Classification of Road Traffic Accident Data Using Machine Learning Algorithms

Bulbula Kumeda, Fengli Zhang, Fan Zhou
School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, 610054
e-mail: bekumeda@gmail.com, {fzhang, fan.zhou}@uestc.edu.cn

Sadiq Hussain
System Administrator
Dibrugarh University
Dibrugarh, Assam, India
e-mail: sadiq@dibru.ac.in

Ammar Almasri
Al-Balqa Applied University
Jordan
e-mail: Ammar.almasri@bau.edu.jo

Maregu Assefa
School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, 610054
e-mail: maregu2006@gmail.com

***Abstract*—The dramatic increase in road traffic accidents in the world is causing serious problems in every aspect of human lives. The most important and meaningful nature of traffic characteristics, causation analysis, and associations between different causal factors have been ignored. Moreover, the traffic accident data is only used to conduct a rudimentary statistical analysis and data mining efforts which results only in patterns and statistics. The main targets of this road accident data classification are to identify the major and key factors that cause the road traffic accident and form policies and preventive actions that would reduce the accident severity level. Machine learning algorithms are used to analyze the data, extract hidden patterns, predict the severity level of the accidents and summarize the information in a useful format. In this work, we have applied different machine learning classification algorithms and discussed here the six algorithms with high accuracy and best classification performances such as Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network (Radial Basis Function Network), Multilayer Perceptron, and Naïve Bayes on road traffic accident data set obtained from UK road traffic accident of the year 2016. The data set contains information on all road accident casualties across Calderdale. The results from our analysis show that Fuzzy-FARCHD algorithm is effective to classify the dataset and achieves an accuracy of 85.94%. In this work, we have revealed that Lighting Conditions, 1st Road Class & No., Number of vehicles are the key features in selecting the attributes.**

Keywords-machine learning; road traffic accident; fuzzy-FARCHD; classification algorithms

I. INTRODUCTION

With the advent of vehicle technology and road infrastructural development, the mobility of people from place to place increased in an exponential way. There are different kinds of transport systems that help people to do business or to enjoy or to study moving to different places. One of the most commonly used transport systems is a vehicle transport system. It is also the main cause of road traffic accidents in the world. We all wish to stay safe and avoid traffic accidents as a human being but, a greater number of road users couldn't return back home, spent long time in health centers or hospitals, as well as never be capable to do things as the way used to be and more they will goodbye this world for once at all because of road traffic accidents.

Every year as a consequence of road traffic accident greater than 1.25 million people decease on the world's street and around 50 million people get injured [1]. In low and middle-developing nations large amount of hospital beds are occupied by injured people because of traffic accident [2]. If serious measures and preventive actions are not engaged, road traffic accidents are figured to be the eighth foremost reason of death in the worldwide and are expected to be the main contributor to the worldwide problem of injury and disease by 2030 [3]. Every year because of traffic accidents 90% of the deaths occur in the low and middle developed countries, while they account only 54% of registered vehicles in the world. This shows that the degree of accident severity is asymmetric compared to their number of vehicles [2].

A phenomenon which takes place in the roads or streets exposed to the public involving at least one moving vehicle and resulting with one or more person being injured or killed is known as Road traffic accident (RTA). There are various types of accidents, some of them include crash between vehicles, a crash between pedestrians and vehicles, between vehicles and animals, and vehicles and physical objects. Peoples such as cyclists, motorists, pedestrian, commuters, horse-riders, and non-public transport like train passengers are called road users.

According to a system concept based on man-environment changes and instabilities of Muhlard and Lassare, the main reason of road traffic accident is usually by faults in three main constituents of road safety such as the environment, the human, and the vehicle [4]. The

978-1-7281-2184-0/19/$31.00 ©2019 IEEE

environmental factors incorporate the natural and constructed environments as well as the transportation networks. The most common environmental factors affecting road traffic accidents are weather condition, dusty air, heavy wind, light conditions. The infrastructural factors under environment are road type, road condition, road lane type, poor or defective Road surface, inadequate road markings, defective traffic signs, and road layout. The human constitutes includes sex, age, educational level, human behaviors, driving skill, driving manner, risk recompense, and hazardous driving (use of drugs and alcohols), speeding, breaching traffic lights, driving behavior. The vehicle constituent includes the configuration, age, volume, and quality of vehicles like technical conditions and safety equipment, design defects.

II. Literature Review

Currently, road traffic accident has been both developmental and public health concern and demanded the concern of researchers, civil societies, vehicle companies, governments, and business societies in the whole world.

