A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA

Amit Gupta
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

Ali Syed
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

Azeem Mohammad
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

Malka N. Halgamuge
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

***Abstract*—In the last five years, crime and accidents rates have increased in many cities of America. The advancement of new technologies can also lead to criminal misuse. In order to reduce incidents, there is a need to understand and examine emerging patterns of criminal activities. This paper analyzed crime and accident datasets from Denver City, USA during 2011 to 2015 consisting of 372,392 instances of crime. The dataset is analyzed by using a number of Classification Algorithms. The aim of this study is to highlight trends of incidents that will in return help security agencies and police department to discover precautionary measures from prediction rates. The classification of algorithms used in this study is to assess trends and patterns that are assessed by BayesNet, NaiveBayes, J48, JRip, OneR and Decision Table. The output that has been used in this study, are correct classification, incorrect classification, True Positive Rate (TP), False Positive Rate (FP), Precision (P), Recall (R) and F-measure (F). These outputs are captured by using two different test methods: k-fold cross-validation and percentage split. Outputs are then compared to understand the classifier performances. Our analysis illustrates that JRip has classified the highest number of correct classifications by 73.71% followed by decision table with 73.66% of correct predictions, whereas OneR produced the least number of correct predictions with 64.95%. NaiveBayes took the least time of 0.57 sec to build the model and perform classification when compared to all the classifiers. The classifier stands out producing better results among all the classification methods. This study would be helpful for security agencies and police department to discover data patterns and analyze trending criminal activity from prediction rates.**

Keywords*—Data Mining; Classification; Big Data; Crime and Accident

I. Introduction

Technologies provide companies new ways to gather talents of innovators working outside corporate margins. Corporate companies create real prosperity when they combine technology with new ways of doing business and storing data at a standard. There is a need to store data as the Computer technology and the use of Internet has heightened the use of social media such as Facebook and Twitter. The increase in social media urges the need for collecting, storing and processing data for company's development. Analyzing this big data is a challenging process, and therefore the need for certain tools and techniques that are significant in sorting huge amounts of data becomes extremely important. Data Mining is one of the disciplines that is used to convert raw data into meaningful information and knowledge [1]. Data mining searches and analyses large quantities of data automatically by discovering, learning and knowing hidden patterns, trends, and structures [2] and it answers questions that cannot be addressed through simple query and reporting techniques [3]. Data Mining is broadly classified into two categories [4], Predictive Data Mining: that deals with the use of few attributes from a dataset and foretells the future value, or it could also be said that the developing model of the system as per given data. On the other hand, Descriptive Data Mining: finds patterns that describe the data, in other words, presenting new information based on the available dataset trends available.

With the use of new tools and techniques, the offenses and accidents are tracked, monitored and reduced; but at the same time, people are getting more knowledgeable about different crimes and ways to perform them with information available online at their fingertips. The use of technology such as surveillance cameras, speed detection devices, fire and burglary alarms, has helped various monitoring and tracking easier than ever. The types of software that are used today, stores huge amount of data that is collected every day [5]. A particular data set related to crimes and accidents from Denver city, USA has been obtained, and data mining techniques are applied to analyze and find information. The criminal activities and accidents show that there is an increase in death rates in the USA [6]. The major cause of road accidents is drink driving, over speed, carelessness, and the violation of traffic rules [5]. Assessing the cause of crimes is extremely important as it makes taking precarious measures easier.

Education or informing police depends on these assessments. Additionally, the cause of these accidents is only preventable if they are tracked and evaluated to inform police in taking measures for minimizing it and bringing awareness to public. This paper is organized as follows. In Section II, we introduce the dataset and attributes in it, and how the data was collected and pre-processed. It also lists and explains the selected classification algorithms. Section III outlines the results obtained by using two different test methods and also the dataset is analyzed on different criteria's giving us insight on trends and patterns of incidents that have occurred in the due course. Section V concludes the paper.

II. Materials and Methods

This paper has used the predictive method of data mining where the particular attribute value is predicted based on other related attributes. A few classification algorithms: BayesNet, NaiveBayes, OneR, J48, Decision Table and JRip are used in this paper to predict the outcomes of collected statistical data.

A. Data Collection

Data is collected from statistical websites: US City open data census and official government site of Denver city from the year 2011 to 2015, and this data is based on the National Incident-Based Reporting System (NIBRS) where the data is updated every day. This dataset excludes crimes related to child abuse and sexual assault as per legal restrictions law. This Dataset contains 15 attributes and 372,392 instances.

