International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Development of Naïve Method to Analyse the Road Accidents Based on Data Mining Techniques

Sonam Singh, Shailesh Singh
Sheat College of Engineering, Varanasi, India

Article Info
Volume 9, Issue 1
Page Number : 357-366

Publication Issue
January-February-2022

Article History
Accepted : 05 Jan 2022
Published : 20 Jan 2022

ABSTRACT

"Road Accident is an all-inclusive calamity with the continually growing trend. In India, according to the Indian road safety campaign every minute, there is a traffic accident and about 17 people die each hour in road accidents. There are several types of car accidents such as rear-end, head-on, and rollover accidents. India, according to the Indian road safety campaign every minute, there is a traffic accident and about 17 people die each hour in road accidents. There are several types of car accidents such as rear-end, head-on, and rollover accidents. The state-recorded police reports or FIRs are the records that provide information regarding the accidents. The event may be self-reported by the individuals or documented by the state police. Using Apriori and Nave Bayesian approaches, recurrent patterns of road accidents are predicted in this research. The government or non-profit organizations might use this pattern to enhance road safety and implement preventative measures in high-accident areas. From 2020 to 2021, a total of 11,574 accidents happened on the roads in the Dehradun district. Based on the variables of accident type, road type, lightning on road, and road feature, K modes clustering found six clusters (C1–C6). Each cluster and the EDS have been used to construct rules using association rule mining. Rules with high lift values are used in the study. Using the rules for each cluster, it is possible to learn about the causes of accidents in that cluster. When compared to EDS-generated rules, this comparison reveals that EDS-generated rules do not give relevant information that may be linked to an accident. If additional features linked with an accident are accessible, more information may be discovered. We also did monthly and hourly trend analyses of all clusters and EDS to reinforce our technique. According to trends, clustering before an analysis helps us locate better and more helpful outcomes that we wouldn't otherwise be able to find."

Keywords - Hierarchical Clustering, K-modes clustering, Self-Organizing Map, Three-and-a-half inches of Birch, Invite Vector Machine Support, Latent Class Clustering, Intentional Bayes

I. INTRODUCTION

A. Overview

The proposed framework may be implemented in the form of an ongoing software program. This topic may be implemented in the government sector. The general public may make use of the current framework to learn about the many types of mishaps that might occur in a certain region or city. The goal of the new piece was to discover any hidden connections and linkages between various elements showing street accidents with fatal consequences. In India, street security may be improved by collaborating with all of the major stakeholders, including the police, the state, and the local government. Precautionary steps will be taken based on a study of the current system, which uses record information and physically dissects it to find out how many accidents there have been. In addition, a slew of gadgets and programs are available to monitor traffic accidents; however, they just collect data from various sources and do not conduct a full investigation. Currently, everything is done manually, which is time-consuming, expensive, and leaves a lot of information hidden. As a result, it is less effective. For the proposed research, we are looking for correlations and links between numerous elements that may have previously gone unnoticed, as well as the prevalence of traffic accidents and their potentially lethal outcomes. Similar information extraction may assist enhance road safety when done in close collaboration with all of the most significant players, such as the police department, state government, and federal government. First, the current circumstances and our motivation for analyzing data on road traffic accidents are explained; next, the survey data set is presented and pre-processing operations are performed to prepare mining data; and last, the chosen methodologies and evaluations are used.

Long-distance riders and travelers may use the data mined by data miners to make informed choices about their journey by using the tools and procedures developed for data mining to the controlled data.

B. Databases

This study's test location is a section of the G60 Freeway in Shanghai, India. The road portion is 48.7 km long and has 6–10 lanes (3–5 lanes in each direction).

C. Objective

The general goal of this thesis is to acquire accuracy and discover the elements behind crashes or accidents that might be useful in reducing the accident ratio shortly and could be useful in saving many lives, deteriorating wealth destruction, and many other things. The following section provides an overview of research publications relevant to this thesis.

I. LITERATURE REVIEW

A. Overview

The field of transportation accident investigation is very important (Kumar and Toshniwal, 2016a). Several studies have used statistical methods and data mining techniques (Savolainen et al. 2010; Karlaftis and Tarko, 1998; Jones et al. 1991, Poch and Mannering 1996, Maher and Summersill, 1996) to analyze traffic crash data, and to establish relationships between accident attributes and road accident severity. The findings of this research are very valuable since they shed light on the many factors that contribute to car accidents. To combat the high accident rates in the study region [37-41, 32, 7, 19, 42-43, 79-80], being aware of these accident variables is unquestionably beneficial.