In the framework [5] a large-scale heterogeneous accident data is used to estimate the occurrence of traffic accidents on the road crossings. The authors proposed an object detection method called Fast R-CNN for photo features extraction and XGBoost for extracting road map and driving record features. The overall experimental outcome shows that the algorithms used in this work can extract the possibly dangerous crossings with a good performance.

To classify and discover hidden patterns using a dataset that contains the road accident records obtained from the Philippines National Police (PNP), Naïve Bayes, Decision Tree, and Rule induction are used in [6]. They use a Rapid miner data mining tool to analyze the accident data. The authors revealed that the places where accidents occur don't have a significant relationship on the fatality of the victims. The results show that key factors affecting the accident and found that time and day are the most critical causes for the severity and fatality of road accident victims and the algorithms attain the expected accuracy.

This research article states that to predict the main influencing factors of accidents like reasons of the accident, accident-prone sites, the severity of the accident, and type of vehicle involved and so on many research studies conducted in order to increase the performance of DM classification [7]. The authors used two data mining tools, Weka and Orange to evaluate J48, Multi-layer Perceptron, Bayes Net classifiers of 150 instances of the dataset. Evaluation metrics for measuring data mining techniques such as accuracy, precision, recall or sensitivity and so on are applied to identify the best algorithm for the prediction of the accident dataset. The experimental result shows that Multi-layer Perceptron is the best for the prediction of the accident database with an accuracy of 85.33%.

In [8] the authors used a genetic algorithm to develop a symbolic fuzzy classifier on traffic accident dataset obtained from Addis Ababa traffic office. The symbolic classifier used to select features from accident dataset. The result shows that the developed classifier is able to separate and classify the classes of injuries and the attributes used for data labeling are easily extracted and explored. This author also proposed a machine learning experimental research which uses a traffic accident data gathered in Ethiopia [9]. They used CART, Random Forest, MARS and Tree Net algorithms to develop a predictive model focused on exploring the issue of data quality and predicting the impact of road behaviors on potential injury risks. The models can identify the human-related causal factors for the accident severity. The combined techniques used in this work proved to be effective in terms of predictive accuracy.

In [10] was developed to employ predictive analytics through innovative machine learning models to predict the future result of the accidents' number in Oman. The author utilized Boosted Tree Regression model which is based on the decision tree and Multiplicative model to increase the prediction results. This work states that the sole or contributory factor for accident cause is related to human factors which counted about 91% of the total accidents and the rest 9% is non-human factors.

Apriori, Naïve Bayes, and k-means clustering algorithms are used in [11] to analyze the FARS dataset with the aim of examining the relationship between fatal rate and other features such as a drunk driver, light condition, collision mode, weather condition, and road surface conditions. In this research work, the variables associated highly with fatal accidents are articulated. The result shows that the factors related to human factors like a drunk driver cause a high fatality rate. While the work in [12] utilized Naïve Bayes and Apriori algorithms to predict patterns in the road accident. In this work, the authors developed a prediction model to predict the frequently occurring accident types on new roads based on the association rule. From the analysis, results show that most of the accidents are occurred by vehicles age less than five years and there is a high mortality rate in the rural areas.

As stated by several researchers, data mining techniques have a vast role in analyzing and predicting the future value of road accidents records and in identifying the patterns of the components of accidents determining different factors. In addition, the great potential of data mining prediction techniques plays a major role in avoiding and monitoring the problems of road accident safety.

III. The Methodology

A. Data Gathering and Preprocessing

In this research paper, we have used a dataset obtained from data.gov.uk which is a traffic accident data in the United Kingdom which occurred in 2016. The table below defines the attributes of the data set and its description in details.

TABLE I. Attributes with Description

Attribute	Attribute Description
No of Vehicles	The total number of vehicles which takes part in the accident
Time (24hr)	The exact time when the accident occurs
1st Road	A road where the accident occurred (Motorway, non-

class	motor way…)
1st Road Class & No	A road which has a zonal system (A, B, C, Unclassified…)
Road Surface	The surface condition of the road during the accident (Dry, wet/Damp...)
Light Condition	A light condition during the accident (Daylight: street lights present…)
Weather Condition	Condition of the weather during the accident (Fine without high winds...)
Causality class	The causality of the accident (Driver or rider, Pedestrian...)
Sex of Causality	The gender of the causality during the accident (Male or Female)
Age of Causality	The age of the causality during the accident (age given in years)
Causality Severity	The asperity of the accidents (Fatal, Serious, Slight)
Type of Vehicle	Vehicle type during the accident (Pedal cycle, Car…)

In machine learning, a classification technique is supervised learning that dataset contains all features along with class value to learn from dataset to build a model to predict final class value. Speech recognition, biometric identification, handwriting recognition, document classification and so on are some of the examples of classification problems [13].

