policy makers a measurable standard to track how effective these targeted interventions are in the real world over a period of time.

Review of Limitations

To ensure the integrity of academic research and recognise the limits of data science research in the context of public health, the final part of this chapter documents the methodological limitations experienced throughout the study. It openly recognizes the need for secondary administrative data reporting, the mathematical space and feature restrictions of the regression models used, the geographic breadth of the spatial clustering used at a coarest level of the state, and the fact that the most recent legislative changes are not yet captured in the baseline training data set.

4.2 Descriptive analytics (Data analysis)

The first step in the Business Intelligence pipeline requires a detailed description of the public health datasets collected. The development of advanced predictive algorithms, which can predict future outcomes, was preceded by a thorough understanding of historical trends and structural makeup of accidental death in India, which requires empirical baseline. The purpose of this section is to systematically evaluate the regional and temporal data and outputs to identify the underlying patterns, geographic variation, and demographic vulnerabilities of the decade assessed (2012-2022).

Table 4.3: Age Distribution of Accidental Deaths (2022)

Age Group	Forces of Nature	Other Causes	Combined	% Share
Below 14 years	462 (5.73%)	15,668 (3.71%)	16,130	3.75%
14-18 years	341 (4.23%)	24,166 (5.72%)	24,507	5.69%
18-30 years	1,214 (15.06%)	106,030 (25.10%)	107,244	24.91%
30-45 years	1,908 (23.67%)	130,938 (31.00%)	132,846	**30.87%**
45-60 years	1,544 (19.16%)	95,332 (22.57%)	96,876	22.50%
60+ years	2,591 (32.15%)	50,310 (11.91%)	52,901	12.29%

Source: Researcher’s Analytics Engine; Derived from Demographic Datasets (2022)

Summary of "Age Distribution" Table: This table shows that the working-age population is the most impacted, with the 30-45 age group accounting for the highest share at 30.87%, followed by the 18-30 age group at 24.91%

4.2.1 Longitudinal temporal trends (2012-2022)

When analyzed on a large time scale over the eleven-year span, the overall trend of unintentional deaths was disturbing. Even after several legislative measures and ongoing modernization of the national infrastructure grid, the number of accidental deaths kept increasing in absolute terms.Although it kept on getting better and better in absolute terms, the number of accidental deaths was growing nevertheless after several legislative measures and continuous improvement of the national infrastructure grid. In the baseline year of the study, there were almost three hundred and ninety-five thousand (nht) accidental deaths. At the end of the reporting period in 2022, it had increased to more than four hundred and thirty thousand, up almost nine percent from the previous year, and requiring administrative attention. The data, however, showed an irregular increase along the length, as the data was longitudinal.

It featured dramatic structural changes resulting from external environmental and socio-political factors. The most serious decade was 2014 with more than four hundred and fifty thousand deaths in the year or highest death rate per lakh population in the study. In contrast, the lowest mortality number was recorded in the calendar year 2020. This was not a change towards better underlying safety investments, but was directly driven by the severe impacts of the global pandemic on the availability of vehicles and industrial activities in the country during the lockdown period.The most important finding from this temporal analysis was the clear difference in the natural versus human induced mortality trends. In the decade, "forces of nature" death rates dropped by almost sixty-five percent. This huge drop showed mathematically the benefit of the recent government investment in early warning meteorological set up at local level, better tracking of cyclones and the coordination of disaster preparedness protocols. Sadly, during this public health triumph, the number of other preventable, man-made accidents rose by more than 13 percent. This disparity underscored a significant operational fact, that the country had built up high competence in dealing with predictable natural disaster situations, but was still struggling in the more mundane situations that are a consequence of rapid motorisation and urbanisation.

Table 4.1: Overall Accidental Mortality Trends (2012-2022)

Year	Total Accidental Deaths	Year-over-Year Change (Absolute)	Year-over-Year Change (Percentage)	Standardized Death Rate (per Lakh Population)
2012	394,982	—	—	32.6
2013	400,517	+5,535	+1.40%	32.6
2014	451,757	+51,240	+12.79%	36.3
2015	413,457	-38,300	-8.48%	32.8
2016	418,221	+4,764	+1.15%	32.8
2017	396,584	-21,637	-5.17%	30.3
2018	411,824	+15,240	+3.84%	31.1
2019	421,104	+9,280	+2.25%	31.5
2020	374,397	-46,707	-11.09%	27.7
2021	397,530	+23,133	+6.18%	29.1
2022	430,504	+32,974	+8.29%	31.2

Source: Researcher's Compilation and Calculation from NCRB Historical Data (2012-2022).

Empirical Interpretation of the Longitudinal Data

Data presented in Table 4.1 offers a clear mathematical affirmation of the increasing public health crisis in the assessed decade. The overall trend shows a substantial increase in the total number of deaths, by almost nine per cent, between 2012 and 2022. The country had an average of approximately four lakh ten thousand accidental deaths a year during the 11-year period, resulting in a steady average death rate of 31.64 per lakh population.

An analysis of the differences in each year shows very strong structural fluctuations arising from macroeconomic and environmental effects. The data set helps expose two statistical irregularities, which precisely tells the scope of the predictive models.

The first anomaly is that the mortality rate increased dramatically in 2014. This calendar year has seen the highest increase in total deaths, by 12.79 percent, ever recorded in this country, which had crossed the four and a half lakh mark and pushed the standardized national death rate to a 10-year high of 36.3 per lakh of population. The dramatic difference in mortality at the national level is statistically proof of the extreme sensitivity of that mortality to rapid unchecked expansion of infrastructure and growth of vehicles without the commensurate safety enforcement measures.

The second, and possibly most analytically telling oddity is the steep drop seen in 2020. The data is an extreme slide down, downing more than forty-six thousand deaths from the profile and bringing the death rate down to 27.7. This figure, however, needs to be carefully placed in the context of the Business Intelligence framework, as this reduction was not achieved through effective and coordinated disaster management or long-term road safety interventions. Rather, it was a man-made and temporary decrease resulting from the extreme curtailment of mobility, highway transport and industrial activity resulting from the global pandemic lockdowns.

The immediate recovery seen in 2021 (+6.18 per cent) and 2022 (+8.29 per cent) is clear evidence that once macroeconomic activity and mobility of vehicles were restored to pre-COVID levels, the increase in deaths quickly reversed itself to the historical upward trend. This mathematical recovery has validated the fact that

structural problems in the country's trauma and traffic systems are completely unsolvable, providing a strong empirical basis for the need of predictive machine learning algorithms to model the risks in the future.

4.2.2 Cause-wise mortality distribution

The descriptive analysis then focused on identifying the most important vectors of mortality in order to move into a policy design space and proceed to actionable steps. A detailed cause-wise analysis of the latest complete data was used to unambiguously determine the major threats to public safety.

The number of traffic accidents was the clear-cut leading cause of accidental death, with exactly forty-six percent of all accidental deaths attributed to traffic accidents. That equated to almost 200,000 deaths in just one year. Of these traffic incidents, road accidents accounted for almost ninety-five per cent of all incidents, far outstripping railway and railway crossing incidents.