B. Factors Responsible For Accident

An investigation conducted by Peng and Boyle [59] aimed to gather information on the influence of

commercial driving considerations on the severity of ROR, single-vehicle collisions. Using safety belts lowered the risk of ROR collisions, according to one study. ROR crashes were made more likely by distracted and careless driving. Fatigue, laziness, and speeding all contributed to an increase in the likelihood of serious injury or death in ROR accidents. Drivers of commercial motor vehicles (CMVs) in good condition had a decreased risk of serious injury or death in rear-end (ROR) collisions. It was 3.8 times more likely for a ROR accident to be injurious and fatal if it happened on provincial roads or dry streets. There were no additional criteria that were considered important. Few driver factors, including exhaustion and laziness, speed, diversion, distractedness, and the use of a safety belt in a ROR incident, were shown to have a significant impact on the likelihood of a ROR crash being fatal.

C. Traditional Statistical Approach For Accident Analysis

Not all statistical procedures are data mining methodologies. Statistical techniques or "statistics" are not. They were in use long before the phrase "data mining" was coined to describe their use in corporate contexts. Statistical methods, on the other hand, are driven by data and used to discover patterns and develop prediction models. In road safety studies, statistical methods have also played an essential role. Researchers Karlaftis and Tarko (1998) examined the relationship between rider age and accident frequency. The data on traffic accidents were analyzed using negative binomial models and cluster analysis. Data from road accidents has been the subject of many significant statistical analyses, including those by Savolaiinen et al. (2010); Karlaftis and Tarko (1998); Jones et al. (1991); Poch (1996); as well as Poch and Mannering (1996).

D. Data Mining Approaches For Accident Analysis

It has been claimed that clustering before analysis of traffic and road accident data is highly effective in dealing with the high degree of variability in these datasets, as recommended by Kumar and Toshniwal (2015a). Latent class clustering (LCC) was utilized by Ona et al. (2013) to eliminate data heterogeneity. Using LCC, they found that it may be used to discover many clusters in the data set, as well as varied criteria for determining which clusters should be included in the analysis. In addition, (Kumar and Toshniwal, 2016d) compared accident data from Haridwar, Uttarakhand, India with those from other cities in India. Clustering approaches such as LCC and K-modes (Chaturvedi et al., 2001; Kumar and Toshniwal, 2015b) were utilized in this work to group the data before completing the analysis. In addition, they used the Frequent Pattern (FP) development approach to extract the association rules that explain the accident patterns in each cluster. They concluded that both methods are equally effective in forming clusters and eliminating heterogeneity from the data. There was no evidence that one method was preferable to another, however.

II. PROPOSED METHODOLOGY

A. K-modes clustering

It is the goal of clustering, an unsupervised data mining technique, to group together data items in such a manner that objects inside a group are more similar than those in other clusters. When working with huge sets of numerical data, clustering algorithms like the K-means [67] method are a popular choice. The dataset is divided into k clusters using this method. However, the choice of an effective clustering technique is dependent on the kind and form of data. The primary goal of this research is to identify the location of an accident based on its frequency of occurrence.

B. Self-Organizing Map (SOM)

Self-Organizing Map (SOM) by Teuvo Kohonen [69] gives a visualization of data that assists high dimensional data by decreasing the dimension of data. SOM likewise describes the clustering idea by gathering comparative data together. Thusly one might say that SOM decreases the dimension of data and presentation similitudes among datasets. With SOM,

clustering is executed by having a few units go after the present object. Once the information has been gone into the framework, the network of neurons is prepared by giving data about sources of info. The weight vector of the unit nearest to the present object turns into the triumphant or dynamic unit. Amid the preparation organize, the values for the input factors are progressively balanced trying to safeguard neighborhood connections that exist inside the input dataset. As it gets nearer to the input object, the weights of the triumphant unit are balanced and in addition to its neighbors. At the point when a training set has been forced to the neural systems then their Euclidean separation to conclusive weight, vectors are figured. Presently the neuron weight is around the weight of input data. In this way, this is called the triumphant unit or Best Matching.

C. Hierarchical Clustering

Figure 2 depicts a clustered tree (or "node"), where each segment (or "node") is linked to at least two subsequent segments. Hierarchical clustering After settling and sorting out the pieces into a tree, we have what might be considered an important categorization plot.

On the clustered tree, each node is placed next to another node that has equivalent data. Iteratively connecting each node in a tree until all information in the set can be seen is an effective way to offer users a sneak peek at what they may expect to see once the tree is complete. As soon as you begin the tree-creation process, you have no idea how many clusters you'll end up with.

