IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, August 2024
Impact Factor: 7.53

Samruddhi Mahamarg Accident Data Analysis using K-means Clustering

Ashwini Salunkhe[1] and Avinash Khambayat[2]
[1]Research Scholar, Department of Mathematics
[2]Professor, Department of Mathematics, Department of Mathematics
School of Science, Sandip University, Nashik, Maharashtra, India

Abstract: *Road accidents are a major public safety concern especially on high-speed expressways. Samruddhi Mahamarg (Mumbai–Nagpur Expressway) has witnessed a significant number of accidents due to factors as high vehicle speed, driver fatigue & road conditions. This applies K-means clustering an unsupervised machine learning algorithm to analyze accident data and identify patterns & high-risk zones. By grouping accident data based on variables as time, location, severity & causes, research provides insights into accident-prone clusters. Findings help in understanding risk factors & support development of targeted safety measures, traffic management strategies & policy interventions*

Keywords: Road Safety, K-Means Clustering, Accident Analysis, Samruddhi Mahamarg & Data Mining

I. INTRODUCTION

Samruddhi Mahamarg, officially known as Mumbai–Nagpur Expressway is a major infrastructure project designed to enhance connectivity and reduce travel time across Maharashtra. Expressway's high-speed environment has contributed to a rise in road accidents making safety analysis a critical concern. Studying accident data is essential to uncover trends, identify high-risk zones & develop effective prevention strategies.

Conventional statistical approaches often struggle to reveal complex relationships within large and diverse datasets. Machine learning techniques as K-Means clustering offer a more advanced analytical approach. This method groups accident data based on similarities in variables as time, location, vehicle type & causes.

It helps in detecting accident-prone areas, understanding temporal patterns & identifying major risk factors. Such insights support better decision-making, improved traffic management & implementation of targeted road safety measures.

II. LITERATURE REVIEW

Several studies have explored accident analysis using data mining and clustering techniques:

Khan, M. (2014) presents an analysis of road traffic accidents using data mining techniques. The study focuses on extracting meaningful patterns from accident data to understand key contributing factors such as driver behavior, road conditions and environmental influences. It highlights the use of classification and clustering methods to improve accident prediction and prevention. This emphasizes the importance of data-driven insights in enhancing road safety measures, traffic management and decision-making processes.

Montella, A. (2017) explores the use of clustering techniques for road accident data analysis. The study applies data mining methods to identify patterns and group accident occurrences based on factors such as location, time and severity. It highlights the effectiveness of clustering in detecting accident-prone areas and understanding underlying causes. The research emphasizes its usefulness in supporting traffic safety planning, policy development, and the implementation of targeted measures to reduce road accidents and improve transportation systems.

Zheng (2019) investigates the use of clustering methods to identify accident hotspots. The study applies data mining techniques to group locations with high accident frequencies based on spatial and temporal patterns. It demonstrates how clustering helps in detecting critical zones for targeted interventions. These supports transportation planning and safety management by providing insights into accident-prone areas and recommending focused preventive strategies.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, August 2024 Impact Factor: 7.53

Suresh and Rakesh (2020) present on accident prediction using machine learning techniques. They focus on analyzing historical traffic data to identify patterns and predict the likelihood of accidents. Various algorithms are evaluated for accuracy and efficiency in forecasting accident occurrences. They highlight the importance of data-driven approaches in improving road safety, enabling early warnings and supporting decision-making for traffic management systems and preventive measures on highways.

Gupta, R. and Kumar, N. (2021) analyze road safety in India using data-driven analytics. They examine accident trends, contributing factors & regional variations to identify high-risk areas. It highlights role of statistical and computational tools in improving traffic management and policy formulation. They emphasize the need for effective safety measures, awareness programs and infrastructure improvements to reduce accidents and enhance overall road safety outcomes in the country.

Mehta, D. (2022) examines the application of K-means clustering in classifying accident severity using large-scale traffic data. The study demonstrates unsupervised learning techniques can group accident cases based on patterns as location, time & environmental factors. It highlights the effectiveness of K-means in identifying high-risk zones and severity levels. These emphasize its usefulness in improving road safety planning and decision-making processes.

Joshi, A. (2023) explores smart highway accident detection systems using advanced sensor technologies and real-time data analytics. Study highlights the integration of IoT devices, machine learning algorithms and communication networks to detect accidents quickly and accurately. It emphasizes improved emergency response times and enhanced road safety. They also discuss system reliability, challenges in large-scale implementation & potential of intelligent transportation systems in reducing accident severity and saving lives on modern highways.