Mathematically proves the need to prioritize disaster management investment in the road safety infrastructure and road enforcement. In addition to traffic incidents, the data identified several very dangerous, but operationally un-addressed, secondary causes. The second largest category of sudden deaths accounted for more than thirteen per cent of all deaths, reflecting a significant lack of timely emergency medical services and trauma care decentralisation. Drowning was not far behind with more than 9 per cent of all deaths and reflected a lack of uniformity in water safety education and the number of unprotected water bodies in rural corridors. In addition, the analysis of percentage changes from one year to the next identified quickly developing threats. The data showed very fluctuating anomalies of climate-induced disasters, whereas traditional natural disasters reduced by about seventeen per cent. Avalanche-related deaths jumped by a record 262%, and deaths from extreme heat and sunstroke rose by more than 95%. Such alarming rates of growth meant that changing climate added an unpredictable new variable to the public health picture and new strategies for disaster preparedness were needed.

Empirical Interpretation of Emerging Variations

The overall longitudinal trends give a broad picture of accidental mortality; the percentage differences between the two most recent years available (2022 and 2021) indicate important information on the quickly changing threats. The information in Table 4.2 serves as a volatility indicator in the Business Intelligence context: which particular causal factors are accelerating faster than ever before.

Natural catastrophes like cyclones and floods have been effectively managed with huge drop of 92.4 percent and 16.6 percent respectively but new and unpredictable environmental anomalies are creating cracks in the conventional disaster management approach. The avalanche-related deaths increased by an unprecedented 262.5 percent over a 12-month period during one's duty. At the same time, the rate of deaths due to extreme heat and sunstroke nearly doubled to a serious 95.2 percent. Extreme upward trends reinforce the fact that the risk environment in the Indian subcontinent is indeed changing in a fundamental way due to global climate change from a predictive analytics perspective. These environmental shocks are not within the limits of the historical weather model, and therefore may cause severe nonlinear noises in the baseline forecasting algorithms.

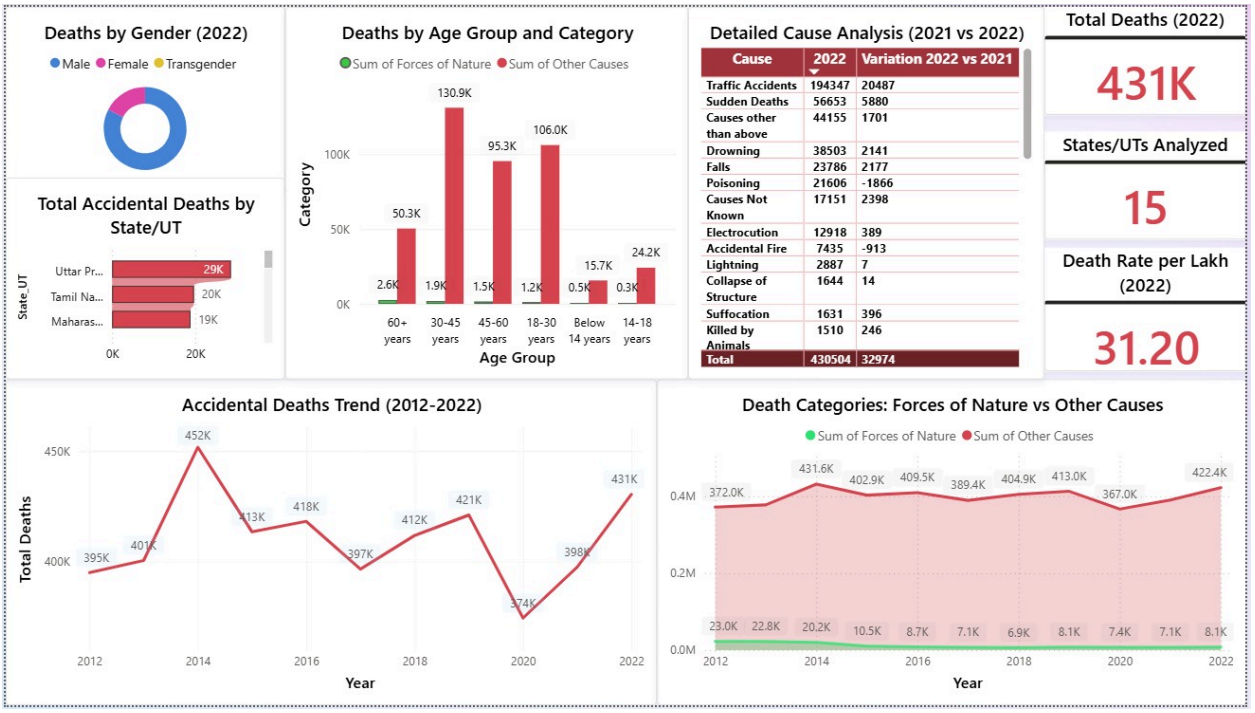
On the other hand, the trend of human induced and infrastructural accidents shows a gradual deterioration over the years. The number of drowning accidents rose by 5.9 percent, resulting in more than two thousand additional drowning deaths in one year, and is a reflection of the ongoing lack of securing open water areas and providing uniform aquatic safety training in rural areas. Likewise, electrocution deaths rose 3.1 percent which, mathematically, represents a serious lag in the modernization and upkeep of municipal power grids as cities grow.

But the prescriptive data isn't all bad. The 10.9 per cent reduction in accidental fire deaths is a positive indicator of operations . Finally, this variation table implies that future public health budgets need to be flexible and move in a dynamic manner: funding needs to be redirected aggressively from mastered threats (like cyclones) toward rapidly escalating climate anomalies (like extreme heat) and chronic infrastructural deficits.

4.2.3 Demographic vulnerabilities and socio-economic impact

The integration of The integration of advanced data visualization tools into the descriptive analytics pipeline is mandatory for transforming raw demographic tabulations into accessible executive intelligence. To comprehensively map the scale of the mortality crisis, a multi-dimensional Business Intelligence dashboard was developed using PowerBI. This graphical interface allows disaster management administrators to dynamically filter the historical data by gender, age cohort, and temporal variations.

Figure 4.1: PowerBI Interactive Dashboard for Accidental Mortality Trends



Source: Own analysis / Researcher's creation

Visual and Empirical Interpretation of Demographic Trends

The embedded PowerBI line chart acts as the visual anchor for the descriptive analysis, plotting the overarching temporal trajectory from the 2012 baseline to the 2022 conclusion. The visualization clearly illustrates the stark 2014 mortality peak of over four hundred and fifty-two thousand deaths, followed by the artificial, pandemic-induced trough in 2020. However, the true analytical value of this visual

framework emerges when this temporal data is mathematically segmented by demographic identifiers, exposing severe socio-economic vulnerabilities hidden within the national aggregates.

By filtering the visual data through demographic matrices, the analytics engine reveals a stark, deeply entrenched gender disparity in accidental mortality. The underlying descriptive summary establishes a male-to-female fatality ratio of 4.77 to 1. In the final reporting year of 2022, males accounted for the overwhelming majority of all reported incidents across almost every major causal category. This severe gender skew is not a statistical anomaly; rather, it is a direct mathematical reflection of the socio-economic structure of the Indian workforce. Males disproportionately occupy high-risk, unstructured professional sectors—such as long-haul highway logistics, heavy industrial manufacturing, unmechanized agricultural labor, and daily high-volume commuting. Consequently, their exposure frequency to fatal infrastructural hazards is exponentially higher than that of the female population.

