International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Utilizing Machine Learning and Data Mining Approaches to Analyze Road Accident Data

[1]Awanish Kumar Singh, [2] V. Uday Krishna,[3] P. Rohitha, [4] N. Venkatesh

[1]Associate Professor,[2,3,4] Scholars
[1] Department of Computer Science and Engineering
[1,2,3,4] Lingaya's Instiitute of Management and Technology,Andhra Pradesh, India

ABSTRACT-

According to the National Safety Council report, approximately 38,300 people were killed and about 4.4 million injured in the United States in 2015. There are a variety of reasons that contribute to accidents some are internal to the driver but many are external. For example adverse weather conditions such as fog, rain or snow can cause reduced visibility and driving on such roads become difficult and dangerous.At the same time, clear weather condition but poor road pavement condition might also lead to accident-prone situations. Predicting any likelihood of crashes as the effect of these features would be a major step towards achieving better public safety. This attempts to create models that account for spatial and temporal features to determine likely road crash conditions. It is expected that the findings would help civic authorities to take proactive actions on likely crash prone weather and traffic conditions.

Keywords: Navie Bayes, SVM, Road Accidents, Datasets

I. INTRODUCTION

In an era where data is abundant and technology omnipresent, the quest for enhancing road safety has found a new ally in the realms of machine learning (ML) and data mining. Every year, road accidents claim millions of lives globally and inflict immeasurable economic losses. Amidst this grim reality lies a beacon of hope — the transformative potential of leveraging ML and data mining approaches to dissect, understand, and ultimately mitigate the underlying factors contributing to road accidents.This introduction serves as a gateway to explore the intersection of technology and road safety. Herein, we delve into the rationale behind utilizing ML and data mining methodologies to analyze road accident data, elucidating how these advanced techniques offer unprecedented insights and predictive capabilities. Firstly,

we outline the burgeoning landscape of road safety challenges, accentuating the multifaceted nature of accidents encompassing human behavior, infrastructure deficiencies, environmental factors, and vehicular dynamics. Traditional approaches to accident analysis often fall short in comprehensively addressing these complexities, necessitating a paradigm shift towards data-driven methodologies. ML and data mining present a compelling proposition by virtue of their ability to distill meaningful patterns

and trends from vast troves of heterogeneous data. Whether it's historical accident records, traffic flow data, weather conditions, or socio-economic indicators, these techniques empower analysts to uncover hidden correlations and causal relationships that elude conventional analysis. Moreover, the dynamic nature of road safety demands real-time insights and adaptive strategies. ML algorithms excel in this regard, capable of continuously learning and evolving from new data inputs, thereby enabling proactive interventions and predictive modeling to avert potential accidents.However, the efficacy of ML and data mining hinges upon the quality and diversity of data inputs.

Hence, we underscore the imperative of data collection frameworks and collaboration among stakeholders to ensure the availability of comprehensive and reliable datasets for analysis.In essence, this introduction sets the stage for a deep dive into the manifold applications of ML and data mining in road accident analysis. By harnessing the power of these cutting-edge technologies, we embark on a journey towards a safer, more sustainable future on the world's roads. The labyrinthine nature of road safety challenges demands a holistic approach that transcends conventional wisdom. From the intricate interplay of human behavior and vehicle dynamics to the subtle nuances of infrastructure design and environmental factors, the tapestry of causality defies simple explanation.

Navigating the modern labyrinth of roadways presents a persistent challenge fraught with peril. Each year, the toll of road accidents reverberates globally, leaving behind shattered lives and fractured communities. Yet, amidst this sobering reality lies a beacon of promise — the fusion of machine learning (ML) and data mining techniques with the vast reservoirs of accident data. This introduction serves as a gateway to explore the transformative potential of these methodologies in decoding the intricacies of road accidents.

Predicting the occurrences of vehicular crashes on State of Iowa roadways based on spatial and

temporal features of weather and traffic. The analysis will Result in Graphical Representation Predictive Features for Vehicular crash:

- Average Speed
- Dew Point Temperature
- Chances of Rain
- Wind Speed in knots
- Traffic volume

The primary focus is to assess the dataset pertaining to accidents. Street accidents claim the lives of thousands of individuals annually. Urban areas and highways are the most common locations for road accidents. The consequences of these accidents are severe and result in the loss of valuable lives.

