2020 IEEE Pune Section International Conference (PuneCon)
Vishwakarma Institute of Technology, Pune India. Dec 16-18, 2020

Road Accident Analysis using Machine Learning

Jayesh Patil
U.G. Student
Department of Computer Engineering
St. John College of Engineering and Management,
Palghar, India
jayeshrpatil3@gmail.com

Mandar Prabhu
U.G. Student
Department of Computer Engineering
St. John College of Engineering and Management,
Palghar, India
mandar.prabhu9@gmail.com

Dhaval Walavalkar
U.G. Student
Department of Computer Engineering
St. John College of Engineering and Management
Palghar, India
dhavalwalavalkar@gmail.com

Vivian Brian Lobo
Assistant Professor
Department of Computer Engineering
St. John College of Engineering and Management,
Palghar, India
lobo.vivian27@gmail.com

2020 IEEE Pune Section International Conference (PuneCon) | 978-1-7281-9600-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/PuneCon50868.2020.9362403

***Abstract*—Accidents through roadways have been a great threat to developed as well as underdeveloped countries. Road accidents and its safety have been a major concern for the world, and everyone is trying to handle this since years. Road traffic and reckless driving occur in every part of the world. Because of this, many pedestrians are affected too. With no fault, they become victims. Many road accidents occur because of numerous factors like atmospheric changes, sharp curves, and human faults. Injuries caused by road accidents are major but sometimes imperceptible, which later on affect health too. This study aims to analyze road accidents in one of the popular metropolitan cities, i.e., Bengaluru, through k-means algorithm and machine learning by scrutinizing accident-prone or hotspot areas and their root causes.**

Keywords—Atmospheric changes, hotspots, k-means algorithm, reckless driving

I. INTRODUCTION

World Health Organization reports that approximately one million three hundred fifty thousand individuals loose their existences and die yearly because of road accidents [1]. Road accidents are avoidable, but small mistakes from an individual's part cause a huge mishap. India's young population lies between 18 to 45 years, which accounts to 70% of road accidents—as per the report by the Ministry of Road Transport and Highways [2]. While road accident awareness has been all over the world, still death rates are alarming. Accidents not only affect a particular person who is injured or hurt but also his/her social contacts. Being a major concern for this era, road accident analysis is neglected. The amount of money invested on roads and their maintenance is still not satisfactory. Highway conditions in rainy season, especially in India, become worse. Potholes are a huge cause of accidents during monsoon. In addition, the conditions of roads in ghats (i.e., sharp curves without any signboards) cause a huge loss to life and loads of traffic for a significant amount of time. This study aims to develop a model that can predict road accidents in one of the popular metropolitan cities, i.e., Bengaluru, which is located in Karnataka, India by considering factors like time, climate, and road type and classifying several regions of the city into low, medium, and high accident-prone areas.

Section II imparts a succinct idea of similar research that is conducted using same or different algorithms. Section III focuses on machine learning (ML) and k-means algorithm. Section IV explains the proposed system by means of a block diagram. Section V explains about the results that are obtained from the proposed system. Section VI concludes the study with scope for future.

II. RELATED WORK

In 2014, Theofilatos and Yannis explicated on the upshots of rush-hour traffic and weather conditions [3]. No clear evidences were found, but some fewer patterns were observable. Road accident non-linearity was observed with the rate of accidents but still some observed linearity. To say about weather conditions and their effect on accidents, factors, as mentioned later on, played a major role but not always. Factors such as visibility, wind speed, road type, temperature did not have a direct impact. According to their study, the use of real-time data made it easy to identify safety impacts on area and weather conditions. Moreover, their paper mentions the use of systematic real-time data.

Meshram and Ghods [4] were of the opinion that speed was one of the reasons that led to road accidents as overtaking between vehicles was observed. Their study clarified over speeding and two-lane road accident issues at rural places where there is no proper provision made to follow a lane. The change in time to collison was due to capacity and load, which led to minimum decrease from 500 to 800 vph and then a slight increase. Their study indicated that the maximum confrontational jeopardy was faced in mid-volume regions.

Kim *et al.* [5] considered a log-linear model in order to simplify the responsibility of a motorist's physiognomies and behavior in a contributory order producing brutal wounds. The authors determined that the use of liquor or pills and absence of strap usage significantly increased the probabilities of tremendously unsympathetic bangs and hurts.