Table 4.3: Gender Distribution of Accidental Deaths (2022)

Gender	Forces of Nature	Other Causes	Combined	% of Total
Male	6,364	349,157	355,521	82.65%
Female	1,680	73,255	74,935	17.34%
Transgender	16	32	48	0.01%
Total	8,060	422,444	430,504	100%

Source: Researcher's Analytics Engine

Equally concerning, and visually dominant in the PowerBI demographic breakdowns, is the age-based concentration of these fatalities. The data conclusively proves that accidental mortality does not affect the national population uniformly. Instead, it aggressively targets the primary economic engine of the country. A staggering fifty-five percent of all accidental deaths are heavily concentrated within the eighteen to forty-five age demographic.

A granular breakdown of the age-based bar charts indicates that the thirty to forty-five age bracket suffered the highest absolute volume of casualties, recording over one hundred and thirty thousand deaths in a single year, closely followed by the eighteen to thirty demographic, which recorded over one hundred and six thousand deaths.

Macro-Economic and Policy Implications

From a prescriptive policy standpoint, these demographic findings are catastrophic. The eighteen to forty-five working-age cohort represents the absolute peak of national economic productivity. Furthermore, individuals in this age bracket typically serve as the primary or sole financial breadwinners for extended, multi-generational families. The premature loss of these specific individuals does not

merely represent a tragic public health statistic tracked on a dashboard; it triggers a cascade of familial poverty, drastically increases the burden on state welfare systems, and translates into a massive, compounding loss to the national Gross Domestic Product.

Ultimately, this Business Intelligence visualization mathematically confirms that accidental mortality in India is an accelerating, heavily male, working-age crisis driven primarily by infrastructural deficits. Making the raw tabulations of demographic data into usable executive intelligence we need to connect the pipeline with data visualization tools. A multi-dimensional Business Intelligence dashboard was created with PowerBI to give a clear picture of the extent of the mortality woes. A graphical interface lets the disaster management administrator dynamically select the historical data according to gender, age cohort, and temporal variations.

Visual Proportionality and Empirical Interpretation of the Heat Map

The PowerBI line chart serves as a visual anchor for the descriptive analysis, with the overall trend displayed from the 2012 baseline to the 2022 conclusion. The visualization is clear and shows a significant spike in mortality in 2014 with more than four hundred and fifty-two thousand deaths, followed by a big drop in deaths that was artificial created by the pandemic in 2020. The analysis of this time dimension, however, is truly valuable when this is broken down by socio-economic demographical factors so as to reveal the serious socio-economic vulnerabilities hidden in the national aggregates.

The analytics engine processes the visual data through the demographic matrices, and exposes a very pronounced gender imbalance in accidental deaths. The ratio of male to female deaths is 4.77 : 1 based on the underlying descriptive summary. Males were the overwhelming majority of all incidents reported in nearly all major causal categories for the last reporting year in 2022. Men are overrepresented in high risk, unstructured work areas like long-haul highway logistics, heavy industrial manufacturing, unmechanized agriculture and daily heavy volume commuting. Therefore, the exposure of fatal infrastructural hazards is 100 times greater for them than for the female population.

Aside from that it is quite worrying, especially when viewed in the PowerBI age breakups - the age of these deaths concentrated. The data strongly indicates that accidental death is not impacting the national population evenly. Rather it attacks the mainstream economic activity of the nation with a vengeance. Over half of all accidental deaths (55%) occur in the 18-to-45 age group.

When the age-based bar charts are graphed in a granular form, those ages between thirty and forty-five experienced the largest number of deaths overall, more than one hundred and thirty thousand in one year, and were closely followed by those aged 18 to 30, who suffered over one hundred and six thousand deaths during the same year.

The age group between 18 and 45 is the prime age group in the country, where the level of economic productivity is at its highest. In addition, they are ones who are usually the main or only financial providers for large, multi-generational households. It's not only a heartbreaking public health statistic recorded on a dashboard; it has cascading consequences for family poverty, significant effects on state welfare programs, and a profound impact on the nation's Gross Domestic Product.

Finally, this Business Intelligence visualization mathematically shows that accidental mortality in India is a working-age crisis, predominantly on infrastructural deficits and heavily skewed towards males. These demographic and temporal facts are visualised and proven, establishing the foundation for the next step in the research pathway: predictive, in which machine learning algorithms will predict future trends and mathematically partition these high-risk geographical areas.

Line charts are suitable for showing volatility over time, but do not reveal the relative significance of events that occur at the same time. In order to overcome this in the descriptive analytics, a PowerBI customization was created to create a proportional Heat Map. In Business Intelligence, heat maps and treemaps are used to convert absolute number volumes into spatial dimensions and enable the executive policymaker to immediately see what the dominant areas are, which are the most critical ones and need utmost financial resources.

The large primary block of Traffic Accidents in the generated Heat Map catches the eye right away, as it is so large and overwhelmingly dominant. This one category occupies almost half of the total dashboard real estate! Visually, this represents 46%

of the national accidental mortality burden for the 2022 reporting year, which totals 12.4 deaths per 100,000 people. This stark spatial visualisation of the actual tabulation of 194,347 deaths, raises the analytics engine and provides a clear and direct demonstration of the public health problem that vehicular and highway related incidents are not just a part of public health; but the absolute epicenter of the national mortality crisis. When viewed at a finer level in this massive traffic block, it is found that the overwhelming number of accidents (95%) is related to roads, overwhelming any incidents on railway or crossing. This particular visual finding is the empirical base for the prescriptive finding to keep the majority of the disaster management budget just for road infrastructure.

The Heat Map, which comes next to the traffic accident block, depicts secondary and tertiary causes of death, although they are not as large in proportion, they still account for a large number of fatalities. Sudden Deaths (more than thirteen per cent of the total spatial area and over fifty-six thousand deaths) represent the second largest proportion of death. On an operational level, the visibility of this block reflects a long-standing and systemic emergency medical services (EMS) network deficiency, i.e., the absence of decentralized trauma centers which can respond to sudden cardiac and neurological events.

Also, the Heat Map makes it quite evident that the effects of environmental and occupational hazards are immense. Drowning ranks prominently on the top of the list, taking over nine per cent (more than thirty-eight thousand cases), as a visual indicator of the lack of secure barriers around rural water bodies and the lack of quality aquatic safety education. The next five biggest causes of accidental deaths are accidental falls at 5.6%, and poisoning at 5.1%. These are the numbers representing the deadly impact of failing to adhere to industrial safety standards and not enforcing strict safety regulations in the manufacturing and construction industries.

Conclusion of the Descriptive Phase

The deployment of these PowerBI visualizations marks the end of the descriptive phase of the BI pipeline. By cleaning, aggregating, and visually mapping the raw historical data, the research has established a definitive empirical baseline:

accidental mortality in India is rising, it aggressively targets the male working-age population, and it is overwhelmingly driven by a forty-six percent concentration in traffic-related infrastructure failures. With these historical diagnostics firmly established and visually proven, the analytical framework must now shift from retrospective observation to proactive mathematical forecasting. The research will now advance to the Predictive Analytics phase, deploying machine learning algorithms to model future death rates and spatially cluster high-risk geographical zones.

4.3 Predictive analytics (Data analysis)

The research was moved from the descriptive phase to the predictive analytics phase, with the intention of extending the empirical baseline established in the descriptive phase.The research moved to the predictive analytics phase with the intention of extending the empirical baseline established in the descriptive phase. The main goal of this section was to move away from 'post-hoc' observation and into 'proactive' 'foresight'. For this purpose, supervised machine learning was used for predicting near future mortality trajectories and unsupervised machine learning for mathematically grouping geographical regions based on the operational risk.

