A data mining approach for classification of traffic violations types

Nor Aqilah Othman [a,1], Cik Feresa Mohd Foozy [a,2], Aida Mustapha [a,3], Salama A Mostafa [a,4,*], Shamala Palaniappan [b,5], Shafiza Ariffin Kashinath [c,6]

[a] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400 Batu Pahat, Johor, Malaysia
[b] Faculty Science Computer and Mathematics, Universiti Teknologi MARA (UiTM), Segamat, Johor, Malaysia.
[c] Traffic Systems Sdn. Bhd., 30-1, Jln Radin Bagus 3, Sri Petaling, 57000 Kuala Lumpur, Malaysia
[1] qiqiiyla.othman@gmail.com; [2] feresa@uthm.edu.my; [3] aidam@uthm.edu.my; [4] salama@uthm.edu.my; [5] shamalap@uitm.edu.my; [6] shafiza@senatraffic.com.my
*corresponding author

ARTICLE INFO

Article history
Received November 18, 2020
Revised March 3, 2021
Accepted August 23, 2021
Available online November 30, 2021

Keywords
Traffic Violations
Data Mining
Naïve Bayes
Gradient Boosted Trees
Deep Learning
Classification

ABSTRACT

Traffic summons, also known as traffic tickets, is a notice issued by a law enforcement official to a motorist, who is a person who drives a car, lorry, or bus, and a person who rides a motorcycle. This study is set to perform a comparative experiment to compare the performance of three classification algorithms (Naive Bayes, Gradient Boosted Trees, and Deep Learning algorithm) in classifying the traffic violation types. The performance of all the three classification models developed in this work is measured and compared. The results show that the Gradient Boosted Trees and Deep Learning algorithm have the best value in accuracy and recall but low precision. Naïve Bayes, on the other hand, has high recall since it is a picky classifier that only performs well in a dataset that is high in precision. This paper's results could serve as baseline results for investigations related to the classification of traffic violation types. It is also helpful for authorities to strategize and plan ways to reduce traffic violations among road users by studying the most common traffic violation types in an area, whether a citation, a warning, or an ESERO (Electronic Safety Equipment Repair Order).

1. Introduction

When a law enforcement official issues a traffic summons (also known as traffic tickets), it is to inform the motorist, which includes anyone who drives a car, truck, or bus as well as anyone who rides a motorcycle, that they have been stopped [1][2]. At some point or another, the majority of drivers are recommended for a moving infringement because they are speeding, running a red light, or committing some other type of criminal traffic infraction [3]. The results of tickets are not calamitous but at the very least, dealing with a ticket requires an investment of time. The fact that many people do not consider street activity offenses to be wrongdoing, for some strange reason, gives the impression that they are not considered wrongdoing; however, nothing could be further from the truth [4]. Because of the high number of fatalities that can result from traffic offenses around the world, they have become a major source of public concern [5][6]. For some inexplicable reason, a significant proportion of people do not regard street activity crimes to be criminal offenses [7].

Data mining allows the processing of large amounts of historical data and the condensing of that information into valuable information that can be used to construct various models, such as prediction

models, clustering models, and anomaly models [8]. Data mining software such as the RapidMiner allows users to analyze data using various data mining approaches until knowledge is extracted as the information is accessible in the diverse organizations to make the best possible move [3][7].

According to the literature, several studies have been conducted on the demographic and socioeconomic features of criminal offenders from varied backgrounds. However, minimal studies characterize traffic offenders and drivers who receive citation tickets or warnings. Understanding traffic offences is critical because it will allow for the development of more effective prevention and enforcement strategies to reduce these offenses and, ultimately, road accidents on the road [4][9]. Moreover, data mining can be applied in many industries to help improve or forecast many things. For traffic violations, prior research has primarily looked at how well drivers can predict the characteristics of other drivers who are ticketed for traffic violations [5][7].

This study aims to build a comparative model for classifying traffic violation types based on a data mining approach. Traffic violation types are categorized into Citation, Warning, and ESERO (Electronic Safety Equipment Repair Order) that referred to [7]. The classification algorithms to be used include the Naïve Bayes, Gradient Boosted Trees, and Deep Learning algorithms [10][11]. This research is scoped to traffic violation data from Montgomery County between the years 2013 to 2016, and the dataset was extracted from a website called `data.world`. The classification models developed in this paper will be measured for accuracy, recall, precision, and *f*-measure.