Mohamed and Hassan [6] used Fatality Analysis Reporting System database to scrutinize the consequences of the intensifying magnitude of light truck vehicle (LTV) enrolments on deadly viewpoint accident drifts in U.S.A. In addition, they examined the total number of annual fatalities that were caused due to angle collisions and collision configuration. Time series modeling findings indicated that mortalities in viewpoint accidents shall rise within the

succeeding ten years, thereby affecting the projected general increase of LTV proportion in traffic.

Bedard *et al.*'s [7] use of multivariate logistic regression ascertained autonomous driver influence, bang, and automobile features to motorists' casualty danger. They were of the opinion that increasing the use of seat belts, decreasing speediness, and lessening the count and sternness of motorist-side influences may perhaps avert death toll.

Evanco [8] performed a multivariate population-based arithmetical examination to establish a connexion concerning death rates and mishap announcement instances. The investigation showed that a mishap's warning period is an imperative element of the count of mortalities for misfortunes on countryside streets.

Ossiander and Cummings [9] applied Poisson regression to evaluate corelation involving deadly smash speed and maximum speed upsurge. The authors discovered that top speed rise was correlated with a great deadly smash proportion, which resulted in additional demises on expressways in Washington. Furthermore, they reviewed the connection amongst motorists' oldness, sex, automobile bulk, and steering pace extent with death toll.

III. Machine Learning & K-means Algorithm

k-means was coined originally by James MacQueen in 1967. Altthough the term and its idea came from Hugo Steinhaus in 1956, the well-known procedure was fundamentally proposed by Stuart Lloyd in 1957, and it was unpublished as a bulletin article until 1982.

ML can be differentiated as supervised, unsupervised, and reinforcement learning. Supervised learning contains classified data having labels. When supervised data is given to an algorithm, it can provide and guess results easily.

k-means is an arrangement of vector quantization that targets to divide *n* instances into *k* groups wherein each instance is a part of a cluster with a closest average functioning as an architype for the cluster.

It is popular for cluster analysis. *k*-means curtails cluster variances but not regular Euclidean distances, which would be a difficult Fermat–Weber problem, i.e., mean enhances squared errors, whereas only a geometric median decreases Euclidean distances. For instance, better Euclidean solutions can be determined using *k*-medians and *k*-medoids.

A. k-means Clustering

It uses an iterative refinement technique.

Given a preliminary set of *k* means m_1, m_2,…, m_k, *k*-means proceeds by switching between two stages.

Stage 1:

Assign each observation to a cluster with the nearest mean—that with the least squared Euclidean distance.

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\}, \tag{1}$$

where each x_p is assigned to exactly one $S^{(t)}$ even if it could be assigned to two or more of them.

Stage 2:

Recalculate mean for observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{3}$$

B. Procedure for k-means Clustering Algorithm

Assume *x* to be a set of data points and *v* to be a set of centers.

1) Choose *k* random points by calculating Elbow point where best *k* value can be determined.
2) Calculate distance between each datum and cluster centers.
3) Group data points into clusters whose distance from a cluster center is minimum of all cluster centers.
4) Acquire a new cluster center using the following equation:

$$V_i = \frac{1}{C_i} \sum_{j=1}^{Ci} x_i \tag{3}$$

where c_i represents the number of data points in an i^{th} cluster.

5) Again, calculate the distance between each data point and new obtained cluster centers.
6) If no datum was reassigned, then stop. Otherwise repeat from Step 3 [10].

k-means converges when assignments no longer change. The algorithm does not assure to acquire an optimum solution.

IV. Proposed System

Figure 1 shows the block diagram of the proposed system for road accident analysis and hotspot prediction.