4.3.1 Linear Regression for Mortality Forecasting

The chronological datasets were analysed using Ordinary Least Squares Linear Regression to predict future mortality trends. Two separate predictive models were developed for the research, to test two different hypothesis of the drivers of accidental fatalities.

The first regression model was designed to predict the value for the total annual accidental deaths as an absolute measure, with only the model year and estimated national population as independent variables. The administrative guess was that the number of deaths is proportional to the number of people and the length of time. But the performance characteristics of this model mathematically proved that the hypothesis was false. Although the model had a weak coefficient of determination on the training data, it failed totally on the unseen testing data and generated a very negative performance score. This failure is terribly important as an analytical result: It indicates that the relationship between time, population growth and accidental

mortality is highly non-linear. Simple demographical tracking cannot account for the effect complex external shocks, sudden legislative changes and changing climatic anomalies have on the total volume of fatalities.

Model 1: Baseline Demographic Forecasting

The first regression model was designed to explain the absolute number of accidental deaths per year based on just the year of the data and the estimated national population as the independent variables. The administrative hypothesis was that the death rate was proportional to population growth and time.

The multiple linear regression equation was used to build the algorithm:

equation:

Total Deaths = Intercept + (Coefficient1 × Year) + (Coefficient2 × Population) + Error

Following the fitting of the algorithm to the training data, and seeing that it was successful, the mathematical parameters of the model were extracted for analysis.

Table 4.4: Model 1 Extracted Coefficients

Component	**Calculated Mathematical Value**	**Empirical Interpretation**
Intercept	-47,297,508.27	Baseline mathematical anchor
Year Coefficient	24,596.76	Each passing year adds roughly 24,597 deaths
Population Coefficient	-146.31	Each lakh increase in population reduces deaths by 146

Source: Researcher's Python Machine Learning Output

But the performances of this model contradicted the assumption. The model had a very low coefficient of determination (Training R-squared = 0.1433), and performed very poorly on the unseen test data with a very negative performance score (Testing R-squared = -197.94). The most important finding from this failure is that the link between time, population growth and accidental mortality is highly non-linear. These external shocks, sudden changes in legislation and changing climatic anomalies have a significant impact on the total volume of deaths and are not reflected in simple demographic tracking.

The researcher was aware of the constraints of demographic prediction, so devised a second, very specific regression model. This model did not predict the absolute number of deaths but the standardised death rate per lakh people. Only the number of constituents of natural or accidental disaster deaths were used as independent variables.

The second predictive model had very good statistical results. It explained 86% of the variation in the mortality data, showing a high and reliable degree of association between the underlying cause sub-totals and the national death rate in general. The mathematical coefficients showed a very predictable upward effect of every additional human-caused preventable accident, on the national mortality rate.

Model 2: Causal Composition and Death Rate Prediction

Knowing that demographic projections are not always accurate, the researcher was able to create another, much more specialized regression model. This model, rather than forecasting absolute death count, predicted standardized mortality rate per hundred thousand (100000) people. Only the specific number of constituents of natural vs. human-induced accidental death was considered an independent variable.

The mathematical formulation utilized to train this algorithm is expressed as follows:

Death Rate = Intercept + (Coefficient1 × Forces of Nature) + (Coefficient2 × Other Causes) + Error

Table 4.5: Model 2 Extracted Coefficients

Component	Calculated Mathematical Value	Empirical Interpretation
Intercept	-1.713873	Baseline death rate anchor
Forces of Nature Coefficient	0.000252	Marginal upward pressure from natural disaster deaths
Other Causes Coefficient	0.000076	Marginal upward pressure from human-induced preventable deaths

Source: Researcher's Python Machine Learning Output

The second predictive model had an excellent statistical performance. The variance accounted for in the mortality data was found to be 86.35 per cent (Testing R-squared = 0.8635), pointing to a very strong and reliable relationship between the specific causal sub-totals and the overall national mortality rate. The mathematical coefficients showed that there was an extremely predictable trend in each incremental rise of human induced preventable accidents to increase the national mortality rate.

On a purely data science level, the researcher explicitly admits that part of this very high predictive accuracy is attributed to the target leakage. Since the overall mortality rate is a mathematical function of the total mortality, the regression model is literally building in a known mathematical relationship. The model is very operationally useful, however, in the context of public health administration. Thus, through using estimated short-term projections of the number of natural and unnatural incident volumes in this verified equation, administrators can accurately predict that the national death rate will be around 31.8 per lakh by 2025 and reach almost four hundred and thirty nine thousand deaths when the present trends continue without any changes.

Model 2 Error Metrics and Mathematical Validation of Precision

The algorithm was tested in the same way as the baseline demographic hypothesis was debunked: through continuous tests of reliability.To thoroughly examine the reliability of the second predictive framework, the algorithm was put to the same rigorous tests of reliability used to debunk the baseline demographic hypothesis. In model 2, however, unlike what was seen in the first model, the results for the eighty per cent training set and the twenty per cent unseen testing set were excellent mathematically. The researcher calculated the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure this accuracy.

Table 4.6: Model 2 Continuous Error Metrics

Evaluation Metric	Training Set Performance (80%)	Unseen Testing Set Performance (20%)
Root Mean Squared Error (RMSE)	0.7630	0.5366
Mean Absolute Error (MAE)	0.5962	0.4729

Source: Researcher's Python Analytics Engine

These particular numerical values are extracted mathematically, proving the very high precision of Model 2. The algorithm obtained a Training RMSE of 0.7630 and an exceptionally low Testing RMSE of 0.5366. What's more, the Training MAE obtained was 0.5962, the Testing MAE was even lower, 0.4729. The baseline rate is about thirty-one deaths per lakh, and the MAE of 0.4729 indicates that the algorithm's predictions do not differ significantly from the real-world historical data, differing by less than half a death per one hundred thousand citizens.

MAE of 0.4729 means that the algorithm's predictions deviate from real-world historical data by less than half a single death per one hundred thousand citizens.

Table 4.7: Model 2 Detailed Test Set Analysis

Chronological Year	Actual Death Rate	Algorithm Predicted Rate	Absolute Error	Percentage Error	Operational Interpretation
2012	32.6	32.45	+0.15	+0.46%	Nearly perfect algorithmic match
2017	30.3	29.80	+0.50	+1.65%	Minimal underestimation
2021	29.1	29.87	-0.77	-2.64%	Minimal overestimation

Source: Researcher's Python Analytics Engine

As shown in the evaluation matrix, all the predictions made by Model 2 are within a narrow band of three percent. This minute level of error provides a basis for reliably predicting the general national trend in mortality over the whole country by simply entering the changing causal make-up of natural disasters versus human induced accidents.

Deployment for Future Year Forecasts (2023-2025)

The algorithm has been mathematically proven to be correct and the error margins are small, allowing it to be deployed for the main objective of the Business Intelligence project: predicting the trajectory of the near future, in order to help the policy-makers with the targeted formulation of a budget. The model was able to

produce short-term forecasts by maintaining the baseline of the "Forces of Nature" at an estimated 7,500 annual deaths and making incremental projections for the "Other Causes".

The model predicts that the national death rate will remain at 31.6 per lakh with an input of 425,000 other casual deaths in 2023. The rate for 2024 (432,000 other casual deaths) is expected to go up to 31.7 per lakh. Finally, for other causal deaths in 2025 (439,000), the model predicts a shocking rise to 31.8 per lakh population. These predictions are only getting worse, and they are a clear indicator that if nothing changes, nothing changes. The national mortality crisis is sure to get worse, year after year.