TABLE I. Attribute Description for Classification

Attribute Name	Description
Incident-ID	Unique identification number for a particular incident.
Offense-ID	Unique identification number related to particular Offense.
Offense-Code	Code associated to each offense type
Offense-TypeID	Different types of offenses
Offence-CategoryID	Offenses grouped / assigned into categories.
First-Occurrence-Date	Date incident first occurred on.
Last-Occurrence-Date	Date incident last occurred on.
Reported-Date	Date on which the incident was reported.
Incident –Address	Address of the location where an incident happened.
GeoX	Geographical location
GeoY	Geographical location
District-ID	Name of the district where an incident took place.
Precinct-ID	Precinct name where an incident occurred.
Neighbourhood-ID	Nearby location to the incident
Incident Type	Type of incident (crime/accident)

B. Data Pre-processing

The raw data obtained does not give any information in the form it appears. The raw data stored could contain errors due to multiple reasons like, missing data, inconsistencies that arise due to merging data, incorrect data entry procedures, and so on [7]. Deriving meaningful information from the raw data requires preprocessing of data that converts real-time data into computer readable format. The phases involved in data processing are as shown in Fig. 1.

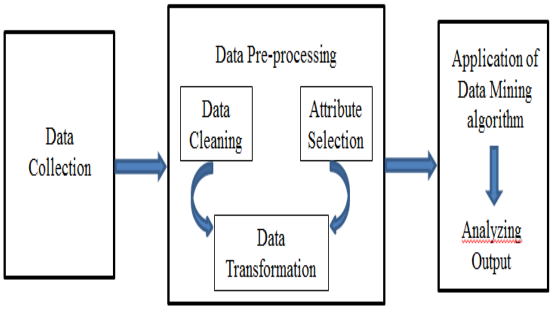

Fig. 1. Data processing of crime and accident dataset obtained for Denver City the USA

The preprocessing is an important phase in data mining. This stage involves the attribute selection, data cleaning, and data transformation [8]. This process starts off with data collection, then the required features or attributes have been selected from the raw data, ready for analysis. Then Data cleaning was performed by eliminating the errors and missing values, with the correction of syntaxes, for example, the address attributes. Finally, the data is prepared and transformed into a suitable and readable format for the data-mining tool to generate.

C. Classification Algorithms

A number of classifications and algorithms are available, and few of them have been selected and used. Below table presents the method used and gives a brief description of the approach and how it is matched with the classifier. The classifiers that are selected are Bayesian, decision trees, and rules based which are outlined in Table 2.

TABLE II. Classification Methods Used in this Study and Description of the Methods

Classifier	Description
NaiveBayes	This supervised learning algorithm is a probabilistic classifier and uses statistical method for each classification.
J48	J48 is an algorithm that generates decision tree using C4.5 algorithms an extension of ID3 algorithm and is used for classification.
JRip	It implements a propositional rule learner called as “Repeated Incremental Pruning to Produce Error Reduction (RIPPER)” and uses sequential covering algorithms for creating ordered rule lists. The algorithm goes through 4 stages: Growing a rule, Pruning, Optimization and Selection [9].
BayesNet	Bayes Net model represents probabilistic relationships among a set of random variables graphically. It models the quantitative strength of the connections between variables, allowing probabilistic beliefs about them to be updated automatically as new information that becomes available. It is a directed acyclic graph (DAG) G that encodes a joint probability distribution, where the nodes of graph represent random variable and arc represent correlation between variables [10].
OneR	A simple classification that produces one rule for each predictor in the data and then the rule with smallest total error is selected [11].

Decision Table	Builds a simple decision table majority classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation.

D. Data Analysis

This study deals with applying the stated classification algorithms in Table 2, to the crime and accident dataset obtained from Denver city, and compared the outputs/results of the classification methods. The analysis is performed based on varied outputs attained from identified number of correct instances and less execution time taken to build the model. The evaluation also helps to gain insights onto which incidents are high in number overall, during a given period of time, and how the trends have been for the last five years.

The software used for this analysis and application of algorithms is Weka (Waikato Environment for Knowledge Analysis, version 3.7). This software allows people to compare different machines to learn algorithms on datasets [11] that contain a collection of visualization tools and algorithms. It is useful for predictive modeling and analyzing data, along with graphical user interfaces for easy access to this functionality [12].

III. Results and Discussions

Results obtained this study are based on different test options: k-fold cross-validation and percentage split criteria.

A. Prediction: k-fold validation

This study has used K-fold cross validation (k=10) method. This method runs the test 10 times, and the first 9 times is used for training, and the final fold is for testing [3] [13], and we have also used the percentage split approach for comparing the outputs and performance of used algorithms. Performances and outputs of each classifier method obtained are compared and presented in Table 3.

