AN EMPIRICAL STUDY ROAD ACCIDENT IN INDIA WITH DATA MINING APPROACH

LehaSri Boyapati[1],Raj Kamla Kuchipudi[2], GOPISETTI Gurukesava Dasu[3]

[1]PG Scholar, Dept. Of CSE, ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY, WEST GODAVARI, A.P.

[2]Assistant Professor, Dept. Of CSE, ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY, WEST GODAVARI,A.P.

[3]Professor And HOD Dept. Of CSE, ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY, WEST GODAVARI,A.P.

Abstract:

The accident is an impromptu episode that prompts injury to individuals, harm to a plant, apparatus or some different misfortune. The objective of this paper is an analysis of road accidents at a nation level and statewide in India. The analysis shows that accidental fatalities and wounds change as per age, sexual orientation, month and time. Analysis of road accidents assumes a significant job in the transportation framework. Road traffic wounds and fatalities are normal in nature it is unrealistic to anticipate balanced connections among the security gauges in road accidents, wounds, and fatalities. Road security is a significant worry for both the national and worldwide levels. Data mining instruments and strategies are utilized to anticipate accident-inclined areas. For at regular intervals, one passing is happened because of road accidents in India. The pivotal thing is an analysis of road accident data is its heterogeneousness. The connection between road surface conditions, road type, seriousness, light conditions, and so on are examined.

Keywords: Cluster Analysis (K-Means), Classification (Decision Tree), Cross Validation, Road Accidents, Data Mining

I. INTRODUCTION

The stark truth of the developing accident rate works up the requirement for a critical spike in road wellbeing everywhere throughout the world. The duty doesn't generally come upon the driver. The different elements that are ascribed to prompting such an accident can run anyplace from vehicular imperfection and climate conditions to transportation conditions. According to the report by WHO, the defenseless road clients (motorcyclists, cyclists and people on foot) represent half of the world's accident drove deadly ities; with bike inhabitants representing about 31% of passings. It proceeds to express that grown-ups represent 59% of the all out fatalities all inclusive [17]. Right now, study the states and the association regions of India against the contributing causes and the instructive background of the driver to reach effective inferences so as to encourage road security in the nation. We are centered around taking the guide of clustering to amass comparative objects of this dataset so as to gather districts based on weakness [16]. The clusters so framed are marked to be additionally arranged utilizing a decision tree to give the district savvy predominant reason [3].

Road transportation is a prevailing vehicle in India, as far as traffic offer and commitment to the national economy. Road accidents can be characterized as "accident happened on a way or street, results show that at least one individuals are killed or harmed, one, one or more vehicles are included [1]. In this way crashes among vehicles; among vehicles and people on foot; among vehicles and creatures; among vehicle and land snags."

The lack of road networks prompts road accidents and road crash fatalities [2]. A ton of vehicles going on the roads consistently and accidents might be going on whenever anyplace a few accidents may prompt wounds some lead to passings. These days, road accident wounds are one of the most significant reasons for death, inabilities and hospitalization in India. Data mining applies various systems and calculations to decide the connections among the traits in the data set. The serious issue is the heterogeneity, in this way heterogeneity must be considered during the analysis of the data in any case, some connection between the data may stay covered up [3]. The thickness of auto collisions in India is the most elevated in the world. Right now, states and the association domains of India and to know restricted causes and instructive background of the driver so as to make conceivable road security in the nation [4]. Delhi and Chennai register that numerous quantities of accidents than different states in India. In Indian roads, the significant accident-inclined time is at the hour of evening and night. World Health Organization (WHO) detect that the vast majority of the car accidents happened because of driver over speed, drunk and driving, languor, and not wear caps and safety belts

II. RELATED WORK

Different investigations turning on the traditional factual techniques have been completed so as to dissect

road accidents. The measurable methodology used to make an accident forecast model neglects to consider the vulnerability factor related with it. The relapse up-and-comer models (Negative Binomial Model and Poisson Model) utilized in arriving at a solid resolution, go into forming the model space. Of these applicant models, any one is chosen to anticipate the recurrence of accidents. Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC) and Akaike Information Criterion (AIC) are a portion of the criteria used to choose the most fit model that effectively takes into thought the back model probabilities. Logical (free) factors' choice is the most fundamental piece of building an accident analysis and forecast model. The BMA approach is utilized utilizing the a far cry calculation to execute the relapse model. Fake Neural Networks (ANN) have been effective in beating the vulnerability statement presented by the regular factual methodology, yet lead to overfit of the data.

