Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)
IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

ANALYSIS OF ROAD ACCIDENTS IN INDIA USING DATA MINING CLASSIFICATION ALGORITHMS

Ms. E. Suganya
Ph.D Research Scholar
Department of Computer Science
Bharathiar University, Coimbatore-641046
elasugan1992@gmail.com

Dr. S. Vijayarani
Assistant Professor
Department of Computer Science
Bharathiar University, Coimbatore-641046
vijimohan_2000@yahoo.com

***Abstract*- Classification is a model finding process which is used for segmenting the data into different classes based on some constraints. This work analyzes the road accidents in India data set using classification algorithms namely linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm. Performance measures used are accuracy, error rate and execution time. This analysis is done in R data mining tool. The performance of KNN is better than other algorithms.**

Keywords- Classification, SVM, Regression, Decision Tree, Naïve Bayes, Gradient Boosting Algorithm,

I. INTRODUCTION

Data mining is the process of analyzing data from different location and summarizing it into useful information and it can be used for making intelligent business decision. It is not specific to any industry, applied in almost all areas to explore the possibility of hidden knowledge. Important data mining techniques are classification, clustering, time series analysis, association rule mining and regression.

Classification is an important data mining technique which analyzes the data and classifies the data into a predefined set of classes. There are one or more machine learning algorithms are available; they are linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm. The main aim of this work is to analyze the performance of machine learning algorithms using R tool. The performance measures are precision, recall, accuracy rate, true positive, false positive, error rate and percentage of correctly classified instances. Experimental results reveals that, KNN outperformed other algorithms with higher accuracy and a lower error rate. This research paper shows that four causes: fatal accident, major injury accident, minor injury accident, total accidents and year wise reports for the road accidents in India.

The remaining portion of the paper has seven sections. Section II describes the methodology of this analysis work. Section III presents the classification accuracy. The conclusion is given in section IV.

II. METHODOLOGY

A. Classification Algorithms

A classification technique is an important component of machine learning algorithms in order to extract rules and patterns of data that could be used for prediction [5]. Classification is the process of finding a set of functions that describe data classes and concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Different techniques from machine learning, statistics, information retrieval and data mining are used for classification [2]. They include Bayesian Methods, Bayesian Belief networks, Decision Trees, Neural Networks, Associative Classifiers, Emerging Patterns and Support Vector Machines (SVM). This work aims to compare the performance of machine learning algorithms such as linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm.

a. Linear Regression

It is used to estimate real values like cost of houses, number of calls, total sales etc. based on continuous variables. Here, establishing relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.

Where Y is dependent variable, a is slope, X is independent variable and b s intercept. These coefficients a and b are derived based on minimizing

the sum of squared difference of distance between data points and regression line.

b. Logistic Regression

It is used to estimate discrete values i.e. binary values like 0/1, yes/no, true/false based on given set of independent variables. In simple words, it predicts the probability of occurrence of an event by filtering data to a logic function. Hence, it is also known as logistic regression. Since it predicts the probability, its output values lies between 0 and 1.

c. Decision Tree

It is a supervised learning algorithm which is most commonly used for classification problems. It can work both categorical and continuous dependent variables. This algorithm is based on most significant attributes/independent variables to make as a distinct group as possible.

d. SVM (Support Vector Machine)

It is a classification algorithm which is plot each data item as a point in n-dimensional space where n is a number of features with the value of a particular coordinate.

e. Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between the predictors. A Naïve Bayesian model is easy to build and can be used for very large data sets. It can also handle numeric attributes using supervised discretization [1] [2]. The Naive Bayes algorithm is based on conditional probabilities. Bayes' Theorem determines the probability of an event occurring given the probability of another event that has already occurred. Figure 2 describes the posterior probability, P(c|x) is calculated from P(c), P(x), and P(x|c) [1]. The effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors [13]. This assumption is called class conditional independence. One way of classification is by determining the posterior probability for each class and assigning c to the class with the highest probability [8].

f. KNN (K-Nearest Neighbors)

It can be used for both classification and regression problems. It is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its k nearest neighbors measured by a distance function such as Euclidean, Manhattan, Minkowsi and Hamming distance. If k=1, then the case is simply assigned to the class of its neighbor.

