International
Journal
Of Advanced Research in Engineering& Management (IJAREM)
ISSN: 2456-2033 || PP. 49-59

Identification of Factors Influencing Injury Severity of Motorized Two Wheeler Crashes in Patna

Sachin Kumar Gupta[a], Ajai Kumar Singh[b]
[a]*Research Scholar, MNNIT Allahabad, Pin- 211004, Uttar Pradesh*
[b]*Professor, MNNIT Allahabad, Pin- 211004, Uttar Pradesh*

Abstract: Worldwide road crashes pose significant threat to social and economical life. WHO report – 2015 on road safety mentioned that the situation was more dangerous in the developing countries because of lack of proper enforcements and techniques in improving road user behavior and decreasing injury severity outcomes. In India as the urbanization is increasing, high speed road facilities have promoted the motorcyclists and other vulnerable road users (VRUs) to select these facilities. This has led to the increase in high VRUs road crashes. Lack of proper crash reporting system and complex nature of crash in heterogeneous traffic flow conditions have worsened the problem of road safety, particularly in the Patna, capital city of Bihar where crash severity in 2015 was 39.20 as compared to 29.10 of India.

The objective of this study was to identify the explanatory variables affecting the crash severity of motorized two wheelers in the Patna with the help of data mining. Two years (2014- 2015) of crash data, collected from police FIR reports, was used in the analysis. Total 17 categorical and numerical attributes such as time of a day, traffic signs, street lights, gap in medians and roadside features were used as independent variables. Roadways were divided into homogeneous segments in terms of land use pattern and vehicle mix of that area. Crash severity was divided into fatal, sever (incapacitating injury) and minor (Non- incapacitating injury). Decision tree models (J48 and random forest) were generated in the analysis of crash data because it can identify and easily explain the complex patterns associated with crash risk and do not need to specify a functional form. J48 and a Random Forest model were used using default parameters of Weka using 80% percentage split. The advantage of tree-based methods is that they are non-linear and non-parametric data mining tools for supervised classification and regression problems. They do not require a priori probabilistic knowledge about the phenomena under studying and consider conditional interactions among input data.

Both J48 and Random forest models were effective in predicting the crash severity with classification accuracies of 54 and 59 % respectively & having kappa statistics values of 0.32 and 0.37 respectively which falls in the fair agreement range.. Time of a day, no. of access/km, median openings, land use, on-street parking and street lights were found to be significant in predicting the injury severity levels in the city.

Keywords: Injury severity, explanatory variables, Data mining, J48, Random Forest models.

1. Introduction

Road crashes are a leading cause of death globally so the UN General assembly adopted the resolution and declared these ten years (2011-2020) as 'The Decade of Action for Road Safety" and called on the countries to implement the measures identified internationally to make their roads safer. Report further revealed that the number of deaths in road crashes- 1.25 million in 2013 - has reached a stable state with some little changes since 2007 but 68 countries have seen a rise in the deaths since 2010, of which 84% are low-or middle-income countries. Low-income countries have fatality rates more than double those in high-income countries. This is evident in the metropolitan cities of developing countries, more specifically in India, as they are witnessing the high growth in the vehicles due to the population increase and the growing economic conditions which leads to the various development activities. An increase in the expansion of high speed road facilities such as national and state highways coupled with rapid increase in the urbanization have led to the high proportion of vehicles using these facilities. World Health Organization (WHO, 2015) stated that with more and more people tend to choose high speed road facilities approx 50% of the total road crash victims were Vulnerable Road Users (VRUs) such as motorized two-wheelers (MTWs), pedestrians and cyclists. Out of the three, share of motorized two- wheeler fatalities are 50% alone. In the developing countries like India situation is alarming where 5 lacs injuries (1 accident every minute) &1.46 lacs killed (1 death every 3.6 minutes) occurred in 2015 (MORTH, 2015). 54.10% of road accidents victims were in 15- 34 year age. Share of MTWs crashes are 27.30% out of the total crashes. Rural areas are more prone to accidents accounting 53.80% of total accidents (MORTH, 2015).