For the existing dataset from [11], the data used in this study came from two different sources: the Southwest City Police Department (SWCPD) and the United States Census Bureau in the year 2000. All traffic citation data was obtained from the SWCPD and consisted of all traffic offenses committed between the dates of January 1, 1999, and October 10, 1999, totaling 87,792 traffic violations and 211,689 fines within that time period. In addition to driver demographic parameters (day, date, and time of the violation), the data on these violation occurrences includes information on the types of charges levied against the driver, his or her speed, and the legal speed limit in the area. The dataset consists of the accident day, year, variables, the vehicle involved, and people included. There are 39 qualities chosen from both datasets, and after data cleansing, 573 of accident information causes driver's casualty.

The remainder of this work is organized in the following manner: Section 2 outlines the methodology that was utilized to complete the data mining work, as well as the dataset and the assessment metrics that were employed in the process. Following that, Section 3 summarizes the findings, and Section 4 closes with conclusion and some recommendations for further research.

2. Method

The Knowledge Discovery in Database (KDD) framework is used in this study to classify the different types of traffic violations. The KDD framework is a data mining system that seeks to uncover interesting patterns in the underlying data. KDD is beneficial for a large dataset, and it can process data from the database as indicated by client necessities. KDD also incorporates how information is prepared, what calculations can be applied to obtain a substantial measure of information proficiently, and how the results can be translated and visualized [12][13]. KDD begins with data warehousing, in which a related field is coming from the database. Data warehousing helps set the phase in KDD in two important ways: data cleaning and data access [14].

There are five phases in KDD that need to be implemented to get the results from classification or prediction techniques. The five phases include selection, pre-processing, transformation, data mining, and evaluation [15]. In this research, the data mining part will involve the classification process to predict traffic violation offenders based on the summons issued. The phases in these experiments are shown in Fig. 1 as adopted from the KDD methodology.

Fig. 1. Experimental phases adopted from the KDD methodology [15]

2.1. Data Selection

According to [16], the process of selecting a certain characteristic from an initial dataset that is most relevant to the data mining activities at hand is known as data selection. Because of the removal of irrelevant or repetitive features, the execution time for the data mining operation will be reduced, while the precision will be raised as a result of the process. When it comes to boosting the effectiveness of data mining algorithms, feature selection is crucial since it ensures that only meaningful and beneficial attributes are used. The final collection of features chosen meets key critical requirements for shrinking in terms of overall size [17].

Data gathering is one of the technical stages required and should be taken totally with the goal that it very well may be run and tested later to think about the execution of the order in expectation of criminal traffic offense [18]. The primary reason for this stage is to get an appropriate dataset with proper credit and run the test. In this study, the dataset was extracted from `data.world`. The dataset was contained 35 attributes and 7,700 rows. After completing data preparation using TurboPrep, the dataset consists of 8 attributes and 7,700 rows. The eight attributes are date, description, vehicle type, year, make, model, violation type, and gender. Fig. 2 shows the dataset after data pre-processing.

	Date Of Stop	Description	VehicleType	Year	Make	Model	Violation Type	Gender
2	9/30/2014	DRIVER FAILUF	Automobile	2014	FORD	MUSTANG	Citation	M
3	3/31/2015	HEADLIGHTS ('	Automobile	2003	HONDA	2S	Warning	M
4	9/30/2014	FAILURE TO DI	Automobile	2009	TOYOTA	CAMRY	Warning	F
5	3/31/2015	DRIVER FAILUF	Automobile	2007	ACURA	MDX	Warning	F
6	3/31/2015	STOP LIGHTS ('	Automobile	2003	NISSAN	MURANO	ESERO	M
7	3/31/2015	DRIVING MOT	Automobile	2007	HONDA	CIVIC	Citation	F
8	3/31/2015	DRIVING VEHI	Automobile	2007	HONDA	CIVIC	Citation	F
9	3/31/2015	FAILURE OF IN	Automobile	2007	HONDA	CIVIC	Citation	F
10	3/31/2015	FAILURE TO DI	Automobile	2007	HONDA	CIVIC	Citation	F
11	3/31/2015	PERSON DRIVI	Automobile	2007	HONDA	CIVIC	Citation	F
12	3/31/2015	PERSON DRIVI	Automobile	2007	HONDA	CIVIC	Citation	F
13	3/31/2015	PERSON DRIVI	Automobile	2007	HONDA	CIVIC	Citation	F
14	9/30/2014	FAILURE TO DI	Automobile	2002	TOYOTA	CAMRY	Warning	F
15	9/30/2014	STOPLIGHT INC	Automobile	2002	TOYOTA	CAMRY	Warning	F