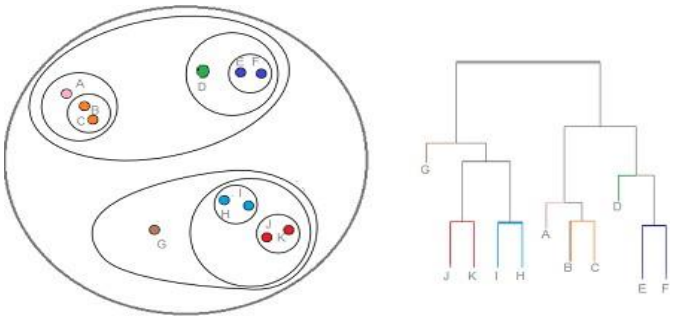

Fig. 3.1 A dendrogram (right) depicting nested hierarchical clusters (left)

D. Latent Class Clustering (LCC)

Using LCC for road and traffic accident data analysis is a common clustering technique. LCC's greatest strength is its versatility, as it can be used to the category, numerical, or mixed attribute data. As an added benefit, the LCC utilizes many cluster selection criteria, such as Akaike Information Criteria (AIC) (Akaike, 1987), Bayesian Information Criteria (BIC) (Raftery, 1986), and CAIC (consistent AIC) (Fraley and Raftery, 1998) to help determine the optimal number of clusters to build.

E. Three-and-a-half inches of Birch

To produce efficient clusters, BIRCH, a well-known dynamic clustering approach, employed hierarchical cluster analysis on a huge dataset. Prior clustering algorithms were limited in their ability to handle huge datasets because of memory constraints. This means that BIRCH is a very efficient algorithm in terms of both CPU and memory use.

F. Invite Vector Machine Support

Analogous to a regression or classification technique, SVM is a supervised learning approach. SVM uses decision planes to describe the boundaries of the decision space. Decision planes are a way of distinguishing between a group of items that are distinct in terms of their class. Using hyperplanes in n-dimensional space, a classifier approach is used to do a classification problem. SVM can handle many categorical and continuous variables and help with classification and regression tasks.

G. Intentional Bayes (Ib)

A classifier based on Bayes' hypothesis with concerns about autonomy across indicators. Due to its simplicity and lack of astounding measure approximation, this approach is very useful for dealing with enormous datasets.

H. DECISION TREE

ID3's upgrade, J48, extends its capabilities. J48's extra components indicate data that has been omitted. J48 is a Java-based open-source implementation of the C4.5 computation in WEKA.

When it comes to tree trimming, the WEKA offers several different options. Pruning may be used as a technique for accuracy if an overfitting situation arises. Many different types of random forest algorithms may be used for regression and classification, and they all work by generating a large number of decision trees during training time and then determining which one is the most accurate. For the routine of decision trees to overfit their training set, random decision forests are ideal to use. Every leaf must be cleaned or pure before the categorization can be completed recursively in various computations. Dynamic conjecture of a choice tree's equilibrium between adaptability and exactness is the aim here.

I. Perceptrons with many layers

Using a non-linear modification of the input data, an MLP may be seen as a logistic regression classifier. Here, the input dataset is moved into space and placed in a linearly separable location

J. Association Rule Mining

Association rule mining [30] is an extremely well-known data mining approach that emphasizes fascinating and veiled links between distinct characteristics in a massive informative index. Association rule mining generates an arrangement of standards that describe the fundamental instances in the information collection.

The associativity of two attributes of crashes is dictated by the recurrence of their event together in the informational collection. A run X → Y demonstrates that if X happens then Y will likewise happen.

K. Interestingness computation

An association rule is considered a solid control if it fulfills the base limit criteria, i.e., support and confidence.

L. Cluster Selection Criteria

The most difficult part of cluster analysis is determining how many clusters to create from the data. The number of clusters may be identified using the given information criteria with LCC. For clustering purposes, we employed the Gap statistic (Tibshirani et al., 2001) as well as the AIC, BIC, and CAIC information criteria.

M. Data Collection

[70] Leeds UK's internet data source is used to collect traffic and road accident statistics. From 2011 to 2015, there were a total of 13062 incidents included in this data set. Data preprocessing yields 11 characteristics appropriate for further investigation. The following variables will be examined: the number of cars involved, the time of the accident, the road surface, the weather, the presence of lightning, the casualty's gender, age, and the day and month of the incident. Table 3.1 shows the accident data.