Data Mining Techniques help defeat the previously mentioned weaknesses. The thought is to decide the prevailing component perpetrating the different conditions of India dependent on the dataset that we have; and to sketch out all the potential results off the different elements that legitimately or in a roundabout way had a section in molding the dataset. The data mining strategies that we intend to use for the equivalent are clustering (K-Means) and classification (Decision Tree). Cluster analysis has been utilized in determining the person on foot fatalities [4] [7] and decision trees have been utilized to foresee reasons for accidents [11] [12]. The data is pre-handled and made fit to be exposed to the data mining procedures [9].

Our point is to altogether break down the unstructured and unlabelled dataset utilizing solo data mining systems.

The focal point of solo data mining methods isn't on foreordained trait. Likewise, no objective worth is anticipated. Or maybe, unaided data mining looks shrouded structure and connection among data [10].

This analysis is required to assist us with determining the accident inclined locales, the predominant factor delivering the different conditions of India and to give a somewhat lucid help with sketching out every single imaginable relationship between the different components that straightforwardly or in a roundabout

According to the appraisals made by the World Health Organization (WHO) in the Global Safety Report, India represents in excess of 200,000 fatalities [17]. Different examinations have been done to decide the expansion experienced in the field.

The roads of India haven't lessened their commitment in the auto collision fatalities. The accident pace of India has been on an expansion since the time the beginning of the century. Data mining investigations can help distinguish the significant causes and help the vehicle experts in improving security prerequisites [6].

As indicated by an evaluation taken in 2013, the quantity of accidents in India arrived at a protruding '1, 37,000'. It was surveyed that one demise happened at regular intervals [15]. The issues of road accidents is exceptionally delicate and generally talked about in Indian media[18][19]. In the year 2014, the road networks of India represented 63% of complete road accidents and saw a 3% raise from 2013, with consistently seeing 16 fatalities [13]. In 2015, the rate saw a hike of 5% and the quantity of passings every hour spiked to 400[14].

III. DATA SET COLLECTION AND PRE-PROCESSING

The dataset gives an intricate record on the road accidents that unfolded during that time 2012 in the numerous conditions of India, and is exposed to the different data mining systems so as to draw rather convincing data. The dataset is chosen from data.gov.in. The data incorporates from all the association domains and states and broke down for 58 characteristics like all out number of accidents, number of individuals killed and number of individuals harmed because of different variables like liquor, speeding, driver's issue, kind of vehicles and so on [1].

way have a section in molding the dataset; to be specific, the instructive background of the drivers engaged with the accidents, the sort of vehicle that drag cause, the different causes that irregularly contributed in the procedures through the span of a year and the numerous lives that missed the mark regarding their time.

Breaks data down, making it fathomable Makes it simple to recover esteems from the dataset

Quick recovery improves the productivity of the calculations being done.

I. PROPOSAL WORK

K-MEANS Clustering is utilized to amass comparative item off of the heterogeneous data.

According to this calculation, an article can be assigned to just one cluster. Euclidean separation is the measure used to characterize the centroid of a cluster. K is the quantity of clusters and is generally given a little whole number worth (1, 2, 3…). 0 K points are then picked haphazardly ideally the underlying ones-which speak to the centroids of k clusters with no individuals and are set to the cluster with the centroid closest to it [2].

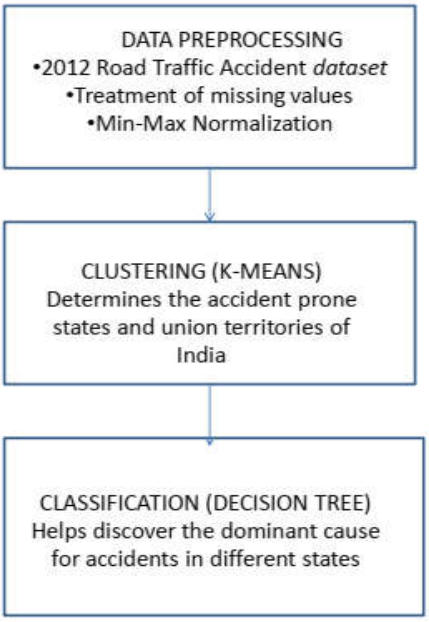

Fig. 1. The data flow of the analysis.