g. Random Forest

Random forests (RF) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [4][3]. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Random forest is a group of classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. It runs efficiently on large data bases and can handle thousands of input variables without variable deletion [8]. Generated forests can be saved for future use on other data. Random Forests give many classification trees without pruning.

h. Gradient Boosting Algorithm

It is used to deal with plenty of data to make a prediction with high power. Boosting is actually an ensemble of learning algorithms which combines of the prediction of several base estimators in order to improve robustness over a single estimator. It combines multiple week or average predictors to build strong predictor.

III. EXPERIMENTAL RESULTS

For experimental results analysis, data is collected from Open Government Data Platform. We have used three performance factors. They are classification accuracy, error rate and execution time. By comparing linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm. The KNN algorithm has produced more accuracy with lower error rate than other algorithms.

A. Classification Accuracy

The following table shows the accuracy measures of classification algorithms. They are correctly classified instances, incorrectly classified instances and time taken. Table 1 and figure 1 shows the experimental results of classification algorithms such as linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm.

TABLE 1: CLASSIFICATION ACCURACY

Algorithms	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)	Time Taken (Secs)	Accuracy
Linear Regression	76.8	23.2	0.05	76.8
Logistic Regression	78.7	21.3	0.06	78.7
Decision Tree	83.6	16.4	0.02	83.6
SVM	75	25	0.06	75
Naive Bayes	83.7	16.3	0.02	83.7
KNN	**93.7**	**6.3**	**0**	**93.7**
Random Forest	78.6	21.4	0.03	78.6
Gradient Boosting Algorithm	82.4	17.6	0.02	82.4

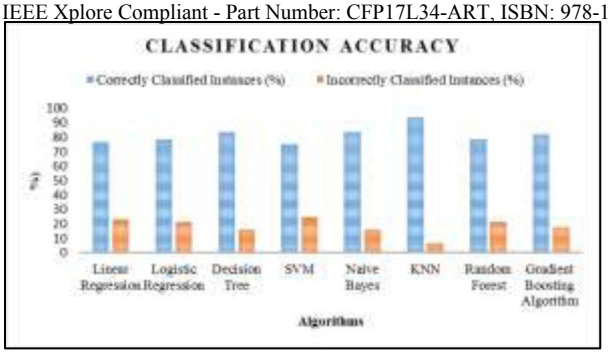

Figure 1: Classification Accuracy

B. *Error Rate*

Table 2 and figure 2 shows MAE (mean absolute error), (RMSE) root mean squared error, RAE (relative absolute error), and RRSE (root relative squared error). Mean absolute error and root mean squared error will be found in numeric value only. Relative absolute error and root relative squared error are shown in percentage for reference and evaluation. Table 2 mainly shows the results of the simulation. Here the KNN classification algorithm has the lowest absolute mean squared error.

TABLE 3. ERROR RATE

Algorithms	MAE	RMSE	RAE (%)	RRSE (%)
Linear Regression	0.0453	0.1552	94.3	94.2
Logistic Regression	0.1247	0.1980	78.4	89.3
Decision Tree	0.0156	0.1765	79.45	87.4
SVM	0.1456	0.1567	89.3	89.3
Naive Bayes	0.1657	0.0789	83.5	84.3
KNN	**0.0024**	**0.0045**	**68.5**	**82.44**
Random Forest	0.1456	0.1467	75.34	96.2
Gradient Boosting Algorithm	0.0788	0.1553	76.4	84.3

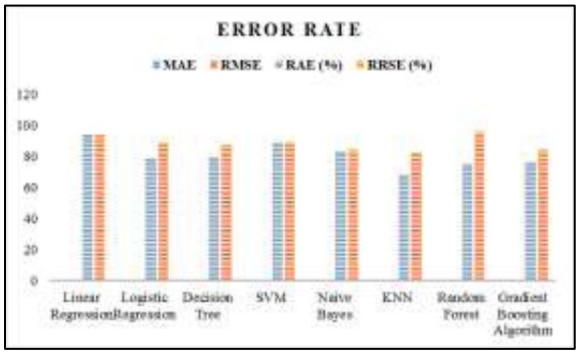

FIGURE 2: ERROR RATE

C. *Weight Average*

Table 3 and figure 3 shows True positive (TP) rate, False positive (FP) rate, precision, recall, and F-Score. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. High recall means that an algorithm returned most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant.