TABLE III. Classifiers Accuracy on the Dataset Based on 10-Fold Cross Validation Test Mode

Classification Method	Correctly Classified Incidents	Incorrectly Classified Incidents
NaiveBayes	66.80%	33.19%
Bayes net	68.74%	31.25%
J48	73.54%	26.45%
OneR	64.95%	35.04%
Decision Table	73.66%	26.34%
JRip	73.71%	26.28%

JRip classifier has identified a number of incidents correctly with 73.71%, followed by Decision Table having correct classification rate of 73.66% compared to other classifiers and OneR has determined least correct instances with 64.95%.

TABLE IV. Classifier Execution Time and Root Mean Square Error on the Dataset Based on 10-Fold Cross Validation Test Mode

Classification Method	Time to Build the Model (Seconds)	Root Mean Squared Error
NaiveBayes	0.57	0.460
Bayes net	4.34	0.461
J48	0.87	0.440
OneR	0.81	0.592
Decision Table	18.6	0.435
JRip	21.27	0.440

Execution time is higher for JRip with 21.27 sec and Decision Table with 18.6 sec, while NaiveBayes time to build the model was the least with 0.57 sec, with J48 and OneR time for a model build is 0.87 sec and 0.81 sec, respectively.

There are different performances and measures that are calculated based on the confusion matrix produced by the algorithms. Fig. 2 portrays the model of confusion matrix also known as contingency table. In this matrix, each row exhibits the actual class and column exhibits the predicted class [11].

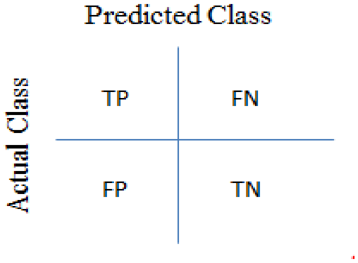

Fig. 2. Confusion Matrix representation

TP (True Positive) and TN (True negatives) are instances correctly classified as a given class and FP (False Positive) and FN (False Negative) are the instances falsely classified as a given class. Other measures are: Precision - % of selected items that are correct and are calculated as Precision (P) = TP / (TP+FP) and Recall - % of correct items that are selected and the calculation for it is Recall (R) = TP / (TP+FN) [14]. With the help of Precision and Recall is calculated F-Measure (F) - the Harmonic mean of precision and recall, calculated as F=2*R*P/(R+P).

TABLE V. Performance Measures Calculated Based on Confusion Matrix Using 10-Fold Cross Validation

Classifier	TP Rate	FP Rate	Precision (P)	Recall (R)	F-Measure (F)
NaiveBayes	66.80%	53.30%	66.50%	66.80%	66.60%
Bayes net	68.70%	55.20%	66.90%	68.70%	67.70%
J48	73.60%	73.60%	54.20%	73.60%	62.50%
OneR	65.00%	12.50%	85.00%	65.00%	66.50%
Decision Table	73.70%	73.30%	68.10%	73.70%	62.70%
JRip	73.70%	73.10%	70.50%	73.70%	62.90%

Above Table 5 shows the TP and FP rate of each classifier, the weighted average of Precision, Recall and F-Measure, obtained by using the 10-fold cross-validation approach.

Decision Table and JRip have the highest TP Rate (True Positive) by 73.7% and Recall values73.7%, followed by J48 having TP rate and recall value of 73.6%. OneR has greater precision when compared to other algorithms.

B. Prediction: Percentage Split

Another test option of split criteria available is also used to compare and evaluate the classifier outputs. In the percentage split method, the algorithm is trained in a certain percentage of data first, and then the learning is tested on the remainder of the data. Table 6 presents the result of classifier output based on split criteria.