Table 3.1: Road Accident Attribute Description

S. No.	Attribute	Code	Value	Total	Casualty Class		
					Driver	Passenger	Pedestrian
1.	No. of vehicles	1	1 vehicle	3334	763	817	753
		2	2 vehicle	7991	5676	2215	99
		3+	>3 vehicle	5214	1218	510	10
2.	Time	T1	[0-4]	630	269	250	110
		T2	[4-8]	903	698	133	71
		T3	[6-12]	2720	1701	644	374
		T4	[12-16]	3342	1812	1027	502
		T5	[16-20]	3976	2387	990	598
		T6	[20-24]	1496	790	498	207
3.	Road Surface	OTR	Other	106	62	30	13
		DR	Dry	9828	5687	2695	1445
		WT	Wet	3063	1858	803	401
		SNW	Snow	157	101	39	16
		FLD	Flood	17	11	5	0
4.	Lightening Condition	DLGT	Day Light	9020	5422	2348	1249
		NLGT	No Light	1446	858	389	198
		SLGT	Street Light	2598	1377	805	415
5.	Weather Condition	CLR	Clear	11584	6770	3140	1666
		FG	Fog	37	26	7	3
		SNY	Snowy	63	41	15	6
		RNY	Rainy	1276	751	350	174
6.	Casualty Class	DR	Driver		7657	0	0
		PSG	Passenger		0	3542	0
		PDT	Pedestrian		0	0	1862
7.	Sex of Casualty	M	Male	7758	5223	1460	1074
		F	Female	5305	2434	2082	788
8.	Age	Minor	<18 years	1976	454	855	667
		Youth	18-30 years	4267	2646	1158	462
		Adult	30-60 years	4254	3152	742	359
		Senior	>60 years	2567	1405	787	374
9.	Type of Vehicle	BS	Bus	842	52	687	102
		CR	Car	9208	4959	2692	1556
		GDV	Goods Vehicle	449	245	86	117
		BCL	Bicycle	1512	1476	11	24
		PTV	PTWW	977	876	48	52
		OTR	Other	79	49	18	11
10.	Day	WKD	Weekday	9884	5980	2499	1404
		WND	Weekend	3179	1677	1043	458
11.	Month	Q1	Jan-March	3017	1731	803	482
		Q2	April-June	3220	1887	907	425
		Q3	Jul-Sep	3376	2021	948	406
		Q4	Oct-Dec	3452	2018	884	549

Table 3.2 summarises the frequency of traffic-related deaths and injuries.

Table 3.2: Road Accident Attribute Description

S. No.	Attribute	Attribute Values	Code	Total
1.	Number of Victims NOV	1 victim	1	1200
		2 victim	2	855
		3 or more victim	+2	245
2.	Age of victim: AOV	0-18 years	CHD	305
		18-30 years	YNG	722
		30-50 years	ADU	815
		50 or more years	SNR	458
3.	Gender: GEN	Male	M	1589
		Female	F	711
4.	Time of day: TOD	[0-6]	T1	155
		[6-12]	T2	660
		[12-18]	T3	626
		[18-24]	T4	859
5.	Month: MON	Jan-Mar	Q1	611
		Apr-Jun	Q2	468
		Jul-Sep	Q3	590
		Oct-Dec	Q4	631
6.	Lighting condition: LIG	Day Light	DLT	1180
		Dusk	DUS	365
		Road Light	RLT	270
		No Light	NLT	485
7.	Roadway Feature: ROF	Intersection	INT	985
		Slope	SLP	320
		Curve	CUR	458
		Straight	STR	537
8.	Road Type: ROT	Highway	HIW	1459
		Local	LOC	841
9.	Accident Severity: ASV	Killed or Severe	KSI	712

N. Description of the Dataset for Result No. 5

Federal Aviation Administration (FAA) statistics on aircraft crashes (p. 71) All aircraft accidents from "1908 to 2020" are included in this dataset. There were a few characteristics in the dataset, including Fatalities, Aboard, Registration, Type, Route, Flight, Operator, Location, Time, Date, and Ground. Because the summary column was all text, we decided to remove it from this dataset. Our goal was to focus on text analysis.

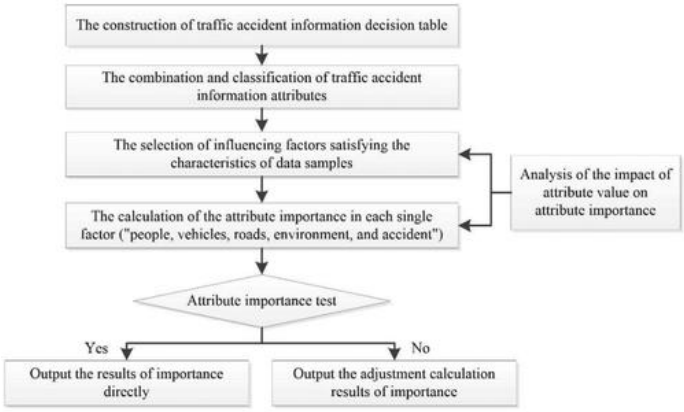

Fig. 3.3: The process of mining algorithm of traffic accident based on rough set theory.