II. LITERATURE SURVEY

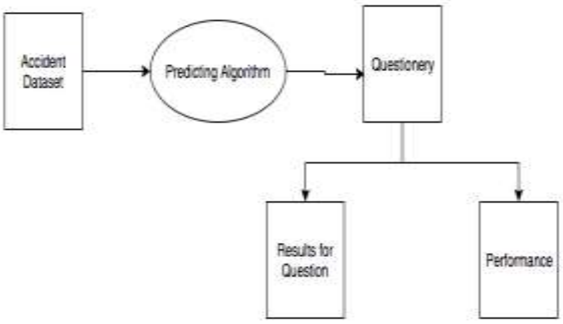
The literature Survey on real-time traffic crash prediction and analysis incorporates advanced modeling techniques to leverage data from traffic flow, weather, and road condition sources. In "Calibrating a Real-Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data," Abdel-Aty and Pemmanaboina develop a model using real-time traffic flow variables and rain data for crash-likelihood prediction. By utilizing PCA and LR, they create a rain index and a matched case-control logit model to identify crash potential based on loop data and rain index, providing insights into significant variables affecting crash occurrence.

Chen, Chen, and Ma in "Crash Frequency Modeling Using Real Time Environmental and Traffic Data and Unbalanced Panel Data Models" focus on modeling crash frequency using real-time environmental data such as visibility and road surface conditions. Their approach involves zero-inflated, negative binomial models with unbalanced panel data, revealing the influence of real-time and site-specific factors on crash frequency.

Yu, Abdel-Aty, Ahmed, and Wang's "Utilizing Microscopic Traffic and Weather Data to Analyze Real-Time Crash Patterns in the Context of Active Traffic Management" investigates the I-70 Freeway, employing AVI and weather detection systems to categorize crashes and develop hierarchical logistic regression models for crash pattern analysis. The paper highlights the need for different active traffic management strategies for various road segments, considering factors like seasonality and road characteristics.

Together, these studies contribute to a comprehensive understanding of the factors influencing traffic crashes and offer sophisticated modeling approaches for real-time crash prediction and management.

III. BLOCK DIAGRAM



IV. METHODOLOGY

In these proposed method The data methods were involved:

- Data collection
- Data Validation
- Data pre-processing
- Feature extration

- Evaluation model

IV.I Data Collection:

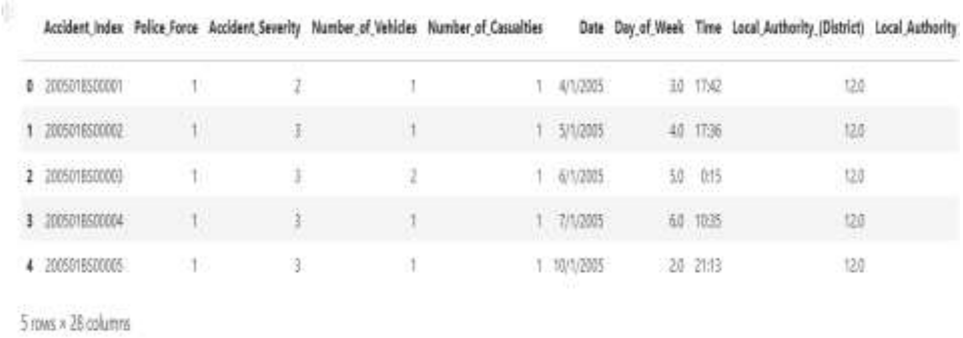
The data used was a series of product reviews collected as part of a traffic accident analysis. This step is about selecting a subset of all available data that you want to use. Machine learning problems are best started with data, a lot of data (examples or observations) for which you already know the target answer. Data for which you already know the answer to the target is called labeled data. We used 3 data sets for analysis. The data collection process for this project involved gathering a comprehensive dataset of traffic accident information from a variety of reliable sources. The data collection phase prioritizes obtaining a broad data set over several years to ensure that the analysis captures important trends and patterns in traffic accidents. Special attention is paid to identifying and correcting any inconsistencies or missing values by conducting data validation and verification processes to ensure the quality and completeness of the data collected.