TABLE VI. RESULT OF CLASSIFIER ACCURACY BASED ON SPLIT CRITERION TEST MODE

Classifier	Train Data (%)	Test Data (%)	Correctly Classified (%)	Incorrectly Classified (%)
BayesNet	90	10	79.53	20.46
	80	20	78.59	21.40
	70	30	77.63	22.36
	60	40	76.79	23.20
	50	50	75.81	24.18
	40	60	74.63	25.36
	30	70	73.29	26.70
	20	80	72.42	27.57
	10	90	72.00	27.99
NaiveBayes	90	10	75.85	24.14
	80	20	76.18	23.81
	70	30	61.77	38.22
	60	40	61.92	38.07
	50	50	66.03	33.96
	40	60	61.48	38.51
	30	70	68.33	31.66
	20	80	30.04	69.95
	10	90	30.90	60.09
OneR	90	10	65.07	64.92
	80	20	63.02	36.97
	70	30	60.68	39.31
	60	40	57.92	42.07
	50	50	55.11	44.88
	40	60	51.40	48.59
	30	70	47.24	52.75
	20	80	41.93	58.06
	10	90	35.14	65.85
J48	90	10	73.61	26.38
	80	20	73.67	26.32
	70	30	73.62	26.37
	60	40	73.71	26.28
	50	50	73.68	26.31
	40	60	73.70	26.29
	30	70	73.61	26.38
	20	80	73.61	26.38
	10	90	73.64	26.35

Figures 3, 4, 5 and 6 demonstrate the graphical representation of the corresponding classifier output. Figures 3, 4 and 5 indicate Bayes net, NaiveBayes and OneR perform identically. When the percentage of data tested is less the results are more accurate. As the amount of test data increases the percentage of correct classification decreases as a result. This is because a number of data samples trained are less. As seen from Fig 6 it shows that J48 has correctly classified the higher number of instances when the test and trained data is almost equal, and lowest classification rate are when test data is either least or most.