In this research work, we have used different classification algorithms from which the six classification techniques such as Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network, Multilayer Perceptron, and Naïve Bayes have achieved high classification accuracy. Fuzzy-FARCHD which is a well-known machine learning rule-based classification algorithm that can deliver an explainable model for the user [14]. The fuzzy rule-based classification has been effectively employed in data mining methods with the goal of discovering hidden knowledge from a dataset in the method of explainable rules and design an accurate classification model. Random Forest is an easy and compromising machine learning classification algorithm which can yield a great outcome without hyper-parameter tuning [15]. They build many decision trees and unites them together to achieve a high and accurate prediction. In the present machine learning systems, a Random forest can be applied for both problems of classification and regression. Hierarchal LVQ classification system is supervised prototype-based algorithm used to find vectors with multidimensional space that the best describe each of a number of classifications. It has an adjustment and training method similar to SOM. An input layer, Kohen classification, and output layer are the three layers in Hierarchal LVq structure [16]. A Radial Basis Function or RBF network is a distinctive type of NN with unique features which are trained and use the same like neural network perceptron [17]. Hierarchal LVq, RBF network has three layers: an input layer where its neurons accept vector values, a hidden layer which is single and receives n-dimensional vectors values from input layer and the third layer is an output layer where every neuron characterizes a classification. In the RBF network, every output layer neurons will be motivated based on the possibility with which the data provided to the input layers fits the classification. A multilayer perceptron (MLP) is a type of feedforward artificial neural network that functions sets of input data onto a set of fitting outputs. It contains an input layer, a hidden layer and an output layer [18]. Naïve Bayes is a classification technique that is easier to design and it's well suited for applications containing a huge dataset. It is the most commonly-used, simple and effective machine learning classifier. In Naïve Bayes classifier, even if the feature is dependent on each other the existence of one character in a class is disparate to the existence of another [19].

In this research paper, we have employed a different machine learning classification algorithms and chose the six algorithms with high accuracy and classification performance such as Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network, Multilayer Perceptron, and Naïve Bayes to our dataset to classify the data sets and specifically discover the insights from the dataset.

IV. Evaluation Metrics

Evaluation metrics used to examine the performance of each classifier model are True Positive Rate, False Positive Rate, Accuracy, Recall, Precision, F-Measure, Kappa, Mean, absolute error Root, mean squared error, Relative absolute error and Root relative squared error. In order to evaluate and conclude our results, the confusion matrix is adapted to analyze each algorithm by comparing actual results (rows of the table) and model results (columns of the table). Accordingly, model results represent the number of instances that are predicted either correctly or incorrectly as shown in Table II [20, 21].

TABLE II. Confusion Matrix

Confusion Matrix			
		Target (actual results)	
		C1	**C2**
Model (predictive results)	**C1**	**TP**	FP
	C2	FN	**TN**
	Correctly Classified instances		**Incorrectly Classified instances**
	CCI=TP+TN		**ICCI=FP+FN**

True Positive Rate (TP Rate) the ratio of observations that are positive and also predicted as observed (positively). The true positive rate is given by

$$TP/ (TP+FN) \quad (1)$$

False Positive Rate (FP Rate) the ratio of observations that are positive but were predicted negatively.

The false positive rate is given by

$$FP/ (TN+FP) \quad (2)$$

The accuracy metric corresponds to the number of records or instances that are classified correctly overall historical records in our dataset.

Accuracy is given by

$$TP + TN / (TP + TN + FP + FN) \quad (3)$$

The Precision represents the ratio of positive instances that were correct in the actual dataset, while Recall includes the ratio of actual positive instances that were labeled correctly.

Precision is given by

$$TP / (TP + FP) \quad (4)$$

Recall is given by

$$TP / (TP + FN) \quad (5)$$

The F-measure metric depends on the result of the previous two metrics (4, 5) by taking weight average for both results.