IV.II DATA VALIDATION

Import the library package and load the specified dataset. Analyze variables by data form and data type to identify and evaluate missing and duplicate values. Validation datasets are data samples that are retained when training a model and are used to provide an assessment of model functionality when optimizing the model and process, allowing you to take full advantage of the validation and test datasets when evaluating the model. Data cleaning/preparation to analyze univariate, bivariate and multivariate processes by renaming specified datasets, removing columns, etc. Data cleaning steps and techniques vary depending on the data set. The main goal of data cleaning is to detect and remove errors and anomalies to increase value of data in analysis and decision-making.

IV.III DATA PRE-PROCESSIG

The data preprocessing phase played a crucial role in preparing the collected road accident datasets for analysis. This phase encompassed several essential steps aimed at enhancing the quality, consistency, and suitability of the data for subsequent analysis. Initially, the raw data underwent thorough cleaning procedures to identify and handle missing values, outliers, and erroneous entries. Techniques such as imputation, where missing values were replaced with estimates derived from the remaining data, were employed to ensure completeness. Additionally, outlier detection methods were applied to identify and address data points that deviated significantly from the norm and could potentially skew the analysis results. Organize selected data by formatting, cleaning and sampling.



	Accident_Index	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of_Casualties	Date	Day_of_Week	Time	Local_Authority_(District)	Local_Authority
0	200501BS00001	1	2	1	1	4/1/2005	3.0	17:42	12.0	
1	200501BS00002	1	3	1	1	5/1/2005	4.0	17:36	12.0	
2	200501BS00003	1	3	2	1	6/1/2005	5.0	0:15	12.0	
3	200501BS00004	1	3	1	1	7/1/2005	6.0	10:35	12.0	
4	200501BS00005	1	3	1	1	10/1/2005	2.0	21:13	12.0	

5 rows × 26 columns

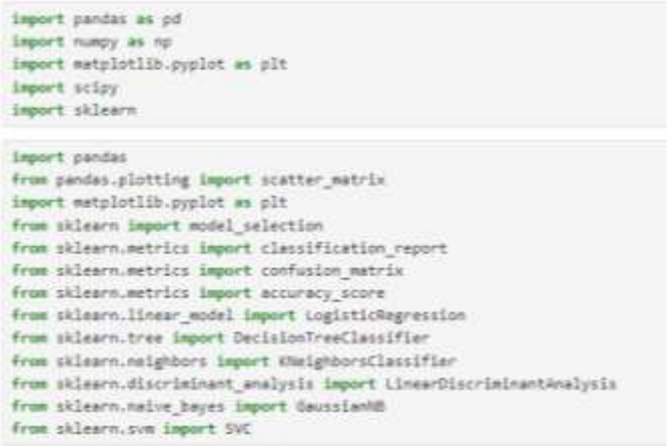
Fig 1 Data Pre-processing from datasets

IV.IV FEATURE EXTRATION

Next is feature extraction, which is the attribute reduction process. Unlike feature selection, where existing attributes are ranked according to their predictive significance, in feature extraction, the attributes are actually transformed. The transformed attributes or features are linear combinations of the original attributes. Finally, our model is trained using a classifier algorithm. We use the classification module of the Natural Language Toolkit library in Python. We use the collected labeled dataset. The rest of our labeled data is used to evaluate the model. Several machine learning algorithms have been used to classify preprocessed data. The chosen classifier is Random Forest. These algorithms are very popular in classification tasks.

IV.V EVALUATION MODEL

Evaluation Model is an integral part of the model development process. It helps in finding the best model that represents our data and the future performance of the chosen model. Evaluating model performance on the data used for training is unacceptable in data science There are two methods for evaluating models in data science: holdout and cross-validation. To avoid overfitting, both methods use a test set (unseen by the model) to evaluate model performance. The performance of each classification model is estimated based on its mean value. Results are provided in a visual form. Represent categorical data in chart form. Accuracy is defined as the percentage of correct predictions on test data. This can be easily calculated by dividing the number of correct predictions by the number of total predictions.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy
import sklearn
```

```
import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```

Fig 2 Models for Prediction

ALGORITHMS

The algorithms were involved in this proposed system:

- Navie Bayes
- Decision Tree
- Logistic Regression
- Support Vector Machine

A. Navie Bayes:

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem with the assumption of conditional independence between the features. It is often used for classification problems and can be applied to analyze road accident data effectively. feature selection or engineering is performed to identify the most relevant predictors**.**

Gaussian Naive Bayes:

- This type is used when the features are continuous and are assumed to follow a Gaussian (normal) distribution.
- The mean and variance of the features for each class are calculated and used for classification.