TABLE 4. WEIGHT AVERAGE

Algorithms	TP Rate (%)	FP Rate (%)	Precision (%)	Recall (%)	F-Measure (%)
Linear Regression	76.8	63.3	76.5	76.8	76.6
Logistic Regression	78.7	65.2	76.9	78.7	77.7
Decision Tree	83.6	83.6	64.2	83.6	72.5
SVM	75	62.5	85.3	75	76.5
Naive Bayes	83.7	83.3	78.1	83.7	72.7
KNN	**93.7**	**93.1**	**90.5**	**93.7**	**82.7**
Random Forest	78.6	65.4	76.4	78.2	75.6
Gradient Boosting Algorithm	82.4	82.4	67.2	82.4	72.7

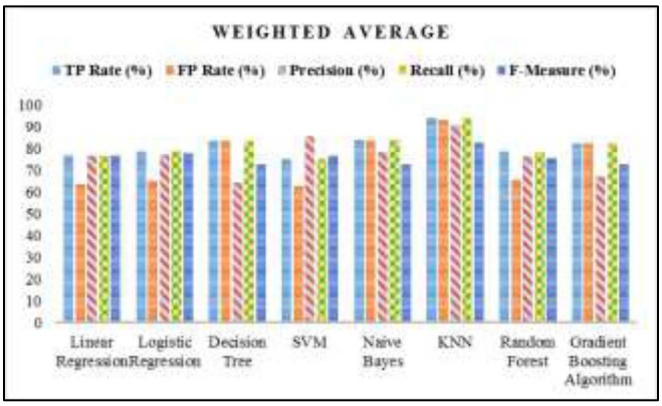

Figure 3: Weighted Average

In this experiment it was found that each classification algorithm shows different accuracy rate. KNN classification algorithm has the highest classification accuracy and the lowest mean absolute error. Table 4 and figure 4 shows the results of road accidents in India. Based on the experimental results, we have observed that from the year 2011 to 2016 most of the accidents occurs in India.

TABLE 4: YEAR WISE ACCIDENTS REPORT IN INDIA

Years	State				
	No.of Accidents				
	1st Place	2 nd Place	3rd Place	4 th Place	5 th Place
2011	Maharashtra	Tamil Nadu	Madhya Pradesh	Karnataka	Andhra Pradesh
	68438	65873	49406	44731	44165
2012	Tamil Nadu	Maharashtra	Madhya Pradesh	Karnataka	Andhra Pradesh
	67757	66316	51210	44448	42524
2013	Tamil Nadu	Maharashtra	Madhya Pradesh	Karnataka	Andhra Pradesh
	66238	63019	51810	44020	43482
2014	Tamil Nadu	Maharashtra	Madhya Pradesh	Karnataka	Uttar Pradesh
	67250	61627	53472	43713	31034
2015	Tamil Nadu	Maharashtra	Madhya Pradesh	Karnataka	Uttar Pradesh
	69059	63805	54947	44011	32385
2016	Tamil Nadu	Maharashtra	Madhya Pradesh	Karnataka	Uttar Pradesh
	69576	62809	54863	44836	31985