Fatalities (61.0%) and injuries (59.10%) were also highest (compared to urban cities) (MORTH, 2015). Injury severity (deaths per 100 crashes) was 29.10 for India whereas for Patna it was 39.20. It should be understood that the road crashes are preventable problems caused by dangerous streets and aberrant behavior of road users. There is a need for the systematic approach so that the measure applied should be optimized and yield great results. 'Exploratory Data Analysis' (EDA) is one of the techniques available which make important part of the systematic approaches. Crashes generate lots of data so analyzing the data and then identifying hidden patterns of it in a better way (generally by visual methods) is what EDA is all about. EDA helps in identifying that whether a factor is safety related or not (Hauer, 2015). It is very important for the general people to easily understand where problem is and what the causes of it are. So EDA helps us to understand that what our data can tell us beyond the formal modeling and hypothesis tasks.

Data mining technique is being extensively used in business, medical, environmental, social and astronomical field. Its implementation in the traffic safety field is relatively few. Mainly data mining task can be divided into descriptive and predictive methods. Predictive Data mining technique available is Classification and Regression Tree (CART) whereas descriptive techniques are clustering and association rules. Classification is used for categorical data and regression is used when for continuous data. CART is a decision tree method. A decision tree comprises of a parent node which splits into the child node and by recursively partitioning we achieve the pure node. Pure node is that in which all class labels (predicted instances) are same. Many researchers have used the decision tree method because it automatically selects the best attribute to split and provides graphical visualization of analysis result.

The objective of the present work was to identify the factors responsible for the severity of two wheelers crashes in the Patna city with the help of data mining. There were two models developed in this study: J48 and Random Forest. It would help in identifying the potential crash locations at National Highways and in urban roads in India largely governed by similar land use, heterogeneity of the traffic, road geometry and environment condition.

The paper gives the brief review of literatures on modeling injury severity using both parametric procedures and non- parametric procedures the next section and then presents a methodology adopted for study then provides the information about the data and models used in this study followed by assessment of result and conclusion.

2. Literature Review

Identification of factors that influences injury severity outcome is one of the areas to improve road safety. As the crashes are random events so many literatures, for identifying factors related to injury severity, have focused on developing statistical models like logistic regression and ordered probit models (Donnell and Conner, 1996: Kockelman and Kweon, 2002: Abdel Aty, 2003: Wang and Kockelman, 2005: Lemp et al., 2011) where for example Donnell and Conner (1996) found that the age and vehicle speed factors contributed to the increase in the probabilities of injury level (serious and fatal). Other factors such as blood alcohol level, type of collision and vehicle used and occupant's position affects the other type of injury level. Kockelman and Kweon (2002) examined the injury severity outcome of two vehicle crashes and single vehicle crashes. Their study found out that the injury outcomes were based on the factors such as type of collision, gender, drunken driving, condition of vehicles involved and its type. In two vehicle crashes pickups and SUVs caused more damage to other drivers. Aty (2003) used the orderd probit method in roadway segments, signalized intersections and toll plaza. Factors such as gender, age, seat belt, type of vehicle, impact point and speed ratios were common between all three models. High probability of sever crash factors were older drivers, male drivers, no use of seat belt, over speeding and collision at driver's side. Dark light conditions and presence of curves were related to roadway crashes. It was found that driver's error was less probabilistic factor for his injury and could be the factor for other vehicles which he struck. Wang and Kockelman (2005) identified that increase in vehicle's weight was not the significant factor for causing injury while light duty trucks had increased the injury severity. Lemp et al. (2011) studied the impact of driver, vehicle and environmental factors on injury severities resulting from large truck crashes.

Computerization of our society has lead to the growth of available data and development of powerful data collecting, storing and analyzing tools. Data mining is one of the steps of knowledge discovery where intelligent methods are applied to extract the data patterns (Han and Camber, 2012). Data mining is now a very popular technique in social, medical, educational, security and business environment. However, its use in road safety is relatively few. Some researchers, by using data mining, have developed the popular decision tree based