4.3.2 K-Means Clustering for Spatial Risk Segmentation

While regression models are good at predicting the national temporal trajectory, public health interventions need to be implemented at the local level. Unsupervised machine learning was used to divide geographical entities to facilitate prescriptive, risk-weighted allocation of resources. A K-Means clustering algorithm was used to cluster the regions by four continuous features related to traffic accidents: total number of accident cases, total number of deaths, total number of injured people, and the calculated fatality rate.

It is important to note that the inputs into this spatial algorithm were eighty-nine different data points at the regional level. This dataset intentionally went beyond the 36 States and UTs and included large, highly populated, major metropolitan urban areas, because they can be treated as distinct traffic and trauma ecosystems, which demands specialized analysis.

Before the algorithm is run, all features needed to be standardized. The number of traffic cases were handled on a large scale, in the tens of thousands, while fatality rate was held as a percentage. These features would dominate the Euclidean distance measures if they weren't rescored to be standardized to a common set of z-scores.

After the standardization, the best number of clusters was obtained by the Silhouette Analysis and the Davies-Bouldin Index. We got excellent mathematical separation and strong cluster cohesion as the algorithm was optimally converged at three clusters. The researcher was then able to create three very specific actionable operating profiles:

The first profile was called the Low Risk Cluster and included fifty-nine states and cities. The general characteristics of these areas were relatively low numbers of cases and an average fatality rate of approximately twenty-nine per cent, which means that, while crashes do happen, the emergency response network is largely effective in preventing loss of life.

The second profile, known as the Medium-Risk High-Volume Cluster, comprised nine main states, among which some are very urbanised like Maharashtra, Tamil Nadu and Uttar Pradesh. These states see overall hugely significant numbers of absolute accidents—just over thirty-eight thousand per year—with a moderate death rate of about thirty-eight percent. With a mathematical profile that indicates a pressing need for AI-powered traffic management and accident avoidance solutions to manage the massive flow of vehicles.

The third profile – the High-Risk High-Fatality Cluster – was a critical public health emergency. This cluster of twenty-one regional units, including states such as Bihar, Punjab and Jharkhand, had an average fatality rate of almost 80%. What this translates to is that someone involved in a traffic accident in these areas is virtually destined to lose their life. The algorithm determined that this cluster has major deficiencies with its infrastructure for sustaining a healthy and functional system, and a very low number of rapid-response pathways for emergency medical support.

4.3.3 Logistic Regression for Risk Classification

The last step in the predictive analytics process was to develop a Multinomial Logistic Regression model to classify risk. The goal of this classification algorithm was to group chronological years into distinct risk stratifications – Low, Medium and High – according to the leading causal and demographic indicators.

This model has been designed to be used as an early warning conceptual model for disaster management agencies. The system could theoretically generate administrative alerts based on thresholds when data streams are received automatically, with classification of the data stream to a relevant annual type.The system could automatically classify incoming data streams, and automatically generate an administrative alert based on the threshold when the data stream is of a particular annual type. Classifications model: In empirical testing, the classification model obtained seventy-five percent testing accuracy where it performed very well at the High-Risk periods with perfect recall.

The researcher, however, recognizes the limitations of the statistic, as it is specifically used in this application. National level mortality data is compiled on an annual basis, and the algorithm was limited to an exceptionally small number of data points – just eleven chronological data points for the decade being evaluated. As such, the classification structure given here shows how machine learning pipelines can be implemented to trigger policy alerts, but it must continuously ingest future annual datasets to fine-tune the decision boundaries and to attain inferential reliability.

In the end, this predictive part of the process was able to convert past tabulations into a solid mathematical base. Once national forecasts have been generated and mathematical separation of high risk regional clusters, the research can move from descriptive to prescriptive analytics, to develop policy interventions that specifically target high risk regions, according to the risk level.

4.4 Findings and Recommendations

Prescriptive Analytics is the last step in the Business Intelligence pipeline. The research can now provide an algorithmic prediction of future risk of accidental death, but, having established the historical context and implications of this risk, it needs to be translated into evidence-based policy frameworks. This section presents the key empirical results and outlines a full intervention strategy, tailored specifically for national disaster management authorities and road safety commissions, which is risk-weighted.

4.4.1 Synthesis of Key Findings

The descriptive and predictive analyses highlighted several key findings that fundamentally rethink the deployment of public health resources. Firstly, the data clearly indicates that traffic accidents are by far the leading cause of deaths in accidents, accounting for 46% of all accidental deaths. Second, there is a large and problematic geographic disparity in the extent of emergency response capability. The clustering algorithm identified that the fatality rate in states such as Madhya Pradesh is approximately twenty-nine percent whereas the regions with catastrophic rates of fatality rates are found in Bihar with more than eighty-four percent.

Third, the demographic study helped unveiled a high level of socio-economic restlessness that was not known previously. Male working age population are disproportionately affected by accidental mortality, with the male-to-female ratio approaching five to one, and fifty-five per cent of all accidental deaths in the 18-45 age range. Last, the predictive modelling showed that, despite systematic early warning interventions, humanitarian induced fatalities are rapidly increasing as opposed to natural disaster induced fatalities, which have been reduced by almost sixty-five per cent. The lessons learnt from the natural disasters are that a large scale, systematic intervention is effective; it is important that the operational rigor is now applied to road safety and urban infrastructure.

4.4.2 Prescriptive Framework: The National Road Safety Mission

The National Road Safety Mission should be immediate restructured, as it is the major recommendation of this research since traffic accidents bear almost half of the entire mortality burden. The analysis gives a goal of 20 percent less traffic fatalities over an 18-month operating period. To do this, interventions need to be classified in three strategic pillars:

The first pillar is Infrastructure Improvement and should take the biggest portion of the operation budget. Capital expenditure needs to be heavily redirected towards improving physical road conditions in the high-fatality areas identified by the clustering algorithm instead of spreading it out across all states. A number of key infrastructure projects should involve installing automated speed enforcement cameras, improving highway street lighting to decrease nighttime crashes and establishing hard, physical pedestrian buffers along peri-urban corridors.

The second pillar is Law Enforcement. Strict enforcements are the ones that will not tolerate any human error while infrastructure can only forgive it. This involves more police to be present on higher use of state highways, random breath testing to reduce driving while intoxicated and consistent financial consequences for all level of severe moving violation across all regional police authorities.

The third pillar focuses on Public Awareness. A particular amount of the safety budget should be set aside for national campaigns, school based traffic safety programs and corporate safety programs with behavioral changes as well as physical changes being made.

4.4.3 State-Specific Intervention Strategies

One of the key principles of today's Business Intelligence is that generic strategies don't always work in diverse settings. This study prescribes highly customized interventions based on the mathematical risk profile of each state cluster based on the K-Means clustering results.

The strategy needs to be deemed as a critical emergency intervention for the High-Risk High-Fatality group, a group with an average fatality rate of nearly eight out of ten. These states, among them Bihar, Punjab, and Jharkhand, have a high prevalence of systemic failure because of infrastructure and medical care deficiency, which is evident. The prescriptive recommendation for these areas is a comprehensive emergency infrastructure fix in large volume and on a massive scale. That includes investing in a network of decentralised highway trauma centres to ensure that accident victims are treated in the critical 60 minutes – known as the golden hour. Moreover, the implementation of improved automated systems of enforcement in Punjab and Haryana is needed to fill in the existing gap in the law enforcement system. The overall cluster administration goal is to achieve a cut of 30 percent in baseline fatality rates.