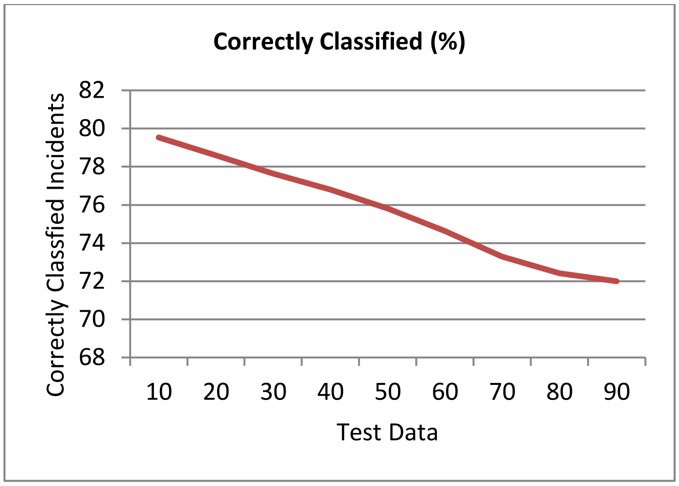

Fig. 3. Bayes net Classification using split percentage test option

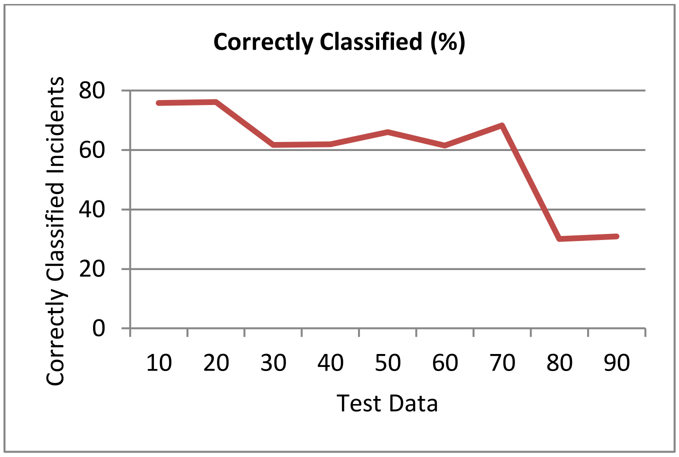

Fig. 4. NaiveBayes Classification using split percentage test option

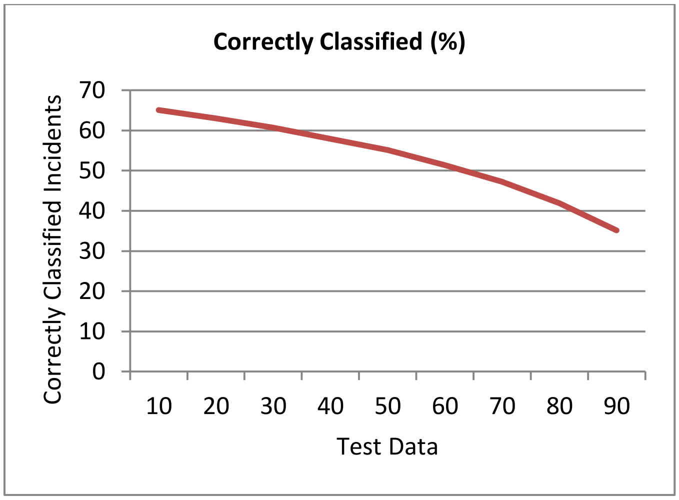

Fig. 5. OneR Classification using split percentage test option

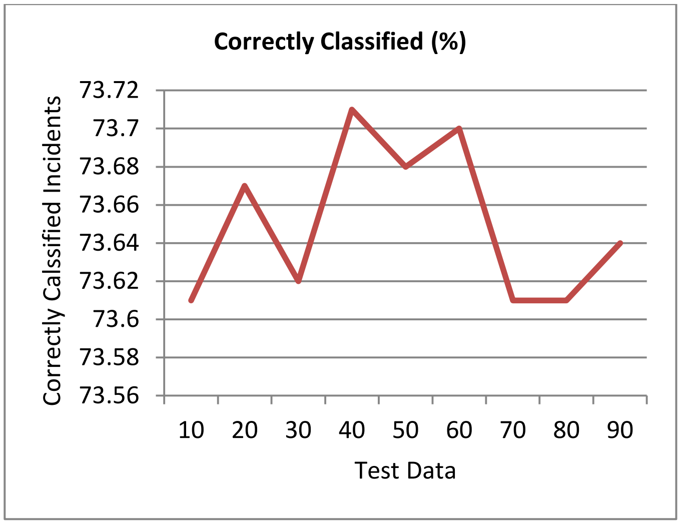

Fig. 6. J48 Classification using split percentage test option

Further analysis of data is performed based on different criteria's.

TABLE VII. CRIME AND ACCIDENT ON WEEKDAY/WEEKEND

	Accident	Crime	Total
Weekday	84,475	189,783	274,258
Weekend	25,106	73,028	98,134
Grand Total	**109,581**	**262,811**	**372,392**

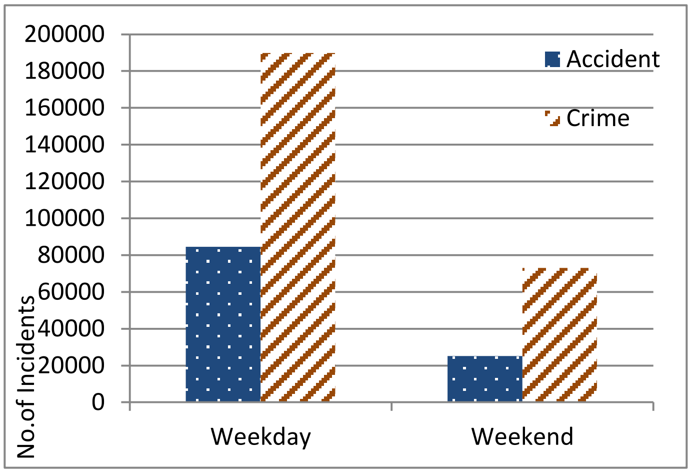

Fig. 7. Crime and accident based on weekday and weekend

TABLE VIII. COUNT OF INCIDENTS ON A MONTHLY BASIS

Month	Crime	Accident	Total
January	24,364	10,525	34,889
February	20,904	10,004	30,908
March	22,010	8927	30,937
April	19,018	8186	27,204
May	20,935	8708	29,643
June	22,085	8781	30,866
July	23,951	8887	32,838
August	24,322	9306	33,628
September	22,833	9203	32,036
October	22,477	9345	31,822
November	20,193	8528	28,721
December	19,719	9181	28,900
Grand Total	**262,811**	**109,581**	**372,392**