models such as CART (Classification and Regression Tree) and MARS (Multivariate Adaptive Regression Splines) to identify the factors responsible for severity. It's one of the obvious benefits are that the individual components can be seen graphically which makes understanding of problem simple. It automatically selects the best input (independent) variables and from them, using the threshold value, classifies the output (target) variables. It also reduces the noise from the data. Tree based models have been used in severity analysis (kuhnert et al., 2000: Tesema et al., 2005: Chang and Wang, 2006: Kashani and Mohaymany, 2011: Montella, 2011: Griselda et al., 2012: Montella et al., 2012: Kashani et al., 2014). For example, Kuhnert et al. (2012) used logistic regression, CART and MARS techniques. Findings suggested that the parametric nature was a demerit of logistic regression model due to the linear terms used for continuous variables. MARS and CART were flexible. MARS model was better than CART and Logistic regression for their particular data. Age, experience and seat belts were important factors for injury. Tesema et al. (2005) used the decision tree approach and rule discovery for predicting injury severity. Findings suggested that the cause of accidents, age, and surface type, condition of road and light and type of vehicle were major factors. Chang and Wang (2006) developed the CART model to identify the factors related to severity. Vehicle type was identified as single most important factor. Type of collision, driver/ vehicle action and condition were the other important factors identified. Kashani and Mohaymany (2011) identified the factors that influence injury severity on two lane two way rural roads in Iran. They used the CART approach for it. They found that the seat belt was the major factor for driver and occupant's injury severity level and accident cause was second most important factor. Montella (2011) studied the different crash causing factors at urban roundabouts in Italy. Association rule discovery was implemented to find out the relation between different crash type and factors related to crash. Geometry of the roundabout was found to be the major factor for crash. In that, radius of deflection and low angle of deviation of entering approach were associated with angle and rear end crashes at entry. Very low angle of deviation and excessive radius of deflection of left approach were related to angle crashes at entry. Griselda et al. (2012) worked on rural roads in Spain and developed the decision trees to identify the factors influences severity outcome. Severity levels were fatal (K) and killed or seriously injured (KSI). CART model accuracy obtained was 54.43%. Montella et al. (2012) used CART and rule discovery procedure on powered two- wheeler (PTWs) crashes in Italy. It was found that road condition was main factor for PTWs crashes because of instability, grip of tyres and presence of curves. Kashani et al. (2014) also worked on two- wheeler pillion passenger crashes in Iran. They used CART to identify the factors responsible for influencing severity outcome. Accuracy of their model was 74%. Study revealed that the affected part of the body and land use pattern were the most important factors which influenced severity outcomes.

3. Methodology

In this part section 3.1 presents the overall study methodology adopted in this study. Section 3.2 describes the working concepts of decision trees. Section 3.3 gives the brief theory of the models used for analysis.

3.1 Study Methodology

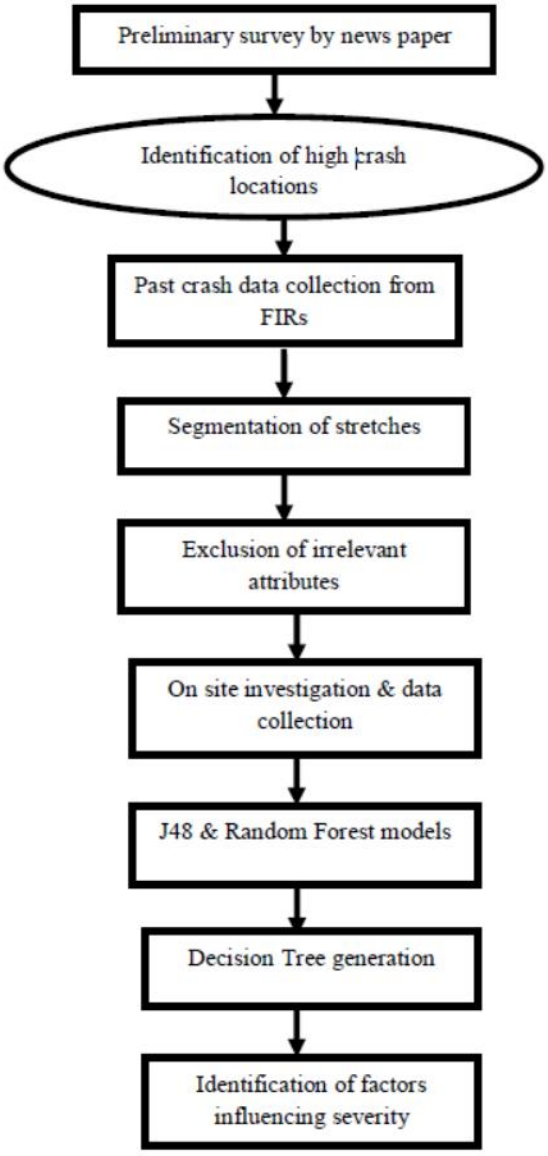

Fig: 3.1 Overall study methodology adopted in this study

3.2 Decision Tree

The process by which decision trees are built by the algorithms such as CART and ID3 are sometimes called 'top down recursive divide and conquer' approach. A decision tree is a representation of the data by some defined procedures. Structure of a decision tree is given in figure 3.2. It starts from a parent node or root node which gets split into branches upon testing on attributes to give child nodes. Each child nodes can again be treated as parent node by selecting the attributes this is called recursive approach. This process continues until we get the attribute which provides the 'pure split'. Pure split is achieved when the node will carry the similar class label. This last node is called the leaf node. So it seems like the tree is growing from top towards down.