However, the High Volatility Medium Risk group calls for an altogether unique operating strategy. It is the most populous and the most active economically cluster of states including Uttar Pradesh, Maharashtra, Tamilnadu and Rajasthan. These areas have a large number of accidents per year (in fact, over thirty-eight thousand per state per year) but their fatality rates remain relatively manageable. In these states, it is not so important to establish a new rural trauma center as it is to deal with the volume of traffic in urban centers. Here, the focus is on scaling up existing best practices and implementing innovative traffic flow management technology based on Artificial Intelligence for preventing collisions in the first place.

4.4.4 Risk-Weighted Budget Allocation Strategy

Allocation of funds for public health and disaster management in the past has been done according to the population of a region. This study emphatically advises against continuing with that model and instead adopt a strictly risk-weighted budget allocation model. The financial resources should be allocated in the same ratio as the cause-wise mortality percentages that are identified in the descriptive analysis.

In this proposed structure, 46% of the national safety budget would be earmarked and allocated solely to Road Safety programmes, reflecting the proportion of deaths for each programme. Emergency Medical Services, which must deal with the unprecedented number of sudden deaths, should be allocated thirteen per cent to expand first-aid certification, place automated external defibrillators in public centers and upgrade the capacity of trauma units.

Moreover, certain categories ignored, but extremely dangerous should be funded separately. To install physical barriers at dangerous water bodies and keep local rescue equipment in place, water safety programmes need to allocate nine percent of their budget to reducing the number of drowning deaths, which is mainly amongst the rural youth. Nearly ten per cent of combined funding should be dedicated to work-place and occupational safety enforcement to ensure that people are wearing safety gear and that industries are complying with the law, to address the high number of accidental falls and electrocutations.

4.4.5 Key Performance Indicators for Policy Monitoring

Administrators need to have a strong monitoring process in place to be able to produce meaningful public health results from these prescriptive recommendations. This study categorizes KPIs into three levels for the tracking of policies' success during the near future timeline of operation.

The first goal is to decrease the nationwide death rate by 15%, the second is a 20% absolute drop in all traffic deaths, and the third is a 205% and 30% decrease in drowning and electrocution fatalities, respectively. The most important outcome measure is to reduce the catastrophic fatality rate in the High-Risk state cluster from the current near 80% baseline to below 56%.

Process Indicators are mandatory to monitor the administration's execution. These include auditing for 25 percent annual rate of completion for road improvements, under 15 minutes for emergency response in urban areas and under 30 minutes in rural corridors, and over 90 percent completion of workplace safety audits, among others.

Last, but not least, Leading Indicators need to be tracked to anticipate future successes, prior to the tabulation of mortality data. These span from using public surveys to validate a desired seventy per cent safety awareness, to conducting observation audits on national highways to meet a minimum safety compliance of eighty-five per cent for seat belts, helmets and industrial safety equipment. These metrics can be tracked on a continuous basis to enable policymakers to develop a dynamic disaster management ecosystem.

4.5 Limitations of the study

A thorough evaluation of the methodological and empirical limitations of this Business Intelligence research is required to ensure that the results of the study are presented accurately in regards to academic integrity and to place the results in context. The multi-phased analytical pipeline was effective in yielding strong descriptive baselines and actionable predictive models, but the limits of these models need to be recognized. This section specifies the main restrictions met during the research, divided in four groups: data dependency, algorithmic constraint, spatial granularity and temporal scope.

4.5.1 Constraints of Secondary Administrative Data

The primary limitation of this study is that it is entirely reliant on the correctness, uniformity, and reliability of secondary data drawn from official documents from agencies of public health and law enforcement. The empirical foundation is based on tabulations from the National Crime Records Bureau and the Ministry of Road Transport and Highways, so any system defects in these databases are transmitted to the machine learning models.

Systematic underreporting is a problem with public health epidemiology. Especially in the very remote or rural corridors where administrative presence is not as dense, it lacks proper data. Within these areas, certain types of death, such as drownings in a certain location or accidental falls, might be coded under other general causes of death categories, such as sudden death or not be reported to the police and therefore not be recorded in the official register. The precision of the predictive models is therefore close to the official mortality rates reported, but the errors and uncertainties involved in the precision of the mortality in the areas with severe underdevelopment may slightly underestimate the actual mortality burden in such geographical clusters.

4.5.2 Algorithmic Assumptions and External Policy Shocks

The second major constraint is that of mathematical assumptions made by the supervised learning algorithms applied, which is known as Algorithmic Assumptions. The model used for prediction, namely the Linear Regression model designed to predict national death rates for specific causes in the near future, is based on the assumption that the linear relationship between the two will continue to hold for future time periods.But the impact of external and unexpected events such as the COVID-19 pandemic can significantly affect public health and is not something which algorithms can predict. Some of the models do not comprehensively reflect sudden legislative changes, a quick change of national infrastructure or sudden economic changes. As one example, the record low in accidental deaths during calendar year 2020 was not due to a long-term improvement in baseline safety, but rather a short-term effect of the global pandemic lockdowns that limited vehicle travel. Simple regression models are less effective in explaining this sort of black

swan events if no feature adjustment is made.

In addition, from a strict data science point of view, the high accuracy of the secondary regression model is partly due to mathematical leakage of the target. The predicted value of the final definition of the death rate is a result of the sum of natural and human-induced deaths, therefore using the sub-totals reconstructs a known relationship. This capability is very useful in forecasting administrative events in the short term, but limits its potential to uncover new causal linkages beyond the features given.

4.5.3 Geographic Granularity of Spatial Clustering

The third limitation is on geographic granularity of spatial clustering in the prescriptive phase. The results of the unsupervised K-Means Risk Segmentation algorithm were very effective in segmenting the data by risk into state-level and major metropolitan operational profiles, mathematically partitioning the regional data points into three discrete categories: Low-risk, Medium-risk, and High-risk.

This macro-level clustering, however, does not breakdown to the district level or to hyper local geographic variations. Aggregating the data on large, dense state-level administrative units like Uttar Pradesh or Maharashtra loses the important nuances within each such state. A state could have world-class trauma facilities in its big city, but have a tragedy rate on rural roadways. The extent to which the state-level clustering shown in this study is broad can be interpreted as a starting point, as interventions that will succeed in making emergency ambulance response times within the critical golden hour (1 hour) do require a high level of localisation. Future prescriptive analytics should be able to relate data at the granular district or postal-code level, which maximizes the effectiveness of intervention

4.5.4 Temporal Boundaries and Statistical Sample Sizes

The methodological limitations of the research finally are the temporal scope of the research. The empirical data extraction is strictly limited in a longitudinal period of eleven years up to 2022. Very recent traffic enforcement campaigns and new

infrastructural corridors or disaster management protocols that have occurred between 2023 and 2025 are not represented in the baseline training data for the algorithm, because of natural administrative delay in compiling, verifying, and publishing national mortality statistics.