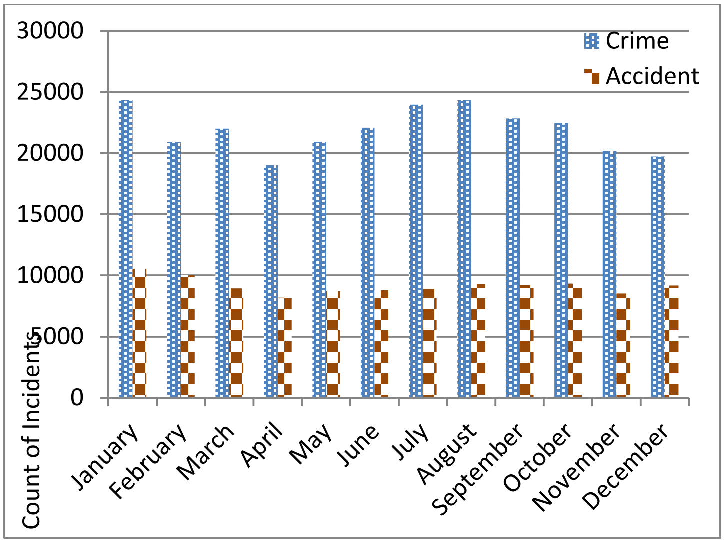

Fig. 8. Count of crime and accidents on a monthly basis

Figure 8 indicates that crime and accidents are more likely to occur during the months of January and February. This is because people start their daily routines after a long vacation of Christmas and New Year. As a result, more public is out in the traffic as people commute and drive to, schools, offices, and work. The trends show an increase of incidents that occur during July and August, as this is the start of the academic year for schools and colleges. During this time, accidents are 60% lower on the weekends when compared to weekdays due to less traffic and crowd on roads. Crime is 60% less on the weekends, as most people stay home relaxing; therefore, crimes such as murder, burglary, and robbery are less likely to occur.

TABLE IX. YEAR-WISE PRESENTATION OF CRIME AND ACCIDENTS

Year	Accident	Crime	Total
2011	20,722	36,419	57,141
2012	19,398	36,258	55,656
2013	19,588	51,820	71,408
2014	21,914	61,340	83,254
2015	23,245	63,632	86,877
2016	4714	13,342	18,056
Total	**109,581**	**262,811**	**372,392**

TABLE X. TYPES OF OFFENSES

Offense Type	No. of Offenses
Murder	210
Arson	533
White-collar-crime	5299
Robbery	5908
Aggravated-assault	8030
Other-crimes-against-persons	13,544
Auto-theft	19,271
Drug-alcohol	21,488
Burglary	24,571
Theft-from-motor-vehicle	32,998
Larceny	40,737
Public-disorder	41,712
All-other-crimes	48,510
Total	**372,392**