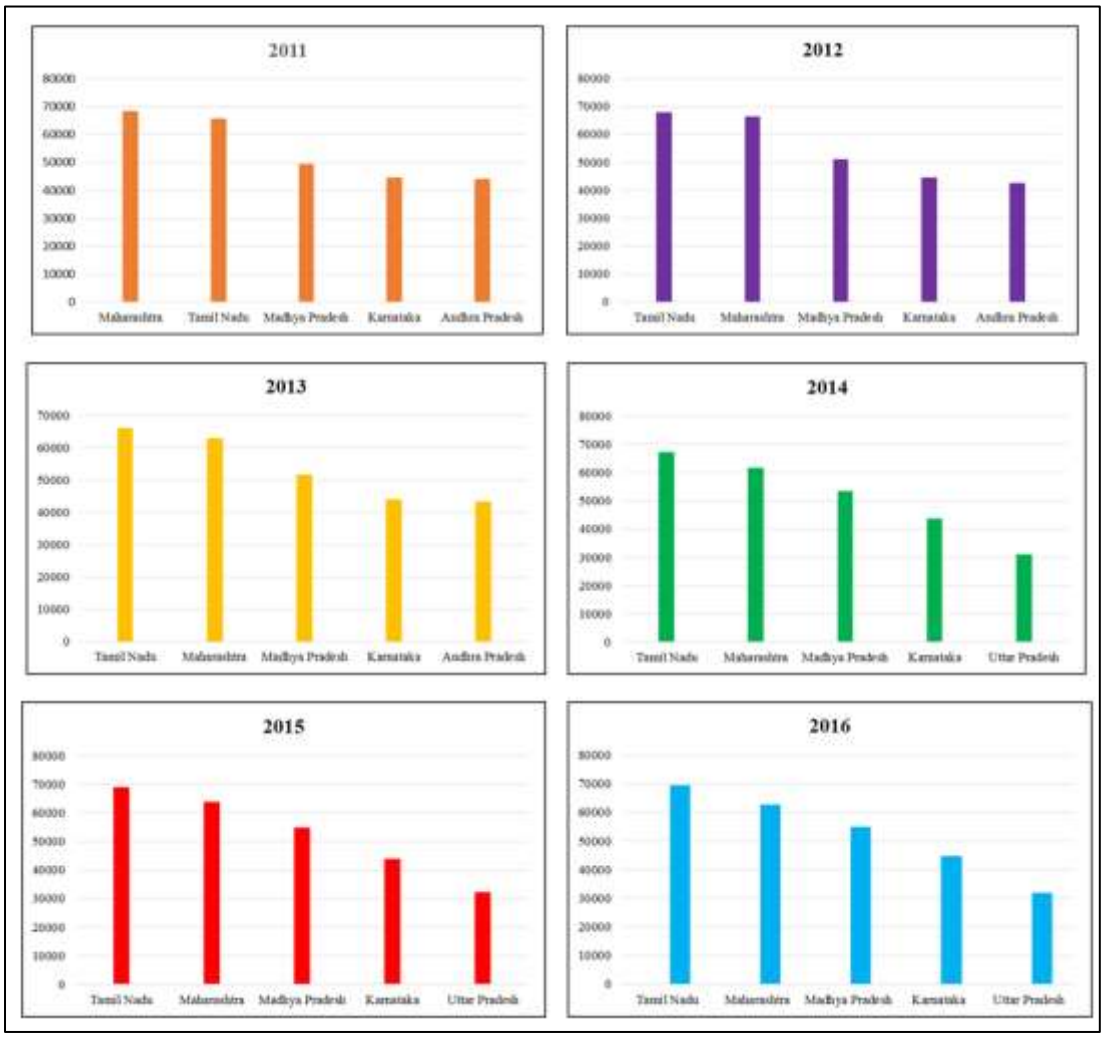

Figure 3: Year wise Accidents Report in India

TABLE 5: TYPES OF ACCIDENT IN TAMILNADU

Years	Accidents			
	Fatal	Major Injury	Minor Injury	Non-Injury
2011	10267	4987	44987	2867
2012	12875	4879	45763	2768
2013	12786	5125	45123	2345
2014	14165	5375	45300	2610
2015	14678	5376	44675	2604
2016	15003	5390	45142	2378