III. RESULTS AND ANALYSIS

A. Overview

There are several machine learning approaches has been implemented to achieve accuracy and determine the most occurrence factor involved in an accident. Results are divided into 5 sections.

B. Result No. 1 (Road-user Specific Analysis of Traffic Accident using Data Mining Techniques)

SVM (support vector machine), naive bays, and decision trees have been used to categorize this dataset based on casualty class. Figure 4.1 displays the accuracy in categorization that was attained. Compared to the other two classifiers, decision trees had the greatest accuracy of 70.7 percent.

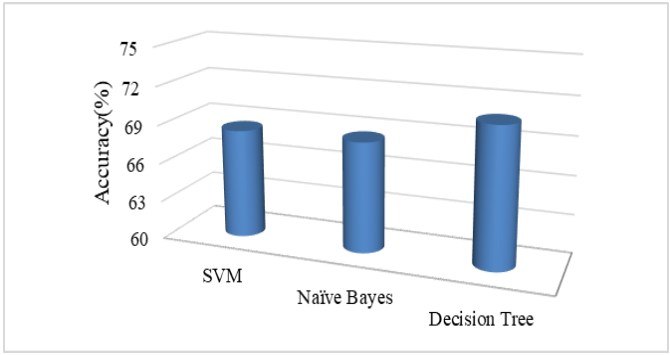

Fig. 4.1: Classification accuracy of different classifiers on accident data

"SOM (Self-organizing map) and K-modes" procedures, two clustering methods, have been used in this study's findings. Using k-modes instead of SOM, the outcome is better, and this shows that classifiers perform better when they are trained on k-mode clusters.

Datasets were sorted into three categories using SVM in this research based on the kind of casualty. There is an improved level of accuracy of 75.5838 percent in this classifier's output compared to a prior dataset that was not clustered, which is roughly 7 percent higher in terms of precision and recall. Table 4 shows the results of SVM on k-mode clusters.

Table 4.1: Performance of SVM

Rate of error= 0.1628								
Predicted values					Confusion Matrix			
Class	Precision	Recall	TPR	FPR	Class	DR	PSG	PDT
DR	0.779	0.909	0.90	0.36	DR	6958	153	546
PSG	0.824	0.375	0.37	0.03	PSG	1828	1330	384
PDT	0.630	0.851	0.85	0.083	PDT	146	132	1584

Based on casualty classification, Nave Bays was used in this research to categorize the dataset, and this classifier classified the dataset into three categories. Again, accuracy, error rate, error, recall, TPR, and other parameters play a significant influence in determining the final product. Clustering improved accuracy to 76.4583 percent, compared to 68.5375 percent without clustering. On k-mode clusters, Nave Bayes performs well, as seen in Table 4.2.

Table 4.2: Performance of Naive Bayes

Rate of error=0.2352								
Predicted values					Confusion Matrix			
Class	Precision	Recall	TPR	FPR	Class	DR	PSG	PDT
DR	0.788	0.86	0.86	0.33	DR	6649	515	493
PSG	0.697	0.43	0.43	0.07	PSG	1624	1535	383
PDT	0.742	0.828	0.828	0.078	PDT	170	151	1541

A Decision Tree classifier was utilized in this work, which resulted in a higher level of accuracy than previously achieved without clustering. An increase in accuracy of 18% was realized. Decision tree performance on clusters derived from k-modes is shown in Table 4.3.

Table 4.3: Performance of Decision Tree

Rate of error=0.1628								
Predicted values					Confusion Matrix			
Class	Precision	Recall	TPR	FPR	Class	DR	PSG	PDT
DR	0.784	0.893	0.893	0.348	DR	6841	422	394
PSG	0.724	0.457	0.457	0.065	PSG	1649	1620	273
PDT	0.683	0.770	0.770	0.060	PDT	231	197	1434

1) 4.2.6 Analysis

Each classification approach has been tested and found to have a satisfactory accuracy, false-positive rate (FPR), true positive rate (TPR), error rate (ER), recall (RR), and a distinct confusion matrix (CCM) for the various classification strategies. The confusion matrix may be used to compare the performance of various classification approaches.

As can be seen from these tables, the "total accuracy of analysis with clustering" is presented with the assistance of Table 4.2.

Fig 4.2 shows the classification accuracy of SVM, Naive Bayes, and decision trees on k-modes and SOM clusters. In comparison to SOM-derived clusters, those formed by k-modes clustering exhibit higher classification accuracy. SOM fails miserably in clustering data with categorical road accident characteristics, whereas the k-modes approach succeeds.