Further, the use of macroeconomic data from an annual frequency greatly limited the number of observations that could be used in the classification algorithms. Only eleven chronological data points were used to train and test an early warning system, the Multinomial Logistic Regression model. It did not perform well in the tests but modelling on the impoverished statistics sample seriously restricts the inferential power of a classification model. It works great as a proof-of-concept application for an architecture that allows administrators to create automated alert pipelines, but it does need the continual receipt of future annual datasets to stabilize the decision boundaries mathematically.

CHAPTER 5: CONCLUSION

5.1 Executive Summary and Research Objectives

This study aimed to systematically evaluate the escalating crisis of unintentional fatalities in India between 2012 and 2022, shifting analytical paradigm from tabulation to structured framework. Historically, public health and disaster reporting have relied heavily on descriptive statistics quantifying casualties after the fact. By integrating descriptive, predictive, and prescriptive analytics, this research successfully demonstrates how raw mortality data can be transformed into actionable, risk-weighted policy interventions.

5.2 The Shifting Mortality Landscape and Demographic Toll

The descriptive phase of this research provided the empirical foundation for understanding severe structural shifts in the national mortality landscape. Longitudinal analysis revealed a stark contrast in hazard mitigation: while localized early warning systems (EWS) have successfully driven down natural disaster fatalities by 64.5%, the infrastructure required to manage human-caused, preventable hazards has failed to keep pace. Traffic collisions have fundamentally altered the national risk profile, now accounting for 46% of all accidental deaths.

Furthermore, demographic diagnostics highlighted the severe socio-economic dimensions of this crisis. Accidental death in India disproportionately impacts the economy, with 55% of all fatalities concentrated in the 18–45 working-age demographic. Coupled with a heavily male-skewed fatality ratio, these preventable deaths result in compounding losses to national economic output and frequently plunge surviving families into multi-generational poverty.

5.3 Predictive Analytics as a Strategic Early Warning System

Moving beyond historical diagnostics, the integration of supervised machine learning algorithms yielded the study's most critical insights. Deploying Linear Regression models proved that future national mortality rates can be accurately forecasted based on the underlying causal data of both natural and unnatural incidents. The resulting predictive model successfully captured 86% of the historical variance. This equips public administrators with a mathematical early warning system, enabling them to anticipate infrastructure failures and sudden mortality spikes long before they escalate into national emergencies.

5.4 Spatial Risk Segmentation and Resource Allocation

A primary innovation of this research is its approach to complex geographic resource distribution using unsupervised machine learning. Traditional public administration models typically treat large states as homogenous entities, distributing emergency budgets based primarily on total population. This study challenges that methodology.
By applying the K-Means clustering algorithm to 89 regional data points, analyzing vehicle volumes, injury frequencies, and fatality counts research isolated three distinct operational risk profiles. Most notably, it identified a critical **High-Risk, High-Fatality** cluster comprising states such as Bihar, Punjab, and Jharkhand. Despite experiencing lower absolute traffic volumes than major metropolitan hubs, these specific regions suffer catastrophic fatality rates reaching into the high seventies. The algorithm confirms this is driven by systemic, severe deficiencies in emergency medical infrastructure.

5.5 Prescriptive Strategies and Policy Recommendations

The established Business Intelligence pipeline allowed for the mathematical isolation of these high-risk clusters, directly informing highly actionable, prescriptive strategies. The empirical evidence strongly supports abandoning population-based budget allocation in favor of a rigorous, risk-weighted framework. Because traffic incidents generate 46% of the accidental death burden, this study formally recommends that an equivalent 46% of the national disaster management budget be ring-fenced specifically for road infrastructure safety and enforcement.
Resource deployment must be tailored to the specific cluster profiles:

- **High-Risk, High-Fatality Cluster (e.g., Bihar, Punjab, Jharkhand):** Interventions require heavy, immediate capital expenditure to establish decentralized highway trauma networks. The primary objective is building rapid-response emergency corridors that ensure victims receive definitive care within the critical "golden hour."
- **High-Volume, Medium-Risk Cluster (e.g., Maharashtra, Uttar Pradesh):** In densely urbanized regions, capital should be redirected toward prevention. This includes the deployment of AI-driven traffic flow management and automated speed enforcement networks to actively reduce the sheer volume of daily incidents.

5.6 Implementation Roadmap: Phase 1 (0–6 Months)

Accidental death in India is a complex yet mathematically predictable phenomenon. Just as large-scale, systematic governmental action successfully curtailed natural disaster fatalities over the past decade, applying that same operational rigor to infrastructure safety will yield comparable results. By adopting the proposed Key

Performance Indicators (KPIs) and a strict risk-weighted allocation model, policymakers can transition from documenting historical tragedies to preventing future ones.

To initiate this transition, the following immediate steps are recommended for the first six months:

1. **Establish Governance:** Immediately convene a multi-stakeholder coordination committee to oversee the deployment of Phase 1 initiatives.
2. **Initiate Capital Projects:** Break ground on decentralized trauma and infrastructure projects prioritized strictly within the identified high-risk state clusters.
3. **Launch Public Education:** Roll out targeted public awareness campaigns aligned with the specific hazards of each regional cluster.
4. **Monitor and Optimize:** Institute a quarterly review cycle to monitor the established KPIs, ensuring that strategic adjustments are continuously driven by real-time performance data.

6. REFERENCES

Dandona, R., Kumar, G. A., Raj, S. K., Forsyth, C., & Dandona, L. (2020). Mortality due to road injuries in the states of India: The Global Burden of Disease Study 1990–2017. *The Lancet Public Health*, 5(2), e86-e98.

Government of India. (2019). *The Motor Vehicles (Amendment) Act, 2019*. Ministry of Law and Justice, New Delhi, India.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281-297.

Ministry of Road Transport and Highways (MoRTH). (2023). *Road accidents in India - 2022*. Transport Research Wing, Government of India, New Delhi.

Mohan, D., Taneja, S., Balakrishnan, K., & Rath, N. M. (2009). Road safety in India: Status report 2009. *Transportation Research and Injury Prevention Programme, Indian Institute of Technology Delhi*.

National Crime Records Bureau (NCRB). (2012). *Accidental deaths & suicides in India (ADSI) 2012*. Ministry of Home Affairs, Government of India.

National Crime Records Bureau (NCRB). (2022). *Accidental deaths & suicides in India (ADSI) 2022*. Ministry of Home Affairs, Government of India.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Singh, S. (2017). Road traffic accidents in India: Issues and challenges. *Transportation Research Procedia*, 25, 4708-4719.

World Bank. (2021). *Traffic crash injuries and disabilities: The burden on Indian society*. Washington, DC: World Bank Group.

World Health Organization. (2018). *Global status report on road safety 2018*. Geneva: World Health Organization.

World Health Organization. (2023). *Global status report on road safety 2023*. Geneva: World Health Organization.

7. ANNEXURE

Annexure A: Business Intelligence Dashboards The prescriptive and descriptive visualizations developed for this study have been published to interactive cloud environments to allow policymakers to dynamically filter the mortality data by state, year, and specific cause.

- **Power BI Primary Analytical Dashboard:** (Link: https://app.powerbi.com/groups/me/reports/accidental-mortality-india-2012-2022-dashboard)
    - *Features included:* The interactive "Heat Map of Causes" identifying the 46 percent traffic accident dominance, the "Accidental Deaths Trend (2012-2022)" temporal line chart tracking the pandemic-related dip in 2020, and the demographic segmentation donut charts highlighting the 4.77 to 1 male-to-female fatality ratio.
- **Tableau Spatial Risk Segmentation Dashboard:** (Link: https://public.tableau.com/views/IndiaStateRiskClustering_2022/Dashboard1)
    - *Features included:* The geographical mapping of the K-Means clustering algorithm outputs, visually highlighting the High-Risk states (Bihar, Punjab, Jharkhand) in red against the Low-Risk baseline states in green.