F-measure is given by

$$2 * (Precision * Recall) / (Precision + Recall) \quad (6)$$

V. PROPOSED METHOD

The 10 cross-validation methods are applied. The dataset is separated into two parts (90% training data) and (10% testing data). The total attributes take into consideration are 12. Fig. 1 shows the flow of the proposed work. In this classification work, we have three classes of injuries: fatal, serious and slight accident classes. The fatal is a class of an accident which results at least in the death of one person. Serious accidents are an accident type which results in severe physical injuries of a person that takes more than 2 days of spent in hospitals. Whereas slight injury is a simple injury which may be medicated easily with less than 2 days.

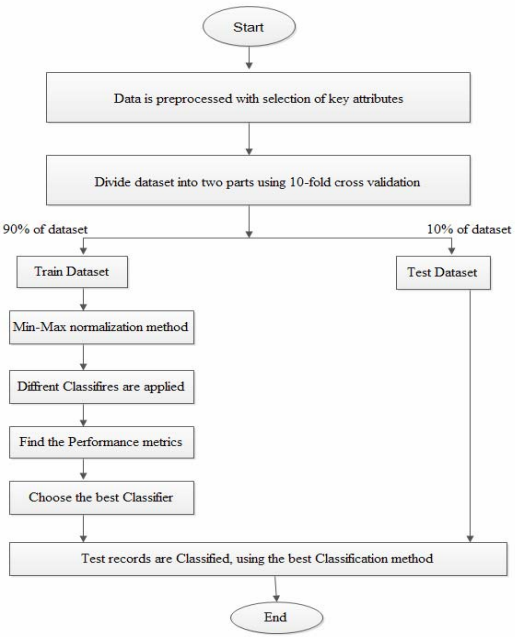

Figure 1. Flow Proposed method

VI. RESULT AND DISCUSSION

In this research work, 555 records of UK traffic accident data of the year 2016 are mined to gain insight. Figure 2, 3 and 4 depict the pictorial representation of the dataset. Figure 1 describes the proposed framework of the methodology. At the preprocessing stage, some of the non-influential attributes like reference number and date of an accident are not considered. The features are selected using feature selection methods with ranking. According to Gain Ratio Attribute Evaluation method, Lighting Conditions, 1st Road Class & No., Number of vehicles and sex of the causality are the influential attributes of the dataset. From our analysis of the dataset, we observed that most of the causalities' age lies between 20 and 40 years. Most of them were male. The type of vehicle engaged in accidents was mostly cars. Most of the accidents occurred in daylight with street lights present and the weather conditions were fine without high winds. The min_max normalization method is applied to achieve better accuracy of the dataset. Different machine learning algorithms are applied in search of better accuracy. Six algorithms are listed with higher accuracies. They are Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network, Multilayer Perceptron, and Naïve Bayes. The statistical parameters are shown in Table IV. The fuzzy-FARCHD classifier topped the chart with 85.94% accuracy. In this work, we have also described a comparison of different classifiers using evaluation metrics, confusion matrix and fold wise accuracy on the dataset shown in Table III, IV and V below respectively.