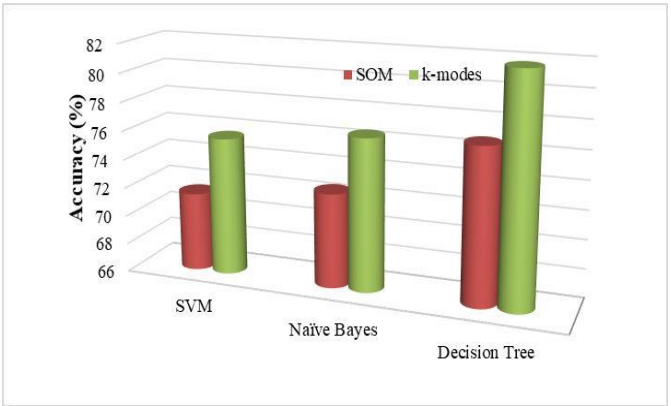

Fig. 4.2: Classification accuracy on clusters obtained from k-modes and SOM

As can be seen from Tables 7 and 12 the accuracy level increased after clustering. It is shown comparison chart in fig. 7 without clustering and with clustering.

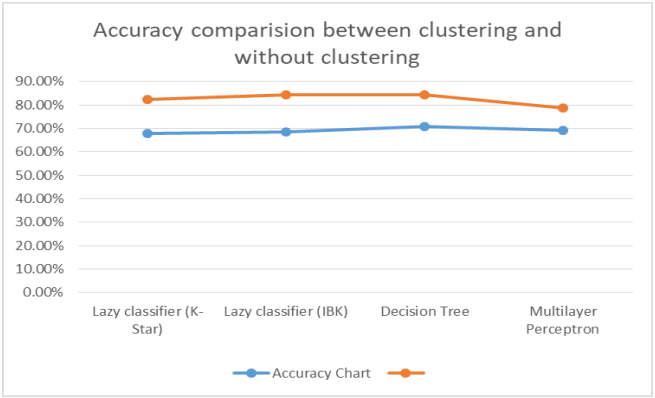

Fig. 4.5: Compared accuracy chart with clustering and without clustering

IV. CONCLUSIONS AND FUTURE WORK

A variety of data mining approaches have been used to analyze the accident dataset, including SOM (Self-Organizing Map), K-modes, Hierarchical clustering, latent class clustering (LCC), BIRCH clustering, and classification techniques such as Support Vector Machines (SVM), Nave Bays, Decision Trees (Random Forest, J-48, etc.), Multilayer Perceptrons, Lazy Classifiers (K-star and IBK), BIRCH clustering. In this research, the goal is to compare the classification performance before clustering methods and after clustering approaches. K-modes clustering was used to divide the dataset into four homogenous groups, which were then examined using Bayesian Networks in a subsequent step. Each cluster and all of the data are represented by a different Bayesian Network. Performance metrics are also used to assess these Bayesian Networks. The results show that the classification accuracy has increased marginally, while the ROC values for several clusters have declined slightly. As a result, the performance of the classifier in terms of accuracy is skewed toward one class value with an estimated high number of occurrences.

Data from traffic accidents in an Indian area was used to compare the performance of three clustering techniques: LCC, k-modes, and BIRCH. The goal of this research was to determine which of the three strategies above generates the most effective clusters for categorization. Different cluster selection criteria are used to identify the right number of clusters in the data. The number of clusters, k=2, was agreed upon by all of the criteria employed. Using the previously identified clusters, we next used three widely used classification approaches (NB, SVM, and RF) to further refine our findings. To compare LCC with NB and SVM, the findings revealed that it produced more accurate clusters and was capable of classifying data with the highest degree of accuracy. Several clustering approaches were examined for their computing performance on a variety of cluster models. k-modes are shown to be more efficient than the other two methods for generating diverse cluster sizes.

Results No. 5 were based on a thorough analysis of accident data. All accidents that occurred between "1908 - 2016" are included in this dataset. SOM, hierarchical clustering, and association rules are a few of the methods used to identify the relationships between the most often occurring phrases. Engine failure, weather conditions, pilot error, and other factors are often cited as contributing factors in aviation accidents. If the causes of crashes are taken into account, they might worsen shortly.

To pick a clustering technique to construct homogenous segments out of accident data based on computation speed or better clusters for classification, this research might be helpful.