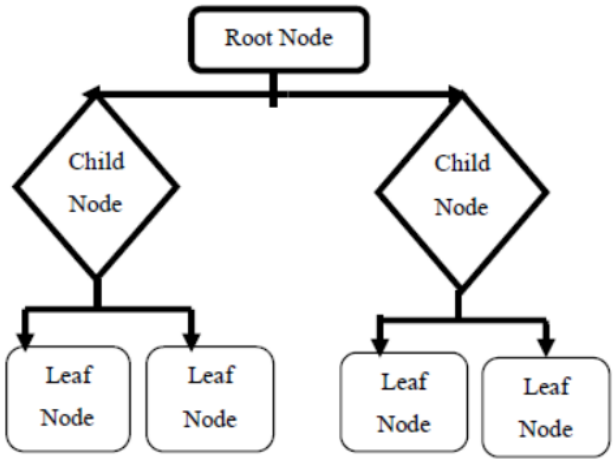

Fig: 3.2 Flow chart of a decision tree

Decision tree selects the best attribute to split based on 'Attribute Selection Measures' (ASM). ASM applies some heuristic procedures to select the best attribute. These procedures are 'information gain' or 'Gini index'. Information gain concept allows the tree to split into multilevel on the other hand Gini index makes tree to split into binary branches. CART uses Gini index and allows only binary split and J48 uses information gain concept to have multilevel splits. Here J48 algorithm has been used so the following paragraphs explain the information gain concept.

a) Information Gain

Information gain works on the concept of 'information theory'. It measures the information gain (purity) after the splitting. So the attribute which provides the highest information gain (has pure class) is selected. The expected information needed to classify the attribute D is given in equation no. 3.1 (Han et al., 2012).

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (3.1)$$

Where P_i, is the non-zero probability that an arbitrary instance in D belongs to class Ci and is calculated as $|C_i|/|D|$. After classifying the first best attribute, to select the best non-root attribute A to arrive at the exact classification use-

$$Info_A(D) = \sum_{j=1}^{v} \left|\frac{D_j}{D}\right| \times Info(D_j) \quad (3.2)$$

Where, D_j / D is the weight of j^{th} partition. $info_A(D)$ says how much information is still required to get the pure subset having pure class.

Information gain is the difference between the information requirement before partition and information requirement after the partition on A. It is given in equation no.3.3 as

$$Gain\ (A) = Info(D) - Info_A(D) \quad (3.3)$$

So, the more gain in the information the more pure the subset will be. It can be concluded that if attribute A is used to make split then it would results into the minimum information requirement.

One of the problems in the ID3 was the information gain problem. Information gain tends to overly favor the attribute which has lots of values in it. So in the improved version of ID3, which is C4.5, this problem was rectified by the use of another term called 'Gain Ratio'. Gain ratio penalizes the information gain for selecting these attributes by using 'split information'.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2 \frac{|Dj|}{|D|} \quad (3.4)$$

It shows the information that would be generated by spiting into v part, corresponding to v outcomes of the test on attribute A. Now the gain ratio is

$$Gain\ Ratio = \frac{Gain\ (A)}{splitInfo_A(D)} \quad (3.5)$$

The attribute which produces the highest gain ratio is selected as the best attribute to split. To avoid the over fitting and to get the generalized model predictions, the tree pruning procedure has to be applied and data should be divided into training and testing part randomly.