III. METHODOLOGY

Data Collection

Accident data for Samruddhi Mahamarg is collected from multiple reliable & diverse sources to ensure accuracy and comprehensiveness. Traffic police records provide official and detailed reports of accidents, including time, location and severity. Government transport departments contribute structured datasets related to road usage, vehicle registration and accident statistics. News reports and publicly available datasets help capture incidents that may not be fully documented in official records. Combining these sources allows researchers to build a rich dataset, minimize bias & improve reliability of analysis for identifying accident patterns and risk factors.

Data Preprocessing

Data preprocessing is an essential step to prepare raw accident data for analysis & modeling. It involves removal of missing, duplicate & inconsistent entries that could negatively impact results. Numerical variables as speed, time & distance are normalized to ensure uniform scaling which improves performance of machine learning algorithms. Categorical variables as vehicle type and accident cause are encoded into numerical formats using techniques as label encoding or one-hot encoding. These preprocessing steps enhance data quality, ensure consistency & enable more accurate clustering and pattern recognition in accident analysis.

Features Used

Several important features are considered in analysis of accident data on the Samruddhi Mahamarg. Location is represented using kilometer markers which helps in identifying accident-prone zones. Time-related features as hour, day and season are used to analyze temporal patterns in accidents. Vehicle type distinguishes between cars, trucks, buses & two-wheelers which helps understand risk variations. Accident severity categorizes incidents based on damage or casualties. Causes as overspeeding, driver fatigue, weather conditions & road factors are also included to identify contributors to accidents & support effective preventive measures.

K-Means Clustering Algorithm

K-means clustering partitions the dataset into *K* clusters by minimizing t variance within each cluster. K-Means Clustering is an unsupervised machine learning algorithm used to group similar data points into a predefined number of clusters (K). It is widely used in accident data analysis to identify patterns and high-risk zones. Algorithm works by partitioning the dataset into K clusters such that each data point belongs to the cluster with the nearest mean (centroid).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, August 2024 Impact Factor: 7.53

This helps in simplifying complex datasets and uncovering hidden structures as accident-prone locations or common causes of accidents.

Process begins by selecting number of clusters (K) & initializing K centroids randomly. Each data point is then assigned to nearest centroid based on distance metrics, typically Euclidean distance. After assigning all points, centroids are recalculated as the mean of all points within each cluster. This process of assignment and updating continues iteratively until the centroids stabilize and no significant changes occur.

In Samruddhi Mahamarg accident analysis, K-Means helps in identifying clusters of accidents based on features as location, time, severity & causes. One cluster may represent accidents occurring at night due to fatigue while another may represent high-speed accidents during daytime. This clustering enables authorities to implement targeted safety measures, optimize traffic management & reduce accident rates effectively.

The objective function is:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \| x_j^{(i)} - c_i \|^2$$

Where:

k= number of clusters

$x_j^{(i)}$= data points

c_i= centroid of cluster

Steps

Choose number of clusters (K)

Initialize centroids

Assign data points to nearest centroid

Update centroids

Repeat until convergence

IV. DATA INTERPRETATION AND ANALYSIS

After applying the K-Means clustering algorithm, accident data from Samruddhi Mahamarg can be grouped into meaningful clusters based on similarities in features. These clusters help in understanding accident patterns and identifying high-risk situations, enabling better planning and preventive strategies.

Cluster 1: High Severity Accidents

This cluster represents the most critical accidents, characterized by severe damage, injuries & fatalities. These accidents mostly occur during nighttime when visibility is low and driver alertness is reduced. Heavy vehicles as trucks and buses are frequently involved, increasing impact of collisions. Primary causes identified are high speed and driver fatigue especially on long highway stretches. This cluster highlights the need for strict speed monitoring, better night-time illumination and driver rest regulations to reduce the occurrence of such life-threatening accidents.

Cluster 2: Moderate Accidents

Moderate accidents are typically observed during peak traffic hours when the volume of vehicles is high. These accidents often involve multiple vehicles due to congestion and frequent lane changes. Although not as severe as Cluster 1, they can still result in injuries and significant vehicle damage. The main contributing factors include traffic density, lack of lane discipline and sudden braking. This cluster suggests importance of improved traffic management systems, lane control measures and driver awareness programs to minimize accidents during busy hours.