Fig. 2. The excerpt of the traffic violation dataset

2.2. Data Pre-Processing

The improvement of data mining cannot be isolated from the fast advancement of data innovation that permits a comprehensive measure of information aggregated in line with the development of data innovation [19]. Mining implies an endeavor to profit from a vast number of fundamental materials. Given the best practice, experts, talented individuals, and individuals who work to discover data in information mining propose some procedure with work process or approach well-ordered easy to expand odds of accomplishment in putting into utilization the examination.

Right off the bat, the dataset got from the site has any sections that are inadequate as missing information, invalid information, or even pointless information. Likewise, additional credits do not apply to the examination in data mining. The information is not significant it is additionally better evacuated because it is nearness can decrease the quality or precision of the data mining later. Data cleaning is essential in every research to detect and remove errors from the raw data [20]. TurboPrep processed the dataset in RapidMiner Tools in the pre-processing data phase. The dataset selected by the operator is read in the RapidMiner tool. Turbo Prep is designed to make data preparation less time-consuming and

difficult [21]. It gives a user interface where a data is continuously visible front and center, so the data can make changes step-by-step and immediately see the results, with an exhaustive run of supporting capacities to get ready so the data for model-building or presentation.

During this process, firstly, data need to choose whether they want to do a prediction or clustering. After that, RapidMiner will display all the details in every attribute in the dataset. In TurboPrep, data can be transformed; for example, rename the attribute, change type, remove the column, and delete all the selected columns from the dataset [22]. Moreover, one more thing is that TurboPrep can replace a missing value in the dataset. The best thing if using TurboPrep is that this tool provides quality measures. It means the user can see at a glance typical data quality problems. They can show the details about the quality measures are calculated in the dataset. The details will show missing value, infinite, IDs, stability, and valid. Users can check the details and then make a data transformation so that all the attributes with a high value of missing value and low stability will be removed from the dataset.

2.3. Classification Algorithms

Graduated boosted trees, Naïve Bayes, and Deep Learning were used in this classification experiment. According to [23], Based on the Bayes theorem, Naïve Bayes is a probabilistic classifier in which all variables or factors are presumed to be independently variable or factor from one another. The algorithm is straightforward to design and performs admirably when dealing with enormous datasets. According to the Bayes Theorem, the probability of $P(A|X) = P(X|A) \times P(A) / P(X) \times P(X)$, where $P(A)$ is the relative frequency of class A samples, and p is increased when $P(X|A)P(A)$ is increased, and p is increased when $P(X|A)P(A)$ is increased [23].

The second classification algorithm used in this experiment is the Gradient Boosted Trees [24]. It is possible to train a boosted decision tree using an ensemble learning method, in which one independent tree corrects the errors of another independent tree. If the first tree makes a mistake, it is corrected by a second one, and so on. The second tree makes a mistake by the first and second trees, and so on. According to [25], boosting is one of the most effective learning concepts to be established in the last twenty years since it may combine a large number of poor learners into a single strong learner with little effort. A gradient boosted decision tree is a classification model that aggregates all tree-based classification models and uses estimations to gradually achieve its prediction outcomes. Boosting may be a nonlinear regression strategy that is adaptive and makes a difference in the precision of trees as they grow in complexity. Improved trees outperform normal trees in terms of accuracy but are slower and less interpretable by humans than standard trees. The Gradient boosting approach is designed to address these concerns.

The latest algorithm used in this research work is Deep Learning, which mimics human intelligence [26], and many recognition problems with huge training samples in numerous representations and high-speed streams benefit from the use of this technique. Deep learning, which is based on base learning technology (particularly, neural networks), can provide a cross-therapy information analysis to allow for better informed treatment decisions.