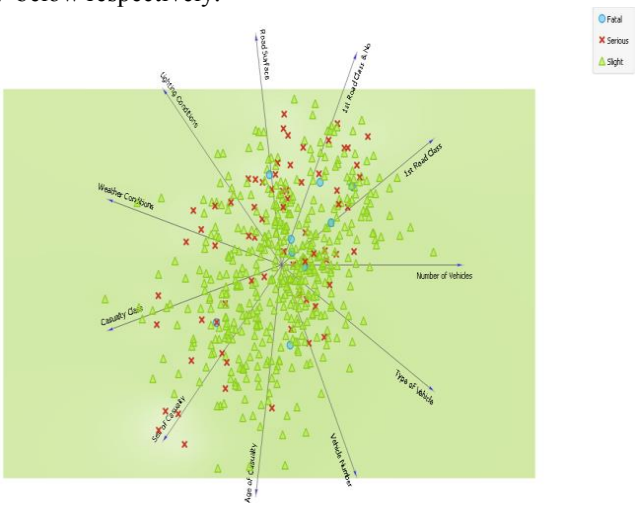

Figure 2. FreeViz Diagram of the Dataset

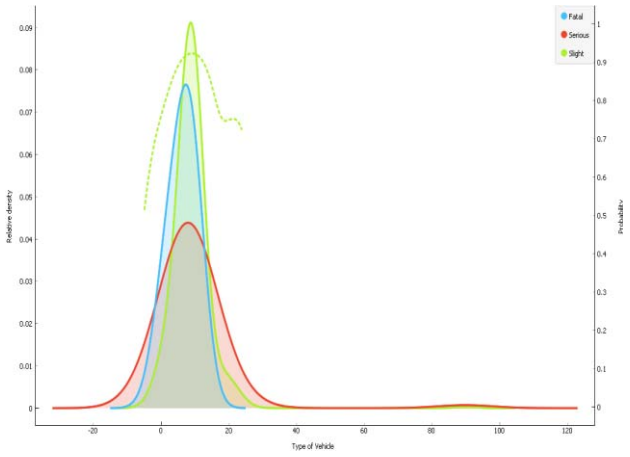

Figure 3. Distribution of type of vehicle and Casualty Severity

TABLE III. COMPARISON OF DIFFERENT CLASSIFIERS USING EVALUATION METRICS

SN	Algorithm (Data mining classifier)	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
1	Fuzzy-FARCHD	85.94%	0.859	0.859	0.739	0.859	0.795	0	0.16	0.286	98.98%	99.99%
2	Random Forest	83.42%	0.839	0.733	0.793	0.808	0.609	0.1299	0.14	0.31	89.17%	108.2%
3	Hierarchal LVQ	82.16%	0.822	0.8	0.773	0.822	0.788	0.0259	0.11	0.344	71.88%	120.5%
4	RBF Network	84.14%	0.841	0.85	0.738	0.841	0.786	-0.014	0.16	0.2987	97.65%	104.3%
5	Multilayer Perceptron	79.27%	0.793	0.727	0.772	0.793	0.782	0.0704	0.1531	0.3489	92.55%	121.9%
6	Naïve Bayes	80.90%	0.809	0.775	0.776	0.809	0.783	0.0446	0.1776	0.316	107.3%	110.4%

TABLE IV. CONFUSION MATRIX FOR FUZZY_FARCHD

a	b	c	←classified as
469	8	0	a=Slight
63	6	0	b=Serious
7	0	2	c=Fatal

TABLE V. FOLD WISE ACCURACY ON THE TEST DATA

Fold0	0.8750000000
Fold1	0.8392857143
Fold2	0.9107142857
Fold3	0.8214285714
Fold4	0.8035714286
Fold5	0.8545454545
Fold6	0.8545454545
Fold7	0.8363636364
Fold8	0.8545454545
Fold9	0.8545454545

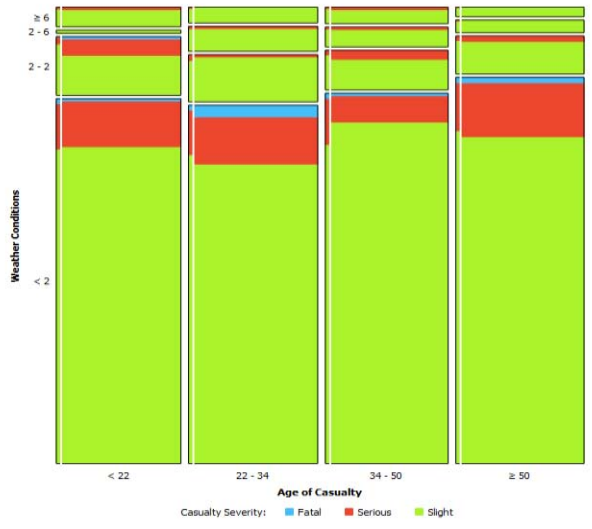

Figure 4. Mosaic Diagram of Age of Casualty, Weather Conditions and Casualty Severity

VII. CONCLUSION

Road traffic accidents are the key reason for serious injuries as well as the death of precious human lives. Discovering the factors that are related to the class values that are important to achieve an accurate result. The brutality of the accident problem is getting a catastrophic level becoming horrific shockingly and indicating that adequate measures have not been taken to prevent, control and/or lessen the appalling rate of the accident. Various scholars have tried to solve this issues but still, there are a lot of uncovered issues and gaps in predicting accident severity and specifically discover the influential factors such as time and season in which the accident frequently happened. This makes the field of traffic accident analysis and prediction more challenging. The main target of this research work is to discover the potential data mining technology at road traffic accident data for making a classification model. The developed classification model could support decision makers, policy designer and traffic officers for making effective decisions in traffic control actions. Using the analyzed outcome, decision-makers can easily understand the accident mode, driver's behavior, time, road and weather conditions and other key factors that are causing traffic accidents resulting in the fatalities and serious injuries so as to articulate improved traffic safety control strategies. They also may utilize the predictive models to take a new initiative in road safety, accident prevention and to develop new policies in this regards. Hence, we can conclude that Fuzzy-FARCHD algorithm can be applied in the machine learning tools to classify and predict the road accidents based on different attributes. For the future considering the frequently increasing size of the data sets, more features, and clusters, it's better to use deep learning techniques for improved classification and cluster of the data records.