In this thesis, we used the K modes clustering and association rule mining technique to develop a framework for studying road accident patterns. From 2009 to 2014, there were 11,574 accidents on the Dehradun district's road network. Based on the variables of accident type, road type, lightning on road, and road feature, K modes clustering found six clusters (C1–C6). Each cluster and the EDS have been used to construct rules using association rule mining. For the analysis, strong rules with high lift values are used. Each set of rules reveals the conditions around the incidents that occur inside that set. To see whether these associations may be made with an accident, the results of this comparison were compared to those created by EDS rules. If additional features linked with an accident are accessible, more information may be discovered. We also conducted trend analysis of all clusters and EDS on a monthly and hourly basis to reinforce our technique.

As previously stated, a large number of collisions occur often in areas of the road known as "accident hotspots." The building of new roads is impractical due to credit limits, thus increasing the safety of existing roads is the most significant measure in minimizing road accidents and road casualties with the largest effect. Several options exist under this strategy, including proactive steps to enhance safety by

addressing dangerous circumstances on the current road network to avoid accidents, as well as reactive actions aimed at addressing problem sites identified as accident hotspots.

Results from models based on artificial neural networks show that the frequency method of accidents is biased toward locations with large traffic volumes and also ignores accident severity. It's not included in the Equivalent Property Damage Only Index, and the divergence goes toward high-speed sites on residential roads. As a consequence, taking into account both of these factors may lead to more accurate findings. As a result of taking "return to mean in accident data" into account, the suggested technique improves estimate accuracy, making it the best choice for pinpointing high-accident areas on suburban highways in comparison to previous approaches.

According to trends, clustering before an analysis helps us locate better and more helpful outcomes that we wouldn't otherwise be able to find. According to the data, rural areas have a greater death rate than urban areas. The kind of vehicle, the driver's age, and the categorization of other road users are all included in the statistical analysis. A graphical depiction of the projected data outcomes is shown. Accident metrics may be better understood by the public via the use of graphic representations."

V. REFERENCES

[1]. F.M.O.I. Forensic Medicine Organization of Iran; Statistical Data, Accidents, online available on: http://www.lmo.ir/?siteid=1&pageid=1347

[2]. A.T. Kashani et al., "A Data Mining Approach to Identify Key Factors of Traffic Injury Severity", PROMETTraffic& Transportation, 23(1), pp. 11-17, 2011.

[3]. L.Y. Chang, H.W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques", Accident Analysis and Prevention, 38(5), pp. 1019-1027, 2016.

[4]. S. Yau-Ren et al. "The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents", Mathematical Problems in Engineering, Volume 2015 (2015), pp. 1-8., 2015. F. Babi and K. Zuskáová • Descriptive and Predictive Mining on Road Accidents Data– 92

[5]. R. Nayak et al., "Road Crash Proneness Prediction using Data Mining". Ailamaki, Anastasia & Amer-Yahia , Sihem (Eds.) Proceedings of the 14th International Conference on Extending Database Technology, Association for Computing Machinery (ACM), Upp-sala, Sweden, pp. 521-526, 2019.

[6]. V. Shankar, J. Milton, F. Mannering, "Modeling accident frequencies as zero-altered probability processes: An empirical inquiry", Accident Analysis & Prevention, 29(6), pp. 829-837, 2000.

[7]. A. Araar et al., "Mining road traffic accident data to improve safety in Dubai", Journal of Theoretical and Applied Information Technology, 47(3), pp. 911-927, 2013.

[8]. S. Vigneswaran et al., "Efficient Analysis of Traffic Accident Using Mining Techniques", International Journal of Software and Hardware Research in Engineering, Vol. 2, No. 3, 2014, pp. 110- 118, 2014.

[9]. L. Martin et al. "Using data mining techniques to road safety improvement in Spanish roads", XI Congreso de Ingeniería del Transporte (CIT 2014), Procedia - Social and Behavioral Sciences 160 (2014), pp. 607–614, 2014.

[10]. P. Flach et al., "On the road to knowledge: Mining 21 years of UK traffic accident reports", Data Mining and Decision Support: Aspects of Integration and Collaboration, Springer, pp. 143-155, 2013.

[11]. H. Zhang et al., "In-Memory Big Data Management and Processing: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 7, pp. 1920–1948, 2015.

[12]. J. Hipp, U. Güntzer, G. Nakhaeizadeh, “Algorithms for Association Rule Mining &Mdash; a General Survey and Comparison”, SIGKDD Explor Newsl 2, pp. 58–64, 2020.

[13]. A.T. Kashani et al., “A Data Mining Approach to Identify Key Factors of Traffic Injury Severity”, PROMETTraffic& Transportation, 23(1), pp. 11-17, 2021.