***ALGORITHM* (K-MEANS)–THE CENTRE OF EACH CLUSTER ISGIVEN BY THE MEAN VALUE OF THE ITEMS IN THE CLUSTER.**

Most importantly, it is ensured that the dataset doesn't have any missing qualities [9]. It is then standardized utilizing the min-max standardization strategy.

$$V' = \frac{[V - \min(a)]\,[newMax(a) - newMin(a)] + newMin(a)}{(\max(a) - \min(a)]}$$

Where V is the main example of the field, Min(a) compares to the base field esteem, Max(a) relates to the greatest field esteem

newMax is 1

newMin is 0

Thusly,

The adequacy of the mining algorithm is essentially improved as

Information -

k: the number of clusters

D: Dataset

n: number of things
Yield A SET OF K CLUSTERS
Strategy

(1) Randomly select 'k' things as the underlying cluster communities from the dataset 'D'

(2)Items are relegated or reassigned to the clusters dependent on the mean estimation of the things in the cluster.

(3)The mean estimation of the things in each cluster is refreshed

(4) Repeat stages 2 and 3 until there is no change.

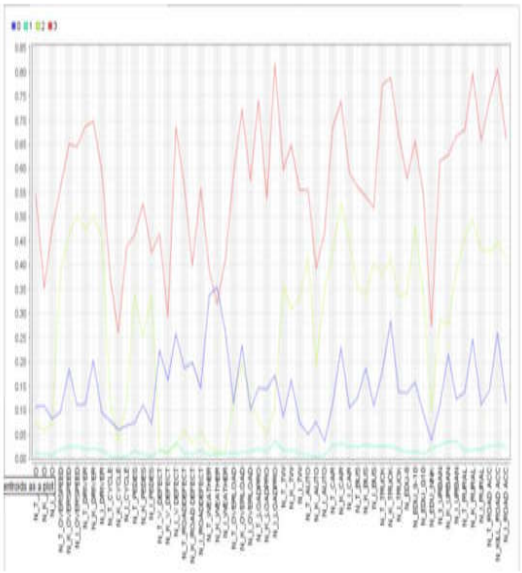

Fig. 2. A centroid plot view of the clusters formed. X-axis gives the names if the attributes and Y-axis gives the normalised values.

The algorithm is applied for various estimations of k=2, 3… 7. The nature of clusters is then assessed utilizing Davies-Bouldin Index. The Davies-Bouldin Index related with the cluster model comparing to k=4 ended up being the most noteworthy.

A concise review of the clusters framed is given underneath:

Cluster 0: This cluster ended up being comprised of 7 states in particular Bihar, Chhattisgarh, Haryana, Jharkhand, Odisha, Punjab and West Bengal. An examination of the cluster drove us to set the "medium" name to these states as these were the states with a normal populace arranged in the northern fields, according to the dataset.

Cluster 1: This cluster ended up being comprised of an aggregate of 19 things including 12 states and 7 Union Territories of India, specifically Arunachal Pradesh, Assam, Goa, Himachal Pradesh, Jammu and Kashmir, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, Tripura and Uttrakhand, Andaman and Nicobar Islands, Chandigarh, Dadar and Nagar Haveli, Daman and Diu, Delhi, Lakshadweep and Puducherry individually. An investigation of this cluster dependent on topography and populace uncovered that all the states gathered were the least populated in the nation and the vast majority of these states had an uneven landscape. These two elements brought about low accident rate in these states. Thusly, the states right now marked "low".

Cluster 2: This cluster was found to contain 3 states in particular Gujarat, Rajasthan and Kerala. Gujarat and Rajasthan are desert states though Kerala is a beach front state. Both these states have a high populace regardless of the less positive topography. Additionally, the availability by means of the roads is acceptable. In this way, the states right now alloted a "High" mark.