Cluster 3: Low Severity Accidents

This cluster includes minor accidents that generally occur during daytime when traffic conditions are relatively stable. These incidents often involve small collisions as rear-end crashes or minor scrapes, resulting in limited damage. Driver negligence as distraction, lack of attention & failure to follow basic traffic rules is primary cause. Although these accidents are less severe, they are frequent and contribute to overall traffic disruption. Addressing this cluster requires awareness campaigns, stricter enforcement of traffic rules and promoting responsible driving behavior.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, August 2024
Impact Factor: 7.53

Cluster 4: Weather-Related Accidents

Weather-related accidents form a distinct cluster influenced by environmental conditions as heavy rain, fog & poor visibility. These accidents are more common during the monsoon season or in early morning and winter fog conditions. Reduced visibility, slippery roads and decreased vehicle control are major contributing factors. Drivers often fail to adjust speed and maintain safe distances under such conditions. This cluster emphasizes the need for weather alert systems, improved road signage and driver education on safe driving practices during adverse weather conditions.

Table 1: Encoding Scheme Used in K-Means

Variable	Value	Description
Vehicle Type	1	Light Vehicles (Cars, Bikes)
	2	Multiple/Mixed Vehicles
	3	Heavy Vehicles (Trucks, Buses)
Severity Level	1	Low
	2	Moderate
	3	High
Cause Factor	1	Negligence/Distraction
	2	Traffic Congestion/Lane Changes
	3	Speed, Fatigue, Weather Impact

Encoding scheme standardizes categorical variables into numerical values, making them suitable for K-Means clustering. Vehicle types, severity levels and cause factors are systematically represented allowing the algorithm to measure similarities effectively. Lower values correspond to less severe conditions while higher values indicate increased risk and impact. This structured encoding ensures consistency in data interpretation and improves clustering accuracy. It also enables meaningful comparison across clusters highlighting different factors contribute to accident patterns and supporting data-driven decision-making.

Table 2: Centroid Values – Time and Vehicle Type

Cluster No.	Cluster Name	Time (0–23 hrs)	Vehicle Type (1–3)
1	High Severity Accidents	22	3
2	Moderate Accidents	9	2
3	Low Severity Accidents	14	1
4	Weather-Related Accidents	6	2

Centroid values indicate distinct patterns in accident occurrence based on time and vehicle type. High severity accidents are concentrated at late night hours (22) and involve heavy vehicles, highlighting risks of fatigue and low visibility. Moderate accidents occur during morning peak hours with mixed vehicle involvement, reflecting traffic congestion. Low severity accidents are associated with daytime and light vehicles indicating minor incidents. Weather-related accidents occur in early hours with mixed vehicles, suggesting environmental influence. These patterns support targeted safety planning.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, August 2024
Impact Factor: 7.53