[14]. P. J. Ossenbruggen, J. Pendharkar et. al., “Roadway safety in rural and small urbanized areas”, Accidents Analysis & Prevention, 33(4), pp. 485-498, 2021.

[15]. R. Agrawal, T. Imieliski, A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, pp. 207–216, 2018.

[16]. R. Agrawal, R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases”, Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 487-499, 2019.

[17]. L. Breiman, “Random Forests”, Machine Learning, Vol. 45, pp. 5 - 32, 2021

[18]. Savolainen P, Mannering F, Lord D, Quddus M. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid Anal Prev. 2021;43:1666–76.

[19]. Depaire B, Wets G, and Vanhoof K. Traffic accident segmentation utilizing latent class clustering, accident analysis, and prevention, vol. 40. Elsevier; 2018.

[20]. Karlaftis M, Tarko A. Heterogeneity considerations in accident modeling. Accid Anal Prev. 2000;30(4):425–33.

[21]. Ma J, Kockelman K. Crash frequency and severity modeling using clustered data from Washington state. In: IEEE Intelligent Transportation Systems Conference. Toronto Canadá; 2016.

[22]. Jones B, Janssen L, Mannering F. Analysis of the frequency and duration of freeway accidents in Seattle, accident analysis and prevention, vol. 23. Elsevier; 2021.

[23]. Miaou SP, Lum H. Modeling vehicle accidents and highway geometric design relationships, accident analysis and prevention, vol. 25. Elsevier; 2020.

[24]. Miaou SP. The relationship between truck accidents and geometric design of road sections–Poisson versus negative binomial regressions, accident analysis, and prevention, vol. 26. Elsevier; 1994.

[25]. Poch M, Mannering F. Negative binomial analysis of intersection-accident frequencies. J Transp Eng. 1996;122.

[26]. Abdel-Aty MA, Radwan AE. Modeling traffic accident occurrence and involvement. Accid Anal Prev Elsevier. 2021;32.

[27]. Joshua SC, Garber NJ. Estimating truck accident rate and involvements using linear and Poisson regression models. Transp Plan Technol. 2000;15.

[28]. Maher MJ, Summersgill I. A comprehensive methodology for the fitting of predictive accident models. Accid Anal Prev Elsevier. 1996;28.

[29]. Chen W, Jovanis P. Method of identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec. 2012:1717.

[30]. Chang LY, Chen WC. Data mining of tree-based models to analyze freeway accident frequency. J Saf Res Elsevier. 2015;36.

[31]. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Pearson Addison-Wesley; 2006.

[32]. Abellan J, Lopez G, Ona J. Analysis of traffic accident severity using decision rules via decision trees, vol. vol. 40. Expert System with Applications: Elsevier; 2013.

[33]. Rovsek V, Batista M, Bogunovic B. Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree, transport. UK: Taylor and Francis; 2014.

[34]. Kashani T, Mohaymany AS, Rajbari A. A data mining approach to identify key factors of traffic injury severity, prompt traffic & transportation, vol. 23; 2011.

[35]. Han J, Kamber M. Data Mining: Concepts and Techniques. USA: Morgan Kaufmann Publishers; 2021.

[36]. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002;97(458):611–31.

[37]. Sohn SY. Quality function deployment applied to local traffic accident reduction. Accid Anal Prev 2011;31:751–61.

[38]. Hung WT, Wong WG. An algorithm for assessing the risk of traffic accidents. J Saf Res. 2012;33:387–410.

[39]. Pardillo-Mayora JM, Domínguez-Lira CA, Jurado-Pina R. Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads. Accid Anal Prev. 2019;42:2018–23.

[40]. Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL, editors. Advances in latent class analysis. Cambridge: Cambridge University Press; 2012.

[41]. Oña JD, López G, Mujalli R, Calvo FJ. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks, accident analysis, and prevention, vol. 51; 2013.

[42]. Kaplan S, Prato CG. Cyclist-motorist crash patterns in Denmark: a latent class clustering approach. Traffic Inj Prev. 2013;14(7):725–33.

[43]. Chaturvedi A, Green P, Carroll J. K-modes clustering. J Classif. 2011;18:35–55.

[44]. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrica. 2020;62.

[45]. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on very large databases; 1994. pp. 487–99.

[46]. Akaike H. Factor analysis and AIC. Psychometry. 1987;52:317–32.

[47]. Raftery AE. A note on Bayes factors for log-linear contingency table models with vague prior information. J Roy Stat Soc B. 1986;48:249–50.

[48]. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J. 1998;41:578–88.

[49]. Wong SC, Leung BSY, Loo BPY, Hung WT, Lo HK. A qualitative assessment methodology for road safety policy strategies. Accid Anal Prev. 2014;36:281–93.