3.3 Models Used

a) J48 Model

Classification algorithm J48 is a successor of the initially build machine learning algorithm ID3 (Iterative Dichotomiser). ID3 was developed by statistician J. Ross Quinlanin in 1979. It was later modified and C4.5 came. Breiman et. al (1984) published a book called 'Classification and Regression Trees' (CART) for the generation of decision trees. Later C4.8 came in 1996 so at that time Weka was built using C4.8 and was written in java so they named this algorithm J48. ID3 and CART are same in their approach as they follow the training data to build a decision tree and then by using the testing set it makes the predictions. By using the information gain, gain ratio and pruning concept it modifies its tree to classify the attributes.

b) Random Forest Model

Random forest model is based on the concept of 'bagging'. Bagging means to average the noisy and unbiased models to create a model with low variance. Random Forest model creates lots of un-pruned decision trees from the training data set to make the classification. So in the end a kind of voting is done on the test data by these individual trees and maximum number of vote given to a particular class is taken as the prediction of the model.

4. Data

Section 4.1 gives the details of data used in this study. Then the next session 4.2 describes the process of segmentation of roadway and various explanatory variables used in the analysis, their representation, maximum and minimum values and descriptions have been shown in the table 4.1. These variables have been taken by doing survey of roads. In the last, the time slots used in this study has been presented in table 4.2. The idea of taking the different time slots was based on temporal volume mix in the city. Subsequently, the graph representing the injury severity against various times of a day has been presented in figure 4.1.

a. Secondary Data

Reported crash data from September 2014 to December 2015 (16 months) from a local daily newspaper were noted as a preliminary analysis. The attributes which were taken from the newspaper were date and time of accident (if mentioned), place of accident, total victims, and vehicle involved in it and at last causalities in 3 levels: fatal, serious, minor. It acted as an indicator of the crash trends at the various locations in the city. It was observed that the two- wheeler crashes were more in the city. Road crash data for 2 years (2014 - 2015) was collected from the different police stations of the city having jurisdiction over any part of the study area. Police stations in the Patna city viz. Didarganj, Zeromile, Patrakar Nagar, Agamkuan, Jakkanpur, Beur, Gardanibag, Sachivalaya, Shashtri Nagar, Khagaul, Airport, Gandhi Maidan, Kotwali and Rajiv Nagar maintained the crash report in the form of hard copy as FIRs. The information regarding crashes mentioned in FIRs were date, time,

location, vehicle involved, type of vehicle, vehicle number, cause, outcome, IPC act clause, number of persons dead and injured ,time of reporting, reporting person's or victim's name and investigating officer's name. Several cases of underreporting were also observed during reporting. After extracting the data of two-wheeler with heavy vehicle crashes from the FIRs, the total data points in this study were reduced to 219.

b. Segmentation and Primary Data

In some cases cause of accident, time and vehicles involved were not present. Exact locations were also not clearly mentioned in many cases. So the segmentation of the road was done. Segments were so chosen that the crash location would lie in it and nearly homogeneous surrounding conditions across roadway would prevail. This was based on the land use pattern, presence of hospitals, schools, petrol pumps, residential areas and commercial areas.

The study area was divided into 23 segments. After the consultation from the guides, the irrelevant attributes (which could not be identified from the FIRs or unavailable) such as horizontal curves, gradient, sobriety condition, qualification and age were removed. So, explanatory variables related to road characteristics, that were recorded physically, were reduced to the total 17.

Primary data includes road and roadside features which were collected physically by onsite survey of the whole study areas. All the measurements and the data points were recorded personally. The explanatory variables include the time, roadway features such as number of access per km, number of medians per km, minimum width of opening, maximum width of opening, shoulders, parking, turning traffic, construction of roadway, presence of street light, presence of traffic signs and land use pattern included attributes such as large or small commercial area, residential area, presence of hospitals, schools and petro pumps. Injury severity was categorized under three cases namely: Fatal, Sever i.e. incapacitated injury and Minor injury i.e. simple hospitalization needed or else no.