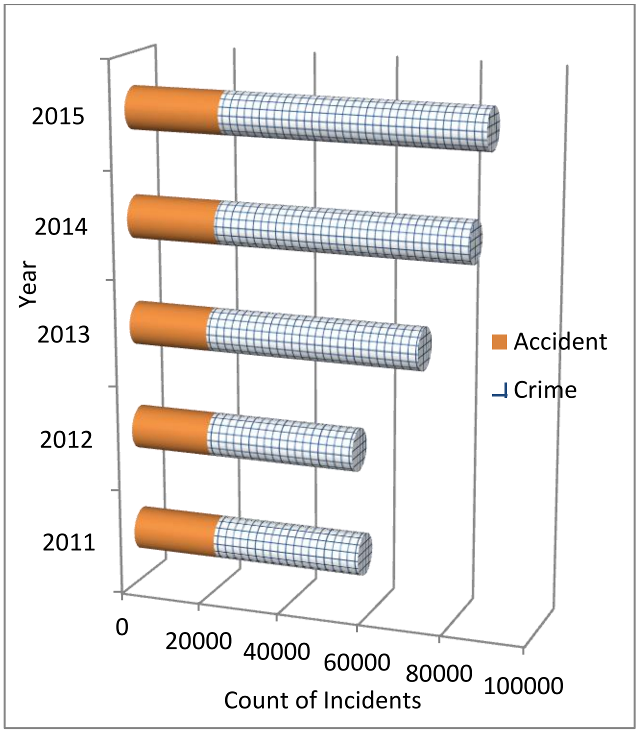

Fig. 9. Number of crime and accidents identified year-wise

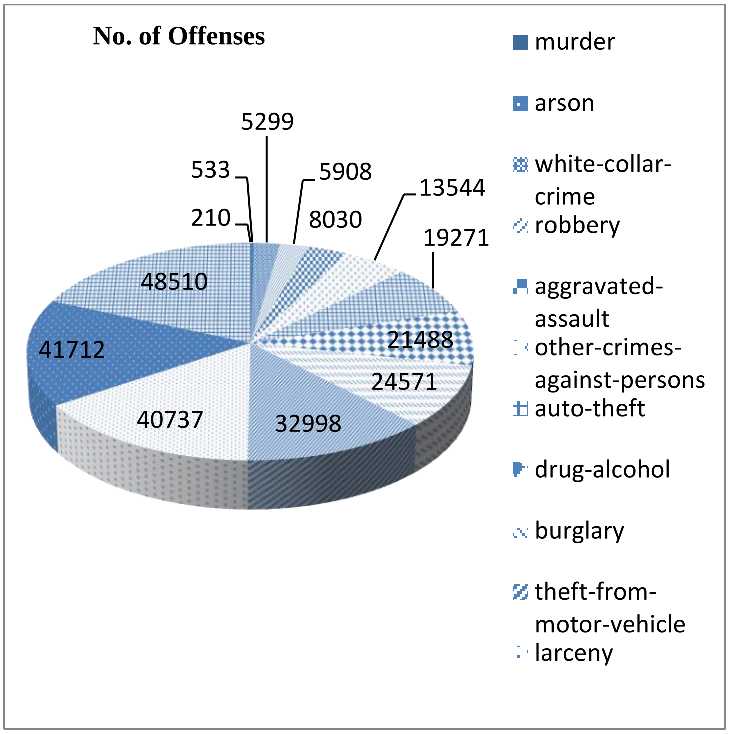

Fig. 10. Different types of offenses indicating number of incidents in each category

TABLE XI. COUNT OF INCIDENTS YEAR-WISE IN EACH OFFENSE TYPE

Offense Category	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**Total**
Aggravated-assault	1314	1467	1522	1599	1755	373	8030
All-other-crimes	1843	1986	9920	15,491	15,589	3681	48,510
Arson	92	92	95	130	107	17	533
Auto-theft	3545	3421	3383	3514	4460	948	19,271
Burglary	4698	4711	4800	4553	4836	973	24,571
Drug-alcohol	1416	1714	4784	6061	6153	1360	21,488
Larceny	5959	6691	8350	9336	8778	1623	40,737
Murder	41	33	39	33	55	9	210
Other-crimes-Against-persons	1286	1427	2617	3649	3840	725	13,544
Public-disorder	6454	5948	8195	9728	9400	1987	41,712
Robbery	1133	1212	1058	1072	1188	245	5908
Theft-from-motor-vehicle	7575	6632	6222	5129	6226	1214	32,998
Traffic-accident	20,722	19,398	19,588	21,914	23,245	4714	109,581
White-collar-crime	1063	924	835	1045	1245	187	5299
Total	**57,141**	**55,656**	**71,408**	**83,254**	**86,877**	**18,056**	**372,392**