Multinomial Naive Bayes:

- This type is used when the features are discrete counts (e.g., word counts in a text classification problem).
- It is commonly used for document classification tasks such as spam detection.

Bernoulli Naive Bayes:

- This type is similar to Multinomial Naive Bayes but is used when the features are binary (e.g., presence or absence of a word in a document).
- It is also commonly used for document classification tasks.

B. Decision Tree:

A decision tree is a popular machine learning algorithm used for classification and regression tasks using In the context of analyzing road accident data, decision trees can help identify patterns and relationships between different factors and the likelihood of an accident.

C. Logistic Regression:

Logistic regression is a machine learning technique commonly used in analyzing road accident data to assess the probability of an event occurring. This model can be employed to predict binary outcomes such as the likelihood of a road accident occurring given a set of independent variables such as weather conditions, time of day, road type, traffic density, and driver behavior. Logistic regression works by estimating the probability of a particular event (e.g., an accident) and categorizing the outcome into binary classes (e.g., accident or no accident). The model learns the relationship between the input variables and the outcome variable from a dataset and can then make predictions on new data. This analysis can help identify key risk factors contributing to road accidents and guide policymakers and safety authorities in implementing measures to reduce accidents and enhance road safety.

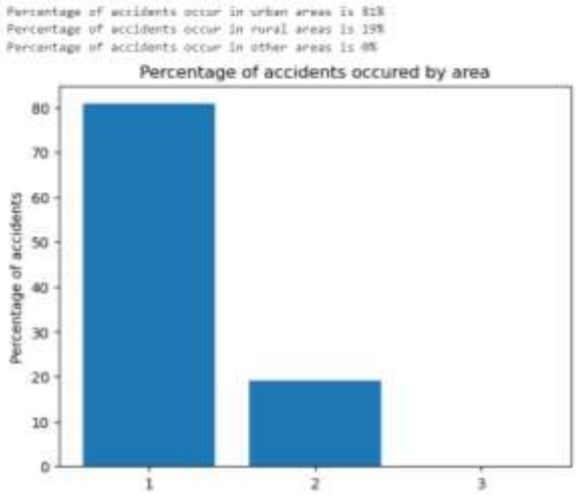
D. Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

- Notations: Consider a linear classifier for a binary classification problem with labels, y and features, x
- y ∈ {+1,-1} (instead of 0 and 1)
- Classifier: Note this classifier will directly predict either +1 or -1
- Without going through intermediate steps of estimating the probability of y being 1 like logistic regression

V. RESULTS AND DISCUSSION

When analyzing road accident data by area, it is important to consider how accidents are distributed across different types of areas, such as urban, suburban, and rural regions. Typically, urban areas tend to have higher traffic density and more complex road networks, which may lead to a higher percentage of accidents due to congestion and the frequency of interactions between vehicles, pedestrians, and cyclists. Conversely, rural areas often have lower traffic volume but may present different risks such as higher speeds and less lighting or signage, which can result in a different set of accident circumstances.



```
Percentage of accidents occur in urban areas is 81%
Percentage of accidents occur in rural areas is 19%
Percentage of accidents occur in other areas is 0%
```

Fig 3 Percentage of Accidents Occurred by Area

Analyzing road accident data by time can provide valuable insights into the patterns and trends associated with when accidents are most likely to occur.In Fig 4 This analysis often involves examining accident data based on specific time intervals such as time of day, day of the week, and even seasons or months. Graphs can visually represent this data, such as bar charts showing the distribution of accidents across different hours of the day, or line graphs depicting the fluctuation in accidents throughout a week or a year. These visualizations can reveal peak accident times, such as rush hours or late-night periods