Table 4.1 Variables definition and summery statistics of the road segments

Variable	**Symbol**	**Description**			
Quantitative		**Min.**	**Max.**	**Mean**	**Remarks**
Stretch	Stretch	0	4.6	2.185	Segment length in km
Median opening	Med. Opn.	0	30	4.515	Gap in median in km
Min. width of opening	MinWO	0	7	2.489	Min. gap in median in m.
Max. width of opening	MxWO	0	80	16.178	Max. gap in median in m.
No. of access/km	NOA/KM	0	12.84	3.599	No. of access along the segment
Qualitative	**Symbol**			**Description**	
Shoulder	PU	PUO	NP	PU= Present unpaved, PUO= Present unpaved occupied, NP= Not Present	
Warning Signs	-	P	N	P= Present, N= Not Present	
Street Light (SL)	PW	PNW	NP	PW= Present working, PNW= Present not working, NP= Not Present	
Large Commercial establishment	LCE	P	N	P= Present, N= Not Present	
Small Commercial establishment	SCE	P	N	P= Present, N= Not Present	
Parked Vehicles	-	P	N	P= Present, N= Not Present	
Construction of Roadway undergoing	-	P	N	P= Present, N= Not Present	
Turning traffic	-	P	N	P= Present, N= Not Present	
Residential Area	-	P	N	P= Present, N= Not Present	
School	-	P	N	P= Present, N= Not Present	
Presence of Petrol Pump	-	P	N	P= Present, N= Not Present	

The data was then arranged in the seven different time slots keeping in mind the traffic volume mix temporal variations.

Table 4.2 Time slots

Time Interval	Notation
23:00 - 04:00	T1
04:00 – 06:00	T2
06:00 – 08:00	T3
08:00 – 12:00	T4
12:00 – 16:00	T5
16:00 – 20:00	T6
20:00 – 23:00	T7

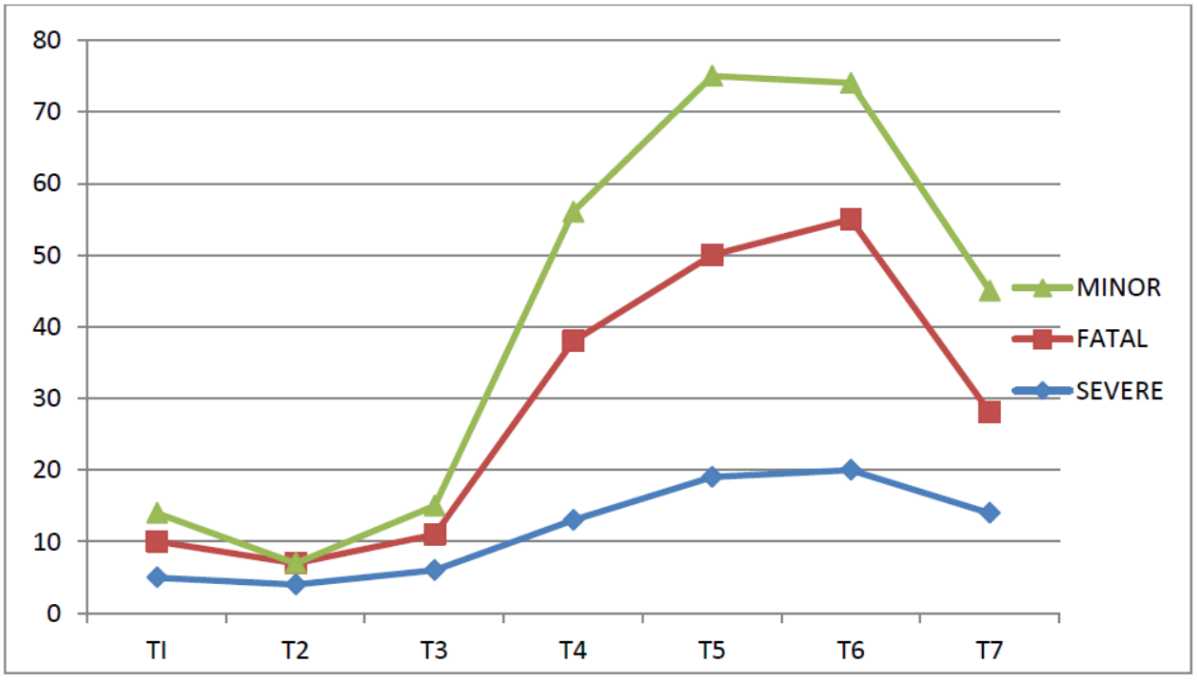

Fig.: 4.1 Plot of crash severity against different time slots

5. Results and Discussions

This section deals with the data analysis and presents the obtained result from both J48 and Random Forest models.

Data was divided into training and testing by random splitting method. From the 219 data points, for training purpose 80% data was retained and for testing purpose 20% of the data was retained.