2.4. Experimental Setup

The software specification will be used to run all the results that had been generated and allow the researcher to get an accurate result. The purpose of building the prediction of traffic violations is by using data mining tools called RapidMiner. RapidMiner is an open-source data mining with the java computer program and stage for data science computer program. It gives a collaborative environment for data planning, deep learning, machine learning, predictive examination, and text mining. Rapid Miner is created in an open center show. RapidMiner combines instruments and appropriateness to supply a user-friendly integration environment of the most up-to-date data mining procedures [27].

2.5. Evaluation Metrics

This section presents the evaluation metrics that need to be applied in this research, for example, accuracy, precision, recall, and F-measure [28]- [30]. Accuracy is defined as the ratio of true positives to the total number of observations in a dataset, while Precision is the ratio of true positives (TP) to all

positives (TP) and false positives (FP) in a sample (FP). It is the true positive (TP) rate, for example, the proportion of positive tuples that are correctly detected, that determines how accurate the recall is. The F-Measure is the weighted average of precision and recall, and this score is calculated by taking both false positives and false negatives in the computation.

3. Results and Discussion

This section presents the comparative results of the classification experiments using Gradient Boosted Trees, Naïve Bayes, and Deep Learning. The results are reported based on accuracy, precision, recall, and *F*-measure. Implementation of the models is described in terms of processes in RapidMiner. In RapidMiner, a process is visualized through a series of connected operators that transform the data for further analysis. On the other hand, operators represent an element that takes input and produces output, such as a function, a formula, or a node. The processes shown in Fig. 5 include feature set, Deep Learning training, cross-validation, model simulator, and explain prediction.

3.1. Naïve Bayes Classifiers

The analysis is performed by using RapidMiner Studio. This research work uses Naïve Bayes in RapidMiner to construct the prediction model. Data was retrieved utilizing the recovery administrator, and information was passed to the administrator named "cross-validation." The set part and discretize operator are used in the pre-processing step. Cross-validation is connected to evaluate and discover the accuracy of the model. The cross-validation operator may be settled; it has two sub-processes testing and preparing.

During the testing and training phase, there is a subprocess for validation. The training model needs to use the sub-processes of validation. After that, the trained model is applied in the testing phase performance also be measured. A Naïve Bayes operator will be used during the cross-validation and testing training phase. The process "Apply model operator" tests the model while "Apply performance operator" evaluates the performance. Fig. 3 shows the processes in Naïve Bayes classification in RapidMiner.