Cluster 3: This cluster was seen as contained 6 states: Andhra Pradesh, Karnataka, Madhya Pradesh, Maharashtra, Tamil Nadu, and Uttar Pradesh. They are thriving states with the most noteworthy populace rate in the nation. The territory in certain areas is plain and others, a level with a not really very much built network of roads. This causes the accident-rate to see a swell. Accordingly, an 'Exceptionally high' mark is alloted to the states present right now.

V. CLASSIFICATION MODEL

Decision tree follows the tree structure and comprises of a root hub, transitional hubs and lead hubs. Decision is contained by every hub and dependent on this decision, the tree advances. Name, worth and activity characterize the totally unrelated spaces made in a tree [2]. These trees are created by means of recursive apportioning. The exhibition of J48, Multilayer Perceptron, Naïve Bayes Updatable, and Bayes Net was assessed to examine road accidents as far as exactness and time. The productivity of J48 is discovered to be better than that of Naïve Bayes [5] [8].

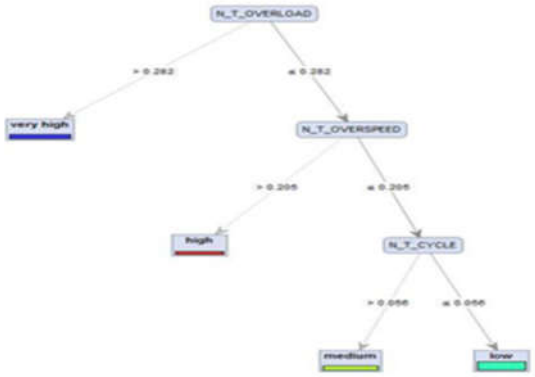

Fig. 3. Decision Tree

Classification techniques are utilized to distinguish the primary driver of accidents. The data set is ordered dependent on the names relegated to it following cluster analysis. Increase Ratio criteria are set to manufacture the decision tree [6].

The increase proportion is:

Increase Ratio(A) = Gain(A)/Split Info(A) :

In the decision tree produced, the most significant variable to part on is the all out number of accidents due to over-burdening/congestion of the vehicle.

Cross-validation of the model so framed gives the exactness at 72.67%.

Table View Plot View

accuracy: 72.67% +/- 16.65% (mikro: 71.43%)

	true very high	true low	true medium	true high	class precision
pred. very high	1	0	0	1	50.00%
pred. low	2	17	4	0	73.91%
pred. medium	0	1	2	0	66.67%
pred. high	0	0	0	0	0.00%
class recall	33.33%	94.44%	33.33%	0.00%	

Fig. 4.A confusion matrix after cross validation.

Algorithm (Build_Decision_Tree) – Build a Decision Tree from the preparation tuples from the data-segment D

Info -

x 'Data Partition' D-a lot of 'preparing tuples' and their relating class-marks

x 'Attribute List'- set of up-and-comer qualities

x 'Attribute determination strategy'- 'parting basis' is resolved utilizing it

Yield Decision tree

Technique – the philosophy decided on structuring the decision tree is given underneath

(1) Create a hub 'n';

(2) If the tuples in D are in a similar class, C at that point return 'n' as a leaf-hub marked C;

(3) If the 'property list' is unfilled, at that point

return 'n' as a leaf-hub with mark of the lion's share class in D;

(4) 'Attribute determination technique' is applied to acquire the most reasonable 'parting paradigm';

(5) Use 'parting foundation' to name hub 'n';

(6) If 'parting trait' is discrete-esteemed and multiday parts are permitted, at that point

expel 'parting characteristic' from the 'property list';

(7) for each result j of 'parting foundation'/separate the tuples and manufacture sub-trees for every division

(8) letDj speak to the arrangement of data tuples in D satisfying the result j/division

in the event that

Dj is seen as vacant, at that point

append the leaf with the name of the greater part class in D to hub 'n';
else

append the hub returned by ' Build_Decision_Tree' to hub 'n';

endfor

(9) Return hub 'n';

CONCLUSION

Right now perceived how cluster analysis encourages us to decide the accident inclined states and domains of India. These clusters are named to be ordered with the assistance of decision tree to finish up the predominant factor, backing the accidents.The states with the complete number of accidents due to over-burdening/packing higher than 5704 breed accidents thickly and accordingly have a genuinely high accident rate, though the states with a less figure than 5704 have accidents due to over speeding as the prevailing element. In future, we can work on building an accident expectation model that applies the Bayesian Model Averaging to a Regression Model to break down the accident dataset and additionally decide the accident recurrence. Furthermore, the investigation of dataset as far as time of event of accidents should be possible to decide the most-defenseless time and the most-powerless period of the year for road accidents and furthermore the casualty rate related with these accidents.