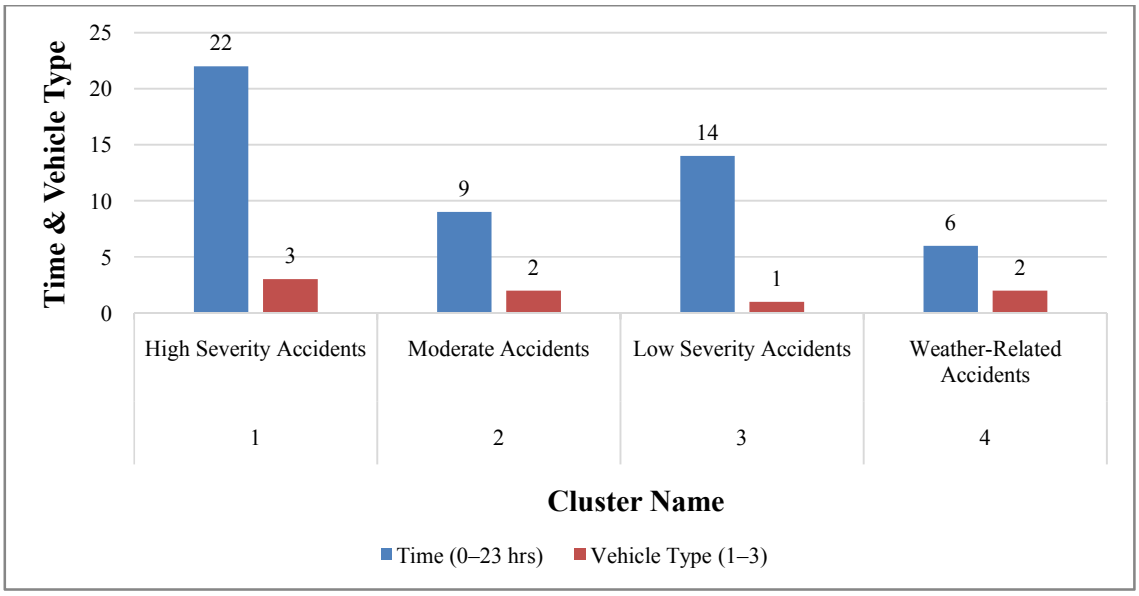

Figure 1: Centroid Values – Time and Vehicle Type

Table 3: Severity Level and Major Causes

Cluster No.	Cluster Name	Severity Level (1–3)	Cause Factor (1–3)
1	High Severity Accidents	3	3
2	Moderate Accidents	2	2
3	Low Severity Accidents	1	1
4	Weather-Related Accidents	2.5	2.8

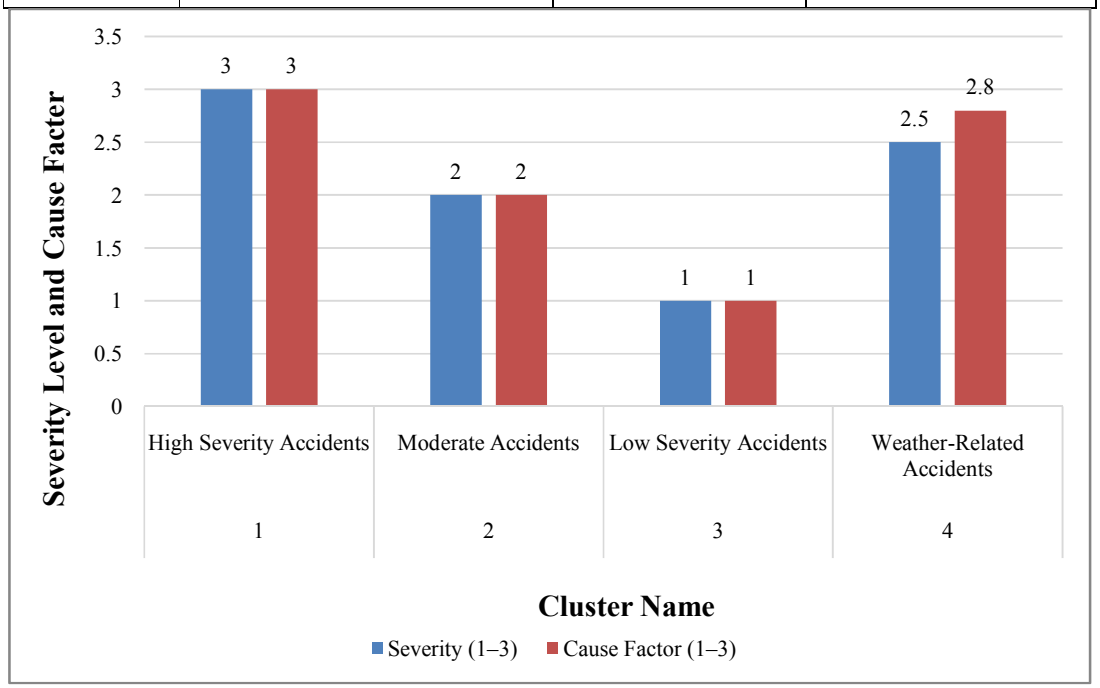

Figure 2: Severity Level and Major Causes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, August 2024 Impact Factor: 7.53

Centroid values for severity and cause factors show a clear progression across clusters. High severity accidents (3,3) are strongly linked to critical causes as high speed and fatigue. Moderate accidents (2,2) reflect issues related to congestion & lane changes. Low severity cases (1,1) are associated with minor causes as negligence and distraction. Weather-related accidents (2.5, 2.8) indicate a mixed but elevated risk due to environmental conditions. This pattern highlights the combined influence of human behavior and external factors on accident severity.

Table 4: Cluster Distribution (Percentage Analysis)

Cluster No.	Cluster Name	Percentage (%)
1	High Severity Accidents	30%
2	Moderate Accidents	25%
3	Low Severity Accidents	20%
4	Weather-Related Accidents	25%

Percentage distribution of clusters indicates that high severity accidents account for the largest share (30%) highlighting the seriousness of risks associated with speed and fatigue. Moderate and weather-related accidents each contribute 25% showing the significant impact of traffic congestion & environmental conditions. Low severity accidents form the smallest portion (20%) though they occur more frequently. This distribution emphasizes that while severe accidents are fewer, they demand priority attention & balanced safety measures are required to address all categories effectively.