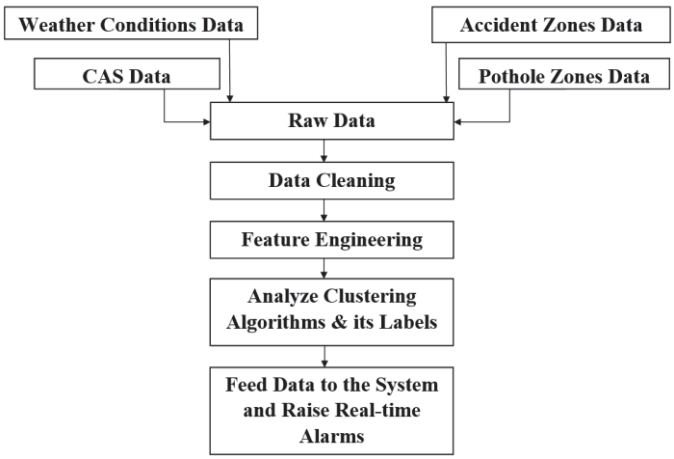

Fig. 1. Block diagram of the proposed system for road accident analysis and hotspot prediction

A. Data Collection

Data was collected from official weather forecast and accident reports from their official website [11]. Government data was used to acquire weather and road conditions to execute the proposed system.

Dataset works with the help of the following attributes:

- *Latitude, longitude, and location name*
- *Speed*
- *Time*
- *Alarm type*
- *Weather*
- *Peak hour timings*
- *Potholes*
- *Accident-prone areas*

Records of the abovementioned attributes were collected on a daily basis for a year.

B. Preprocessing

Sorting data and making it usable for the proposed system is a major task for majority of the models. For our model, we sorted data into 8 attributes as per time and date in a month-wise manner.

Data preprocessing comprises the following steps:

1) **Combination:** Data is integrated from various sources.

2) **Conversion:** *k*-means algorithm is applied to sorted data to obtain expected outputs.

3) **Trimming:** Only needed data is taken.

4) **Sorting:** Inconsistent and noisy data is removed, which in turn makes data highly efficient.

After successful preprocessing for a given dataset takes place, it is converted into a desirable road accident dataset, which is used for training purpose. Table I lists several parameters such as time, temperature, weather, wind, direction, humidity, barometer, and visibility that represent climatic conditions that affect road accidents in one way or the other. This data is used as preprocessed data in the proposed system for analysis.

TABLE I. PREPROCESSED DATA

Time	Temp	Weather	Wind	Direction	Humidity	Barometer	Visibility
14:30	26 `C	Scattered Clouds	19 km/h	↗	70%	1009 mbar	8 km
14:00	26 `C	Scattered Clouds	15 km/h	→	70%	1010 mbar	8 km
13:30	26 `C	Scattered Clouds	19 km/h	↗	70%	1010 mbar	8 km
13:00	26 `C	Scattered Clouds	22 km/h	↗	65%	1011 mbar	8 km
12:30	27 `C	Scattered Clouds	20 km/h	↗	62%	1011 mbar	8 km
12:00	25 `C	Scattered Clouds	17 km/h	→	69%	1012 mbar	8 km

C. Training and Testing

The proposed system is trained under various circumstances to make it efficient. 60%–40% is used to train and test the system, respectively, so that no discrepancy occurs. In addition, the system is adaptable to all changes that can be suggested for further improvement (i.e., if data needs to be preprocessed repeatedly, then it can be preprocessed and merged with an old dataset for further use). Table II lists several locations and recorded time for various aspects of preprocessed data after sorting. It includes speed conditions while accidents occur, i.e., overspeed, low speed, etc.

TABLE II. DATA AFTER SORTING

ceCode_device	Code_location_l	ode_location_lo	e_location_	pde_pyId_al	de_py	Code_time_recordedTime_
8.64504E+14	12.9845953	77.74408722	Kadugodi	PCW	32	2018-02-01T01:48:59.000Z
8.64504E+14	12.9845953	77.74408722	Kadugodi	PCW	32	2018-02-01T01:48:59.000Z
8.64504E+14	12.98723316	77.74111938	Garudachar	FCW	41	2018-02-01T01:50:00.000Z
8.64504E+14	12.98723316	77.74111938	Garudachar	FCW	41	2018-02-01T01:50:00.000Z
8.64504E+14	12.98750305	77.74005127	Hudi	Overspeed	37	2018-02-01T01:50:11.000Z
8.64504E+14	12.98750305	77.74005127	Hudi	Overspeed	37	2018-02-01T01:50:11.000Z
8.64504E+14	12.98752308	77.73670197	Kadugodi	HMW	32	2018-02-01T01:50:50.000Z

D. Working

Data collected from various sources is provided as an input to *k*-means algorithm to execute the model.