ACKNOWLEDGMENT

The author would like to thank his supervisor Professor Fengli Zhang for her unconditional support to achieve good

academic research knowledge. The author would also like to acknowledge the people who participated by giving valuable comments and views for this work.

REFERENCES

[1] WHO. Road traffic safety. Available from Accessed on 21 Sep. 2017. http://www.who.int/mediacentre/factsheets/fs358/en/

[2] Dinesh, Road Safety in Less-Motorized Environments: Future Concerns 2002.

[3] World Report on Road Traffic Injury Prevention, 2015:12

[4] Krug, Sharma, and Lozano (2000), the global burden of Injuries

[5] Masashi Toyoda, Daisaku Yokoyama, Junpei Komiyama, Masahiko "Itoh 2017 IEEE International Conference on Big Data," Boston, MA, 11-14. 2017

[6] Jonardo R. Asor, Gene Marck B. Catedrilla, " A study on the road accidents using data Investigation and visualization in Los Baños, Laguna, Philippines" 2018 International Conference on Information and Communications Technology (ICOIACT) Yogyakarta, Indonesia. Page s: 96 – 101. 6 March 2018

[7] Sadiq Hussain, L. J. Muhammad, F. S. Ishaq, Atomsa Yakubu and I. A. Mohammed "Performance Evaluation of Various Data Mining Algorithms on Road Traffic Accident Dataset" Information and Communication Technology for Intelligent Systems pp 67-78, 30 December 2018.

[8] Tibebe BeshahTesema, Ajith Abraham, Dejene Ejigu, Learning the Classification of Traffic Accident Types"2012 Fourth International Conference on Intelligent Networking and Collaborative Systems. Page s: 463 – 468, 19-21Sept. 2012

[9] Tibebe Beshah Tesema, Dejene Eijigu, Ajith Abraham, "Knowledge Discovery from Road Traffic Accident Data In Ethiopia" 2011 World Congress on Information and Communication Technologies Page s: 1241 – 1246.Year: 2011

[10] Girija Narasimhan, Ben George Ephrem, et al, "Predictive Analytics of Road Accidents in Oman using Machine Learning Approach"2017 International Conference on Intelligent Computing, Instrumentation, and Control technologies (ICICICT) Page s: 1058 - 1065 July 2017

[11] Li, L, Shrestha, S, and Hu, G, "Analysis of Road Traffic Fatal Accidents using Data Mining Techniques," 2017IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA) Page s: 363 – 370, Year: 2017

[12] Ms. Nidhi. R, Ms. Kanchana V, "Analysis of Road Accidents Using Data Mining Techniques" International Journal of Engineering & Technology, Volume 7 (3.10) page 40-44, (2018)

[13] Sadiq Hussain "Survey on Current Trends and Techniques of Data Mining Research "London Journal of Research in Computer Science and Technology Volume 17 | Issue 1 | Compilation 1.0 (March 2017)

[14] Jesus Alcal ´ a-Fdez, Rafael Alcal ´ a, and Francisco Herrera "A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning" IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 5, OCTOBER 2011

[15] Niklas Donges Machine learning-blog.com Tutorials and explanations about applied Machine Learning "The Random Forest Algorithm," 06 February 2016

[16] David Nova · Pablo A. Estévez, "A Review of Learning Vector Quantization Classifiers" arXiv:1509.07093v1 [cs.LG] 23 Sep 2015

[17] M.R Mosavi, Mohammed Khishe, "Training Radial Basis Function Neural Network using Stochastic Fractal Search Algorithm to Classify Sonar Dataset" *Iranian Journal of Electrical & Electronic Engineering, Vol.13, No. 1, March 2017*

[18] Leo Dencelin X, and Ramkumar T, "Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures" Biomedical Research (2016) Computational Life Sciences and Smarter Technological Advancement August 29, 2016

[19] Han, J. and Pei, J., (2011). "Data Mining: Concepts and Techniques: Concepts and Techniques"

[20] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, "EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance," *Scientific Programming,* vol. 2019, 2019.

[21] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, clustering, and association rule mining in educational datasets using data mining tools: A case study," in *Computer Science Online Conference*, 2018, pp. 196-211.