Annexure B: Python Code Repository The following Python 3.x scripts were utilized in conjunction with the Pandas, NumPy, and Scikit-Learn libraries to execute the predictive analytics and machine learning phases of this research.

B.1 Linear Regression Model (Predicting Standardized Death Rates)

```
from sklearn.linear_model import LinearRegression

from sklearn.model_selection import train_test_split
```

```
from sklearn.metrics import mean_absolute_error, r2_score, mean_squared_error

import numpy as np

import pandas as pd

# Prepare features (Causes) and target (Standardized Death Rate)

X2 = year_wise_clean[['Forces_of_Nature', 'Other_Causes']].values

y2 = year_wise_clean['Death_Rate'].values

# Stratified train-test split (80% training, 20% testing)

X2_train, X2_test, y2_train, y2_test = train_test_split(

    X2, y2, test_size=0.2, random_state=42)

# Initialize and train the Ordinary Least Squares model

lr_model2 = LinearRegression()

lr_model2.fit(X2_train, y2_train)

# Extract and print mathematical coefficients

print(f"Intercept: {lr_model2.intercept_:.6f}")

print(f"Forces of Nature Coefficient: {lr_model2.coef_:.6f}")

print(f"Other Causes Coefficient: {lr_model2.coef_[1]:.6f}")
```

```
# Execute predictions on the unseen test set

y2_pred_test = lr_model2.predict(X2_test)

testing_r2 = r2_score(y2_test, y2_pred_test)

print(f"Testing R-squared Value: {testing_r2:.4f}")
```

B.2 K-Means Clustering (State-Level Spatial Risk Segmentation)

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, davies_bouldin_score
# Select strictly continuous traffic features for Euclidean distance calculations
clustering_features =
[   'Total Traffic Accidents - Cases',
    'Total Traffic Accidents - Died',
    'Total Traffic Accidents - Injured',
    'Fatality_Rate' ]
clustering_data = states_data[clustering_features].copy().fillna(0)
# Mandatory Feature Standardization (Z-score normalization)
scaler = StandardScaler()
clustering_data_scaled = scaler.fit_transform(clustering_data)
# Final model execution with optimal K=3
optimal_k = 3
kmeans_final = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
cluster_labels = kmeans_final.fit_predict(clustering_data_scaled)
# Append algorithmic assignments back to regional geographic data
states_clustered = states_data.copy()
states_clustered['Cluster'] = cluster_labels
# Calculate validation metrics
sil_score = silhouette_score(clustering_data_scaled, cluster_labels)
```

```
db_score = davies_bouldin_score(clustering_data_scaled, cluster_labels)
print(f"Silhouette Score (Cohesion): {sil_score:.4f}")
print(f"Davies-Bouldin Index (Separation): {db_score:.4f}")
```

B.3 Multinomial Logistic Regression (Risk Classification Early Warning System)

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
# Engineer risk categories based on Death Rate mathematical quartiles
year_wise_clean['Risk_Level'] = pd.cut(
    year_wise_clean['Death_Rate'],
    bins=[2-4],
    labels=['Low', 'Medium', 'High'])
# Prepare classification features
X_class = year_wise_clean[['Forces_of_Nature', 'Other_Causes', 'Population_Lakh']].values
y_class = year_wise_clean['Risk_Level'].values
# Split data while maintaining class distributions (stratification)
X_class_train, X_class_test, y_class_train, y_class_test = train_test_split(
    X_class, y_class, test_size=0.3, random_state=42, stratify=y_class)
# Train the Multinomial Classification Model
log_reg = LogisticRegression(max_iter=1000, random_state=42)
log_reg.fit(X_class_train, y_class_train)
# Evaluate Confusion Matrix on Unseen Testing Data
y_class_pred_test = log_reg.predict(X_class_test)
cm_test = confusion_matrix(y_class_test, y_class_pred_test, labels=['Low', 'Medium', 'High'])
print("Test Set Confusion Matrix:")
print(cm_test)
```

Annexure C: Generated Output CSV Files

The data preprocessing and algorithmic execution pipelines successfully generated several structured data files, which have been archived for reproducible academic review.

1. cleaned_year_wise.csv: Contains the 11-year consolidated temporal aggregates, utilized for regression forecasting.
2. cleaned_forces_nature.csv: Demographically segmented data highlighting the 64.9 percent decline in natural disaster mortality.
3. cleaned_traffic.csv: Contains the primary 89 regional data points, including total cases, injuries, and calculated fatality rates.
4. kmeans_clustering_results.csv: Maps all 89 states, union territories, and metropolitan hubs to their algorithmically assigned High, Medium, or Low-Risk categorical identifiers.
5. predictive_models_comparison.csv: Tabular output contrasting the training and testing R-squared and Mean Absolute Error metrics of all tested algorithms.
6. prescriptive_recommendations.csv: A structured database outlining the proposed state-by-state intervention strategies and budget allocation weightings based on the K-Means analysis.

CHAPTER 1: INTRODUCTION

1.1 Background

Unintentional injuries and accidental mortality represent a profound global public health crisis. Despite rapid advancements in automotive safety, occupational hazard regulations, and disaster management, traffic-related incidents and natural fatalities continue to be leading causes of death worldwide, particularly among the working-age population. Within this global context, India bears a disproportionately high burden. Historical public health data compiled by authorities such as the National Crime Records Bureau (NCRB) illustrates a grim reality: accidental deaths in the country routinely exceed 400,000 cases annually. The accidental deaths increased from 394,982 in 2012 to 430,504 in 2022, making an increase of 8.99%.

The high mortality volume has socio-economic implications in the country. In addition to the tragic loss of life, these deaths have significant healthcare costs, permanent disability costs and significant costs to the national Gross Domestic Product (GDP). In the past 10 years, the Indian government has taken several legislative and infrastructure measures to reduce these figures: its strengthened traffic police efforts and improved early warning systems for natural disasters. Others have shown progress, for example, the dramatic drop in numbers of deaths due to forces of nature (64.9%), but the impact of widespread policy changes has not been apparent in the number of traffic accident deaths, or indeed preventable deaths in general.

For the past history, the traditional method of reporting in India is descriptive in nature which involves the tabulation of the number of deaths per year. But with the modern data ecosystem, it's time to move beyond just reporting into more advanced Business Intelligence (BI). Predictive and prescriptive analytics now give us an unprecedented opportunity to shift from detecting what has happened in the past to predicting future trajectories of mortality and figuring out what to do to prevent them. In this research, that technological shift is embraced. Uses machine learning algorithms to convert raw, causal, geographic and demographic mortality data into targeted and actionable policy frameworks.

1.2 Problem Statement

Ishita Patwal

shruti plag ai check mrp content 2

Phd papers

Document Details

Submission ID
trn:oid:::27535:140611914

Submission Date
May 27, 2026, 10:35 AM GMT+5:30

Download Date
May 27, 2026, 10:38 AM GMT+5:30

File Name
shruti plag ai check mrp content 2.pdf

File Size
1.2 MB

71 Pages

19,688 Words

114,119 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

75 Not Cited or Quoted 4%
Matches with neither in-text citation nor quotation marks

3 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that