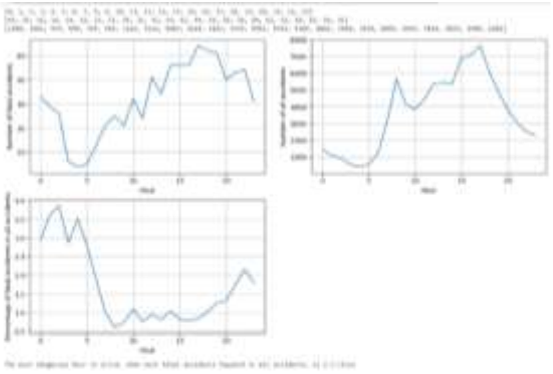


Fig 4 Accidents data analyzed by Time

In Fig 5 Analyzing road accident data over the years using graphs can provide valuable insights into trends and patterns in traffic safety. Graphs that illustrate the number of accidents by year can reveal whether there has been an overall increase or decrease in road accidents over time. A line graph, for example, can help identify periods of improvement or decline in traffic safety. Peaks and troughs in the data may correspond to changes in road safety policies, vehicle technology advancements, or variations in road infrastructure. Moreover, identifying specific years with significant deviations from the average accident rate can prompt further investigation into potential causes such as weather conditions, economic changes, or alterations in driving habits. By analyzing these trends, policymakers can assess the effectiveness of existing safety measures and identify areas for further improvement.

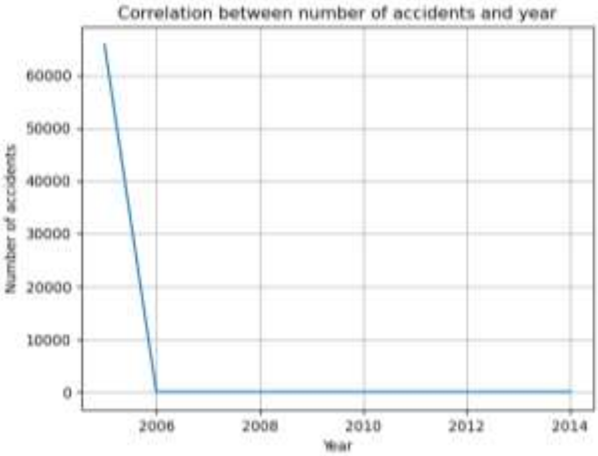


Fig 5 Data showing number Accidents by year

Analyzing accident data in relation to casualty rates per accident and speed limits in Fig 6 can reveal important insights into traffic safety. By plotting casualty rates against different speed limits, one can observe how changes in speed restrictions may impact the severity of accidents. Typically, higher speed limits are associated with higher casualty rates per accident due to increased impact forces during collisions. By visualizing this data in a graph, patterns may emerge indicating specific speed ranges where casualties are more frequent or severe. This information can help inform policy decisions on optimal speed limits to enhance road safety.

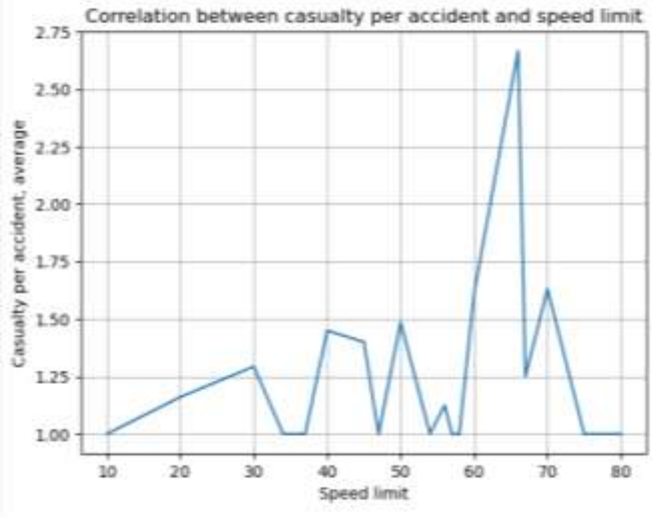


Fig 6 Casualty per accident and Speed Limit

Analyzing road accident data in the context of the number of car accidents by driver age can offer valuable insights into road safety and risk factors. The data could be visualized in a bar chart or line graph, where each age group is represented on the x-axis, and the number of accidents is plotted on the y-axis. This graphical representation allows for a quick comparison of accident rates across different age groups and can help identify which age groups are at greater risk.