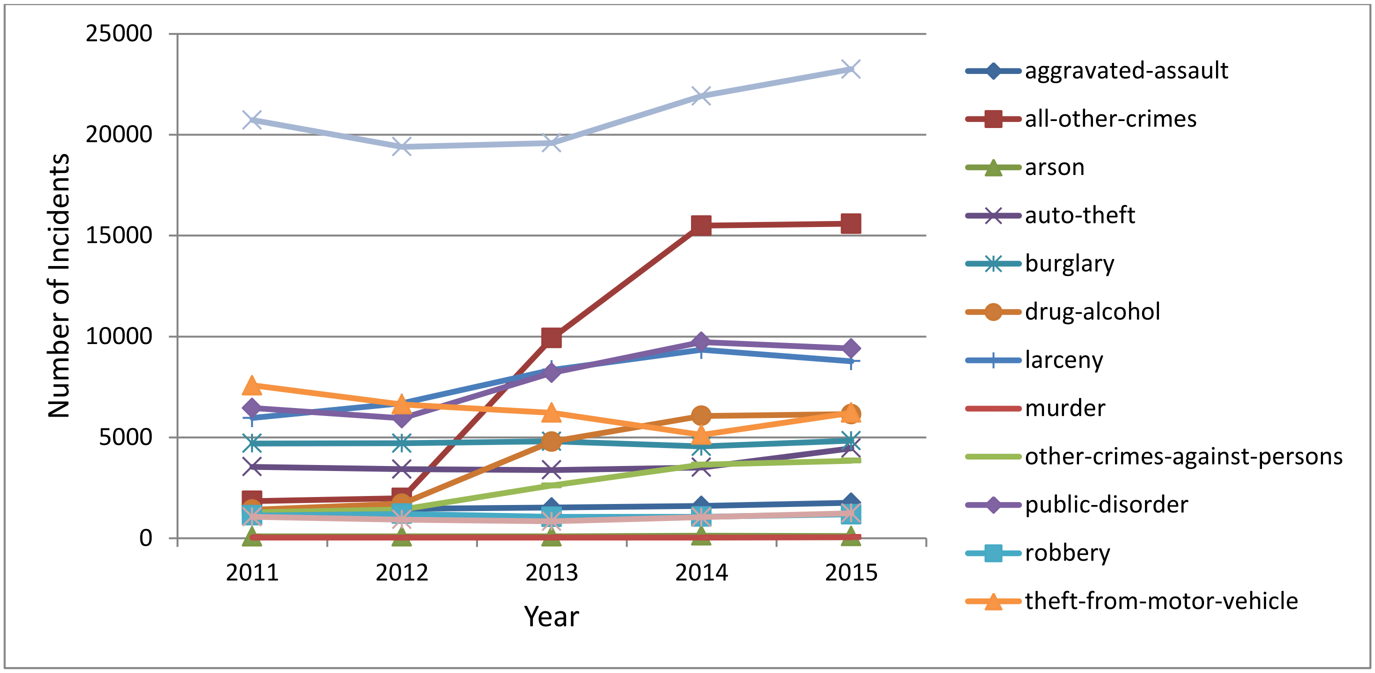

Fig. 11. Number of incidents occurring in each category of offense year-wise

Above Figure 11 shows that drug and alcohol consumption has been increasing year-by-year. In the year 2009, marijuana was legalized in many states of the US, it was allowed on the basis of certain medical conditions. However after a couple of years, it was legalized in Colorado as well. This legalization in 2012 has made the availability of it easier and since then the intake of this drug has been increasing continuously [15]. It is evident from the analysis results as per Fig. 11 from the year 2012-2013 there has been more than 100% increase in drug and alcohol consumption, nevertheless, no strong evidence has found that people consume marijuana truly for medical reasons.

IV. Conclusion

Data Mining techniques and tools have brought tremendous change in the way data is analyzed revealing useful information. This paper has analyzed the application and performance of six classification algorithms that produce different results. Different test methods were used to predict the outcomes for same classification methods. This study has found that various crime patterns have heightened in particular seasons. Results obtained for various classification methods show different outputs and performance measures. Our analysis indicates JRip and Decision Table classified the most number of correct incidents with 73.71% and 73.66%, whereas OneR classified showed the least number of correct incidents with 64.95%. Although JRip is the most accurate classifier, it took the maximum time building the model with 21.2 sec. NaiveBayes model builds the quickest time with 0.57 sec. This study is helpful for various agencies, police department and other organizations aiding them to foresee prediction rate of incidents and develop strategies, plans, and preventive measures for the purpose of crime reduction.

References

[1] J. H. Trevor, R. J. Tibshirani and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2011.

[2] C. C. Aggarwal, *Data Mining: The Textbook*. Springer, 2015.

[3] R. A. El-Deen Ahmeda, M. E. Shehaba, S. Morsya and N. Mekawiea, Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining. In*Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on* IEEE, pp. 1344-1349.

[4] S. Gnanapriya, R. Suganya, G. S. Devi and M. S. Kumar, Data Mining Concepts and Techniques. *Data Mining and Knowledge Engineering*, vol. 2, p. 256-263, 2010.

[5] K. B. Saran and G. Sreelekha, Traffic video surveillance: Vehicle detection and classification. In *2015 International Conference on Control Communication & Computing India (ICCC)* IEEE, pp. 516-521, November 2015.

[6] P. C. Kratcoski and M. Edelbacher, "Collaborative Policing: Police, Academics, Professionals, and Communities Working Together for Education, Training, and Program Implementation", CRC Press: 2015, vol. 25.

[7] S. García, J. Luengo and F. Herrera, *Data preprocessing in data mining*. Switzerland: Springer, 2015.

[8] R. Deb, A. W. C. Liew, Incorrect attribute value detection for traffic accident data. In *Neural Networks (IJCNN), 2015 International Joint Conference* IEEE, 2015, pp. 1-7.

[9] V. Veeralakshmi and D. Ramyachitra, Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset. *Issues*, vol 1, p. 79-85.

[10] Bayes Nets. Retrieved from http://www.bayesnets.com/

[11] I. H. Witten, E. Frank and M. A. Hall, Data Mining: Practical machine learning tools and techniques, 3rd ed., Morgan Kaufmann, 2011.

[12] S. Kalmegh, Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News, February 2015.

[13] C. Sitaula, "A Comparative Study of Data Mining Algorithms for Classification. *Journal of Computer Science and Control System"s*, vol. 7, 29.

[14] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi and A. I. Madbouly, A comparative analysis of classification algorithms for students college enrolment approval using data mining. In *Proceedings of the 2014 Workshop on InteractionDesign in Educational Environments,* 2014, ACM, p. 106.

[15] J. Schuermeyer, S. Salomonsen-Sautel, R. K. Price, S. Balan, C. Thurstone, S. J. Min and J. T. Sakai, Temporal trends in marijuana attitudes, availability and use in Colorado compared to non-medical marijuana states: 2003–11. *Drug and alcohol dependence*, 2014, vol 140, p. 145-155.