REFERENCES

[1] Open Government Data (OGD) Platform India [Online]. Available 2016: https://data.gov.in/catalogs/sector/Transport-9383

[2] Jiawei Han and MichelineKamber, *Data Mining: Concepts andTechniques*, 2nd ed., Morgan Kaufmann Publishers, 2006.

[3] HümeyraBolakar and AhmetTortum, “Clustering of Districts in Erzurum by Number of Injury”, Journal of Traffic and Logistics Engineering vol. 3, pp. 125–128, Dec. 2015

[4] Carlo Giacomo Prato, Victoria Gitelman and ShlomoBekhor, “Mapping patterns of pedestrian fatal accidents in Israel”, Accident Analysis and Prevention, pp. 54–62,Jan. 2012

[5] V. Vaithiyanathan, K. Rajeswari, KapilTajane and Rahul Pitale, “Comparison of different classification techniques using different datasets”, International Journal of Advances in Engineering & Technology, vol. 6, pp. 764–768, May. 2013.

[6] NainaMahajan and Bikram Pal Kaur (PhD), “Analysis of Factors of Road Traffic Accidents using Enhanced Decision Tree Algorithm”, International Journal of Computer Applications (0975 – 8887), vol. 135, pp. 1–3, Feb. 2016.

[7] Svetlana Bačkalić, BoškoMatović and DraganJovanović, “Identification of hotspots road locations of traffic accidents with pedestrian in urban areas”, International Co-operation on Theories and Concept in Traffic Safety, Dec. 2014

[8] S.Vigneswaran, A.Arun Joseph and E.Rajamanickam, “Efficient Analysis of Traffic Accident using Mining Techniques”, International Journal of Software and Hardware research in engineering, vol. 2, pp. 110–118, March 2014.

[9] Sachin Kumar and DurgaToshniwal, “A data mining framework to analyze road accident data”, Journal of Big Data (Springeropen Journal), Oct. 2015.

[10] A. Priyanka and K. Sathiyakumari, “A comparative study of classification algorithm

using accident data", International Journal of Computer Science & Engineering Technology
[11] (IJCSET). vol. 5, pp. 1018–1023, Oct. 2014.
[12] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Hong Kong, vol. 1, March 18 - 20, 2009,
[13] The Times of India website. [Online]. Available 2015: http://timesofindia.indiatimes.com/india/16-deaths-every-hour-Indian-roads-claim-the-maximum-number-of-lives-in-2014/articleshow/48128946.cms.
[14] The Times of India website. [Online]. Available 2015:http://timesofindia.indiatimes.com/india/400-road-deaths-per-day-in-India-up-5-to-1-46-lakh-in-2015/articleshow/51919213.cms
[15] NDTV news portal. [Online] Available 2015: http://sites.ndtv.com/roadsafety/important-feature-to-you-in-your-car/
[16] M.N. Postorino and G.M.L. Sarne, "Cluster analysis for road accident investigation", Urban Transport VIII, vol. 60, pp.- 785-794, 2002
[17] Global Status Report on Road Safety: supporting a decade ofaction, Geneva, World Health Organization, 2013.

About Authors

KUCHIPUDI. RAJKAMAL
B.Tech, M.Tech, (CSE)
ASST PROFESSOR ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY, AP, INDIA.

Dr. GOPISETTI GURU KESAVA DASU PROFESSOR and HOD DEPARTMENT OF CSE ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY, AP, INDIA.
BE(CSE), ME(CSE), PHD(CSE)

LehaSri Boyapati
M.tech(CSE) ELURU COLLEGE OF ENGINEERING AND TECHNOLOGY,
WEST GODAVARI,AP, INDIA