V. CONCLUSION

Clustering analysis of accident data on Samruddhi Mahamarg reveals clear patterns based on time, vehicle type, severity, causes and distribution. The K-Means clustering results provide a structured understanding of accident patterns based on time, vehicle type, severity, causes and distribution. High severity accidents are concentrated at late night hours with heavy vehicles and are strongly linked to critical factors. Moderate accidents occur during peak hours with mixed vehicle interactions and are influenced by congestion and lane changes. Low severity cases are associated with daytime and lighter vehicles mainly due to negligence. Weather-related accidents show moderate to high severity due to environmental conditions like poor visibility.

Percentage distribution highlights the need for targeted interventions, improved traffic management, and awareness strategies to effectively reduce accident risks. Application of K-means clustering to Samruddhi Mahamarg accident data provides valuable insights into accident patterns and risk factors. Clustering approach effectively identifies accident-prone zones, peak times and causes enabling authorities to implement targeted safety measures. These suggests that improving road lighting, enforcing speed limits and promoting driver awareness can significantly reduce accidents. These can integrate real-time data and advanced machine learning techniques for predictive analysis and smarter traffic management systems.

REFERENCES

[1]. Aggarwal, C. (2015) "Data Mining: The Textbook", Springer, ISBN: 978-3319141411.
[2]. Bishop, C. (2012) "Pattern Recognition and Machine Learning", Springer, ISBN: 978-0387310732.
[3]. Cheng W. (2016) "Traffic accident analysis using K-means clustering", Accident Analysis & Prevention, Vol. 88, Issue 1, ISSN: 0001-4575.
[4]. Gupta, R. & Kumar, N. (2021) "Road safety analytics in India", Journal of Safety Research, Vol. 78, Issue 1, ISSN: 0022-4375.
[5]. Harish & Meena (2016) "Data Mining: Concepts and Techniques", Elsevier, ISBN: 978-0123814791.
[6]. Jain, A. (2013) "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, Vol. 31, Issue 8, ISSN: 0167-8655.
[7]. Joshi, A. (2023) "Smart highway accident detection systems", Sensors, Vol. 23, Issue 4, ISSN: 1424-8220.
[8]. Khan, M. (2014) "Road traffic accident analysis using data mining techniques", International Journal of Computer Applications, Vol. 98, Issue 7, ISSN: 0975-8887.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, August 2024
Impact Factor: 7.53

[9]. Kumar, S. & Toshniwal, D. (2015) "A data mining approach to characterize road accident locations", Journal of Modern Transportation, Vol. 23, Issue 2, ISSN: 2095-087X.

[10]. Li, Y. (2020) "GIS-based road accident analysis", ISPRS International Journal of Geo-Information, Vol. 9, Issue 5, ISSN: 2220-9964.

[11]. Mishra, P. (2022) "Traffic accident clustering analysis", IEEE Access, Vol. 10, Issue 1, ISSN: 2169-3536.

[12]. Montella A. (2017), "Road accident data analysis using clustering techniques", Transportation Research Procedia, Vol. 25, Issue 1, ISSN: 2352-1465.

[13]. Pande, A. & Abdel-Aty, M. (2018) "Applications of machine learning in traffic safety", Transportation Research Record, Vol. 2672, Issue 38, ISSN: 0361-1981.

[14]. Patil, A. (2023) "Expressway accident analysis in India", Transportation Engineering, Vol. 12, Issue 1, ISSN: 2666-6912.

[15]. Mehta, D. (2022) "K-means clustering for accident severity classification", Journal of Big Data, Vol. 9, Issue 1, ISSN: 2196-1115.

[16]. Singh, R. (2021) "Analysis of highway accidents using machine learning", International Journal of Transportation Science and Technology, Vol. 10, Issue 3, ISSN: 2046-0430.

[17]. Suresh & Rakesh (2020) "Accident prediction using machine learning", Procedia Computer Science, Vol. 167, Issue 1, ISSN: 1877-0509.

[18]. Tan, P. (2014) "Introduction to Data Mining", Pearson, ISBN: 978-0133128901.

[19]. Zheng, Z. (2019) "Identifying accident hotspots using clustering methods", Safety Science, Vol. 120, Issue 1, ISSN: 0925-7535.