1. We take the hour during which an event occurred and map it to different categories to study if a vehicle was observed during peak hours, early hours, regular hours, and so on.
2. We follow similar strategy here to get the hour to equate it to the hour from previous data frame of events recorded by the collision avoidance system (CAS).
3. We map the areas that we have in our dataset to a nearest accident-prone zone marked in the map based on the number of fatalities reported by the traffic police. We performed this task manually.
4. Temperature, condition, and visibility columns have the same number of records since we replaced them with the mod values of each of those columns.
5. We create a flag for low and high visibility. Anything >10 would be equated to high visibility, whereas the other cases would be low visibility.
6. We now have two datasets, i.e., one that has all the data in words for understandability and the other in numerical format to fit to our prediction model.
7. Since data is unlabeled, we opt for unsupervised learning approach using a clustering technique. We tried different clustering algorithms, but we present the ones (i.e., *k*-means clustering) that offer the most suitable labels as per our intuition.
8. Data is then passed through *k*-means clustering, and we get a segregated data where accidents occur in the form of data as well as graph. Through this, we can take several major accidents to reduce accidents in a particular area taking into consideration the cause for the same.

E. Output

After successful testing, the model could be made available to local authorities for use. Figure 2 shows hotspot regions pinned in red where accidents usually occur.

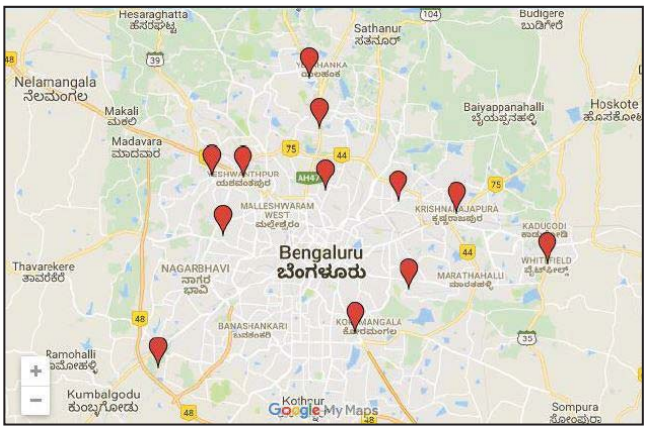

Fig. 2. Accident-prone areas

V. Simulation Results

We can see locations with high accident-prone areas and high potholes are more vulnerable to accidents no matter what the conditions are during the morning hours, as in Figure 5. It can be viewed from Figure 4 that in medium accident-prone zones and pothole severities, over speeding with low visibility during early hours of data can have high chances of an accident as per a given area and an alarm type. In low accident-prone areas, i.e., Figure 3, accidents were because of human error and atmospheric conditions only.