In Fig 7 Analyzing accident data in relation to the light conditions can reveal important insights into the safety of roadways under different lighting scenarios. By graphing the number of accidents per light condition—such as daylight, dusk, dawn, and nighttime—patterns can emerge that highlight the times of day and light conditions when accidents are most likely to occur. For instance, a higher incidence of accidents during nighttime could suggest the need for better street lighting or reflective road markings. Similarly, a spike in accidents at dawn or dusk could indicate issues with visibility as natural light changes. By understanding these patterns, policymakers and transportation authorities can take targeted measures to improve road safety, such as adjusting traffic signals, enhancing street lighting, or promoting safer driving practices during particular times of day.

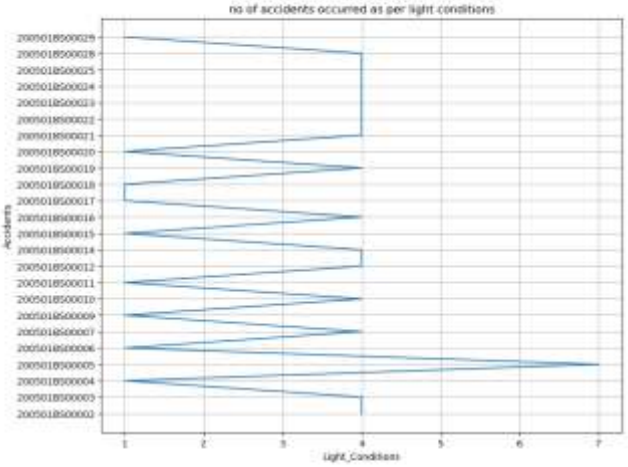


Fig 7 Number of Accidents Occurred in light conditions

VI. CONCLUSIONS AND FUTURE WORK

The study showed that carriageway width, shoulder type,road condition, land use and composition of trucks in traffic stream were the most significant parameters affecting accidents on urban roads. Some of the recommendations that were drawn from the above results and field observations have been enlisted below: Provision of dedicated lanes for the slow moving heavy vehicles and very strict monitoring of the regulation, A minimum width of unpaved shoulders to be provided for the road infrastructure which also contributes to the reduction in the rate of the accidents,Stringent measures to be taken to restrict the movement of heavy vehicles during the peak hours,Restrict the number of minor exits and implement effective design provisions to allow for safe traffic diversion along such exits.

To allow safe convergence of traffic near the fly over exit, provision of signals (ramp meters) to regulate the inflow of traffic from below the fly over. In future, we will take the dataset in a state-wise pattern so that it will become the big project. This project will save the people life in an efficient way. and can Implementation of some other algorithms in Machine learning.

REFERENCES

[1] “Road Accidents in India”, A report by Ministry of Road Transport and Highways, 2011.

[2] R. V. Ponnaluri, “Modeling road traffic fatalities in India: Smeed’s law, time invariance and regional specificity,” International Association of Traffic and Safety Sciences, vol. 36, no. 1, Jul. 2012, pp. 75–82.

[3] Halima DrissiTouzani, "Data mining techniques to analyze traffic accidents data: Case application in Morocco"

[4] Gagandeep Kaur, "Prediction of the cause of accident and accident prone location on roads using data mining techniques"[10] Muhammad Yogi Ilham, "Analyzing Highway Road Accident Characteristic Using Data Mining"

[5] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations”, J. Mod. Transport. (2016)4(1):62–72.

[6] S. Shanthi and Dr. R. Geetha Ramani, “Gender Specific Classification of Road Accident through Data Mining Techniques”, IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012) March 30, 31, 2012.

[7] Miao Chong, Ajith Abraham and Marcin Paprzycki, “Traffic Accident Data Mining Using Machine Learning Paradigms”,Oklahoma State University, USA.

[8] Tessa K. Anderson, “Kernel density estimation and K-means clustering profile road accident ”, Accident Analysis and Prevention 41 (2009) 359–364.

[9] H. Meng, X. Wang, and X. Wang, “Expressway crash prediction based on traffic big data,”. Proceeding Ser., pp. 11–16, 2018.

[10] Dr. P. Pramada Valli, “Road accident models for metropolitan cities of India”, International Association of Traffic and Safety Sciences, vol. 29, no. 1, 2010, pp. 57–64