Weka's explorer has an ability to select the best splitting attribute and making reasonable predictions by using its default parameter values (Witten et al., 2011). So by using the confidence factor 0.25, objects per leaf minimum was set to 2 and pruning was allowed. The first model used here was J48 and using the default parameters the accuracy was achieved at 54.55 % with Kappa statistics value of 0.3189 which is in fair agreement range (Landis and Koch, 1977). Second model was Random forest model and the classification accuracy achieved there was 59.09 % which is slightly better value than J48 model which had also kappa value in fair agreement range.

Table 5.1: Summery of Experiment conducted

S. No	Model Used	Kappa Statistics	Accuracy in Percentage
1	J 48	0.3189	54.5455
2	Random Forest	0.3704	59.0909

Since the present model had overall classification accuracy of both the models by and large the same so further evaluation of the model accuracy was carried out using the other classification accuracy measures such as 'true positive rate, precision, recall, F- measures and receiver operating curve area'. It is important to note

here that in a crash data, class imbalance problem usually occurs. Class imbalance problem is that when data shows the majority of negative class or the class which is not in interest. So even if the overall classification accuracy of a model is high, it is possible to have the class in interest accuracy low. Where, true positive rate (TPR) signifies the percentage of actual positive class correctly classified by the classifier out of the total positive class. It is also called as 'sensitivity'. Precision measures the percentage of the actual true classified class out of the total positive class which includes the false classified positive class also. Recall is same as the sensitivity. F- Measures are simply the harmonic mean of precision and recall. Last is receiver operating curve area (ROC area). Curve measures the prediction accuracy of model. It is a plot between the true positive rate (at y- axis) and false positive rate (x axis). The more ROC curve is at left side and up it is accurate (i.e. towards y-axis). The ROC area value < 0.5 means the poor classification accuracy and close to 1 means best accuracy. ROC area is also used to compare the accuracy of models.

Table 5.2 J48 Model: Detailed accuracy by Class

TP Rate	Precision	Recall	F- Measures	ROC Area	Class
0.429	0.462	0.429	0.444	0.692	Sever
0.556	0.667	0.556	0.377	0.703	Minor
0.667	0.500	0.667	0.386	0.725	Fatal

Table 5.3 Random Forest model: Detailed Accuracy by Class

TP Rate	Precision	Recall	F- Measures	ROC Area	Class
0.643	0.600	0.643	0.621	0.649	Sever
0.722	0.650	0.722	0.684	0.792	Minor
0.333	0.444	0.333	0.381	0.628	Fatal

From the table 5.2, it can be seen that J48 model has been good in predicting accuracy for fatal crashes having high TP rate 0.667 which is good. Table 5.3 shows that the random forest model has good accuracy in predicting the minor crashes and sever crashes in the city having TP rate 0.722 and 0.643 respectively. Precision is high in the J48 Model for the minor class as 0.667 and in random forest model it is high for the minor and sever class as 0.650 and 0.600 respectively which means that the models are classifying correctly high percentage of actual class out of total classified class by them.

Table 5.4 ROC area values of different class

S. No.	Model	Class Value	ROC area value	Average
1	J 48	Fatal	0.725	0.705
		Sever	0.692	
		Minor	0.703	
2	Random Forest	Fatal	0.628	0.701
		Sever	0.792	
		Minor	0.649	

Both the J48 and Random Forest model have shown their prediction accuracy in various crash severities. The overall classification accuracy of Random Forest model was better. To get the details of the individual class vise classification accuracy and to address the class imbalance situation, various other accuracy measurements were adopted such as TPR, precision, recall, F-measure and ROC area. So, from the analysis it was observed that J48 had better accuracy in predicting the fatal crashes in the city whereas sever and minor crashes were better predicted by the J48 model. ROC area value of J48 model for fatal crashes was identified as very good that means J48 model could act as the better classifier for fatal crashes. ROC values of Random forest model was observed and found to be close to 0.80 for sever injuries so it also could be used as a better classifier.

J48 model has been good in predicting accuracy for fatal crashes having high TP rate 0.667 which is good. Table 5.3 shows that the random forest model has good accuracy in predicting the minor crashes and sever crashes in the city having TP rate 0.722 and 0.643 respectively. Precision is high in the J48 Model for the minor class as 0.667 and in random forest model it is high for the minor and sever class as 0.650 and 0.600

respectively which means that the models are classifying correctly high percentage of actual class out of total classified class by them.