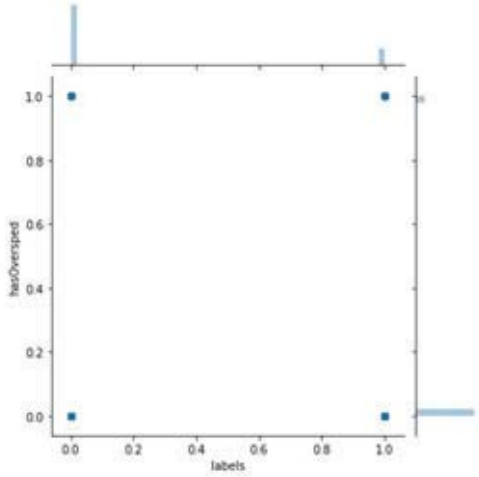

Fig. 3. Low accident-prone areas

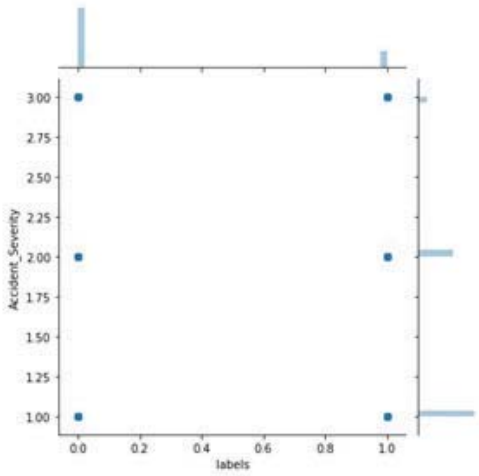

Fig. 4. Medium accident-prone areas

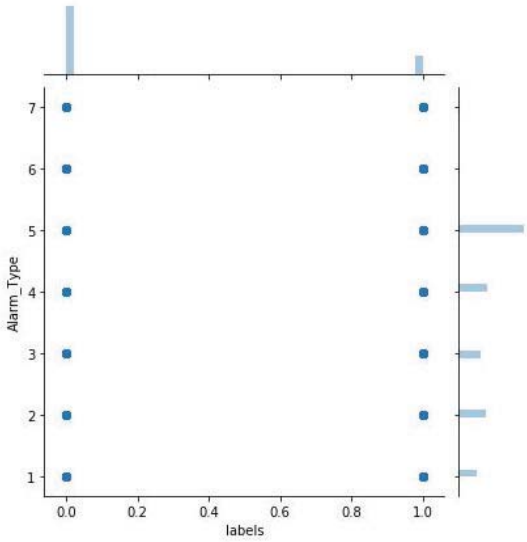

Fig. 5. High accident-prone areas

VI. Conclusion and Future Scope

ML has been helping us to solve many problems in our day-to-day life. It has helped to analyze data provided and provide appropriate solutions to problems that occur. Due to which, the study uses *k*-means algorithm. This study aimed to determine the reason behind the major cause of the increase in the number of road accidents happening around. For the past few years, it was noticed that the rate of road accidents had been increasing at an alarming rate due to various factors like drunk driving, problems related to climate, human error, etc. Considering this, the study of road accidents can play an important role to prevent road accidents that would have happened in the near future. From the data collected, we found a few major reasons for the cause of road accidents and the relationship between them. The data, when represented meaningfully, can help people to avoid places or take caution while going through such areas as per the severity of a location in terms of accidents. Moreover, the study would help road transport systems to improve their work efficiency and help in controlling the number of casualties and loss of personal property taking place due to road accidents.

References

[1] Road Traffic Injuries, "World Health Organization (WHO)", [Online], Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries [Accessed on September 30, 2020].

[2] Ministry of Road Transport and Highways, [Online], Available: https://morth.nic.in/ [Accessed on September 25, 2020].

[3] A. Theofilatos and G. Yannis, "A review of the effect of traffic and weather characteristics on road safety", *Accident Analysis & Prevention*, vol. 72, pp. 244–256, July 2014.

[4] K. Meshram and H.S. Goliya, "Accident analysis on national highway-3 between Indore to Dhamnod," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 2, no, 7, pp. 57–59, July 2013.

[5] K. Kim, L. Nitz, J. Richardson, and L. Li, "Personal and behavioral predictors of automobile crash and injury severity," *Accident Analysis & Prevention*, vol. 27, no. 4, pp. 469–481, 1995.

[6] M. Abdel-Aty and H. Abdelwahab, "Analysis and prediction of traffic fatalities resulting from angle collisions including the effect of vehicles' configuration and compatibility," *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 457–469, 2004.

[7] M. Bedard, G.H. Guyatt, M. J. Stones, and J. P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics

to driver fatalities," *Accident Analysis & Prevention*, vol. 34, no. 6, pp. 717–727, 2002.

[8] W.M. Evanco, "The potential impact of rural mayday systems on vehicular crash fatalities," *Accident Analysis & Prevention*, vol. 31, no. 5, pp. 455–462, 1999.

[9] E.M. Ossiander and P. Cummings, "Freeway speed limits and traffic fatalities in Washington State," *Accident Analysis & Prevention*, vol. 34, no. 1, pp. 13–18, 2002.

[10] N. Sridevi, M.V. Keerthana, M.V. Pal, T.R. Nikshitha, and P. Jyothi, "Road accident analysis using machine learning," *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 5, pp. 859–861, May 2020.

[11] Weather in Bangalore, Karnataka, India, [Online], Available: https://www.timeanddate.com/weather/india/bangalore [Accessed on September 30, 2020].