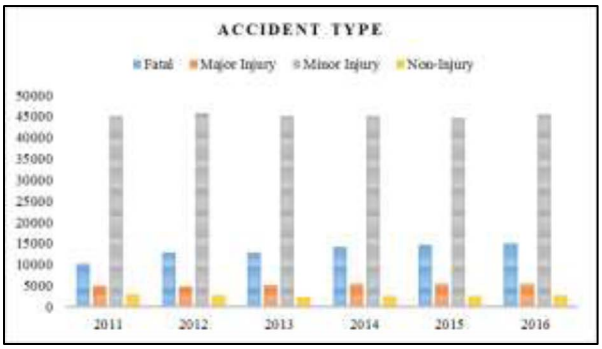

Figure 4: Accident Types

TABLE 6: FACTORS OF ACCIDENT

Attributes	2011	2012	2013	2014	2015	2016
Drunken Drive	1426	1714	4784	6061	6153	1358
Traffic Accident	8857	9264	9375	8987	9235	9578
Careless Driving	7568	7284	7389	7390	7546	7584
Using Mobile Phones	817	485	544	569	756	875
Weather Condition	192	195	195	230	107	115

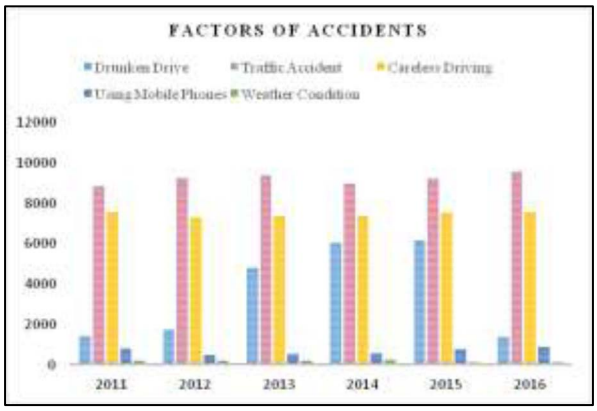

Figure 5: Factors of Accident

Table 5 and figure 4 shows the types of accidents occurs in Tamilnadu. In that, fatal accident and major injury are increased in the year of 2016, minor injury is increased in the year of 2012 and non- injury is increased in 2011. Table 6 and figure 5 shows six factors for the accident. From the table 6, we have observed that the most of the accidents occur at the time of traffic and secondly driving carelessly.

IV. CONCLUSION

The main aim of this study is to evaluate and investigate classification algorithms. The road accident dataset is used to test the performance of the selected classifiers. The algorithm which has the lowest mean absolute error and higher accuracy is chosen as the best algorithm. By considering different parameters of accuracy and the error rate, it is found out that the KNN classification algorithm is the best algorithm with a maximum accuracy of 93.7 than other classification algorithms. Experimental results, road accidents increased in the year of 2016 in Tamilnadu, India.

REFERENCES

[1]Ms S. Vijayarani, Ms M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms" International Journal of Advanced Research in Computer and Communication EngineeringVol. 2, Issue 8, August 2013.

[2] Olutayo V.A, Eludire A.A, "Traffic Accident Analysis Using Decision Trees and Neural Networks", I.J. Information Technology and Computer Science, 2014.

[3] Tibebe Beshah, Shawndra Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road- elated Factors on Accident Severity in Ethiopia",

[4] Meenu Gupta, Vijender Kumar Solanki, Vijay Kumar Singh, "Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review", Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering pp. 197–199, Vol. 10 ISSN 2300-5963

[5] Suwarna Gothane, Dr. M. V. Sarode, "Analyzing Factors, Construction of Dataset, Estimating importance of factor and generation of association rules for Indian road Accident", 2016 IEEE 6th International Conference on Advanced Computing

[6]Priyanka A.Nandurge, "Analyzing Road Accident Data using Machine Learning Paradigms", International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) 2017.

[7] Maninder Singh, Amrit Kaur," A Review on Road Accident in Traffic System using Data Mining Techniques", International Journal of Science and Research, Volume 5 Issue 1, January 2016

[8]https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

[9] http://www.saedsayad.com/naive_bayesian.htm

[10] Ayushi Jain, Garima Ahuja, Anuranjana, Deepti Mehrotra, "Data Mining Approach to Analyse the Road Accidents in India", IEEE explore, January 2016

[11] J. M. Manasa, Shrutilipi Bhattacharjee, Soumya K. Ghosh, and Sudeshna Mitra, "Spatial Decision Tree for Accident Data Analysis", IEEE explore 2014.

[12] S. Shanthi, R. Geetha Ramani, "Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques", Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I.