6. Conclusions

The objective of present study was to identify the factors that could influence the injury severity outcome in two- wheeler crashes. It was observed in many literatures that the mostly used non parametric technique for analyzing crash severity was CART analysis. However, in this study new data mining software 'weka' was used. It is easy to use and does not assumes a functional form of a model in advance which was the drawback of parametric procedures. J48 and Random Forest models were developed and tested on the validation set. From the decision tree it was observed that the attributes such as no. of access/km, time of the day, land use, median openings, shoulder conditions, parked vehicles, warning signs, length of the stretch, street lights and turning traffic were found to be significant out of 17 explanatory variables in different conditions. It was also observed in previous studies that the tree based models were unstable (Chang and Wang, 2006). So it is recommended that the use of decision tree results should be the complimentary to other techniques. For future work, the results of tree based models should be compared with the results of parametric procedures so that the measures implemented provide more significance.

Acknowledgment

Authors want to acknowledge the work of Waikato Institute, New Zealand for developing Weka software which is freely available also to Prof. Ian Witten for their online lectures.

Reference

[1]. Aty, A. M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34 (2003) 597– 603.
[2]. Beshah T., Hill S., (2010). Mining road traffic accident data to improve road safety: Role of road related crash factors on accidents severity in Ethiopia. AAAI Spring Symposium
[3]. Chang, L.Y. & Wang, W.H. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027
[4]. Donnell' O, C. J., Conner, H. D. (1996). Predicting the Severity of Motor Vehicle Accident Injuries using models of Ordered Multiple Choice. Accident Analysis and Prevention Vol. 28, No. 6, pp. 739-753, 1996
[5]. Global Status Report on Road Safety (2015). WHO
[6]. Griselda, L., Juan, O. D. & Joaquin, A. (2012). Using Decision Trees to extract Decision Rules from Police Reports on Road Accidents. SIIV - 5th International Congress - Sustainability of Road Infrastructures. Procedia - Social and Behavioral Sciences 53 (2012) 106 – 114
[7]. Han, J., Kamber, M. and Pei J. (2012). Data mining: concepts and techniques, 3rd edition. Morgan Kaufmann, USA.
[8]. Hauer, E. (2015). Art of Regression Modeling in Road Safety. ISBN 978-3-319-12529-9 (eBook) Highway Safety Manual- 1st edition (2009)
[9]. Kashani, A. T. & Mohaymany A.S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science 49, 1314–1320
[10]. Kashani, A. T., Rabieyen, R. & Besharati, M. M. (2014). A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. Journal of Safety Research 51 (2014) 93–98
[11]. Kuhnert, M. P., Do, A. k., & McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis 34 (2000) 371{386
[12]. Kockelman, M. K. & Kweon, Y. J. (2002). Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34 (2002) 313–321.
[13]. Landis, R. J. & Koch G. G. (1977). The Measurement of Observer Agreement for Categorical Data. International Biometric Society. Biometrics, Vol. 33, No. 1 (Mar., 1977), pp. 159-174
[14]. Lemp, D. J., Kockelman, M. K. & Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit model. Accident Analysis and Prevention 43 (2011) 370–380

[15]. Ministry of Home Affairs., Accidental death and suicides in India- (2014). National Crime Records bureau
[16]. MoRT&H, Basic Road Statistics of India, 2011 – 12, TRW
[17]. MoRT&H, Road Accidents in India (2015).
[18]. Montella, A. (2011). Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accident Analysis and Prevention 43, 1451–1463
[19]. Montella,A., Aria, M., D'Ambrosio, A. & Mauriello, F. (2012). Analysis of powered two wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention 49, 58– 72
[20]. Tesema, T. B., Abraham, A. & Grosan, C. (2005). Rule Mining and Classification of Road Accidents Using Adaptive Regression Techniques. I. J. of Simulation Vol. 6 No 10 and 11
[21]. Tesema, T., Abraham, A., Grosan, C., 2005. Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation systems 6, 80–94.
[22]. Witten, I., Frank, E., Hall, M. A. (2011). Data mining: practical machine learning tools and techniques, 3rd edition. Morgan Kaufmann, USA
[23]. Wang, X. & Kockelman, M. K., (2005). Occupant Injury Severity using a Heteroscedastic Ordered Logit Model: Distinguishing the Effects of Vehicle Weight and Type. Transportation Research Record No. 1908: 195-204, 2005.