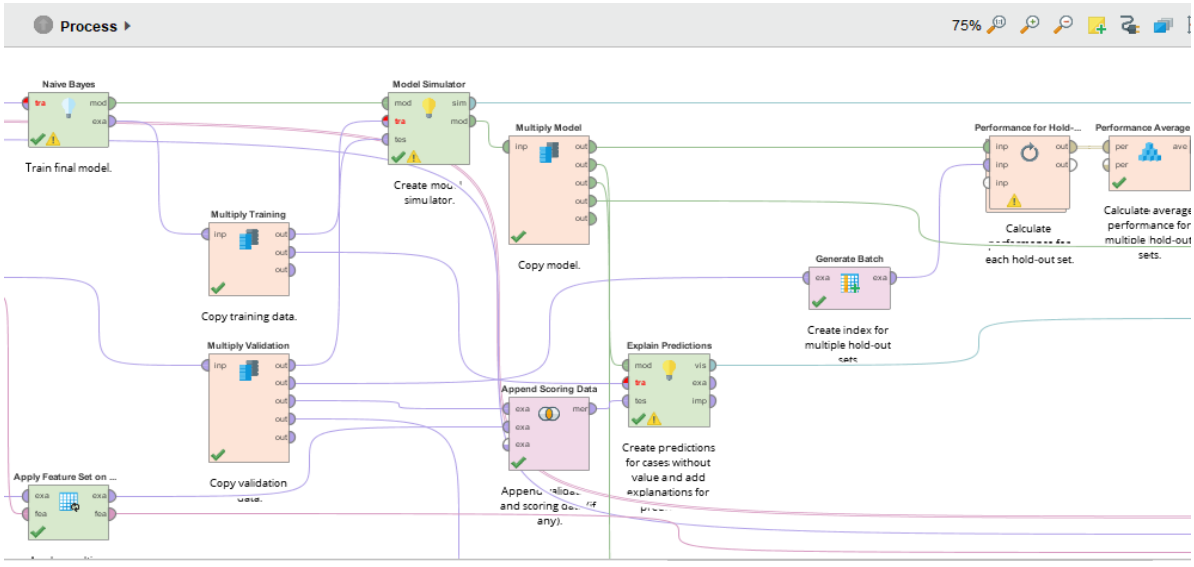

Fig. 3. Process of Naïve Bayes classifier in RapidMiner

RapidMiner performance operator will provide several options to evaluate the performance using the Naïve Bayes model. The classifier accuracy performance for Naïve Bayes was high at 65.75 percent, precision performance was evaluated at 77.24 percent, recall performance was 69.01 percent, and f-measure for this classifier is 72.89 percent.

3.2. Gradient Boosted Trees Classifiers

Fig. 4 shows the Gradient boosted trees classifier by using RapidMiner. The H20 GBT operator is used to predict the traffic violation dataset, which is the dataset that has been used in this experiment. During this phase, the application model and performance operator need to be used to calculate the Gradient boosted trees classifier's performance, accuracy, precision, recall, and *f*- measure.

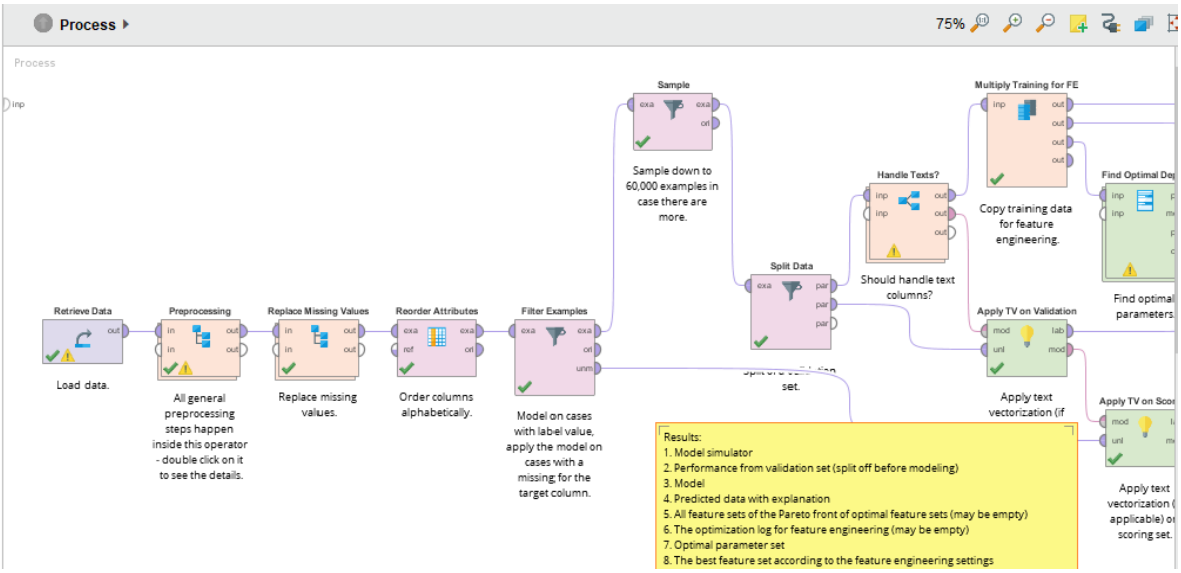

Fig. 4. Process of Gradient Boosted Trees classifier in RapidMiner

The performance of the Gradient Boosted Trees classifier shows that the accuracy is 69.59 percent. The precision performance was evaluated as 70.92 percent, while recall performance was recorded at 91.92 percent, and the f-measure in this experiment was 80.06 percent.

3.3. Deep Learning Classifier

As mentioned before, this research work was proposed to compare all the three algorithms used in this research work. The last algorithm is Deep Learning. Fig. 5 shows the process in RapidMiner.

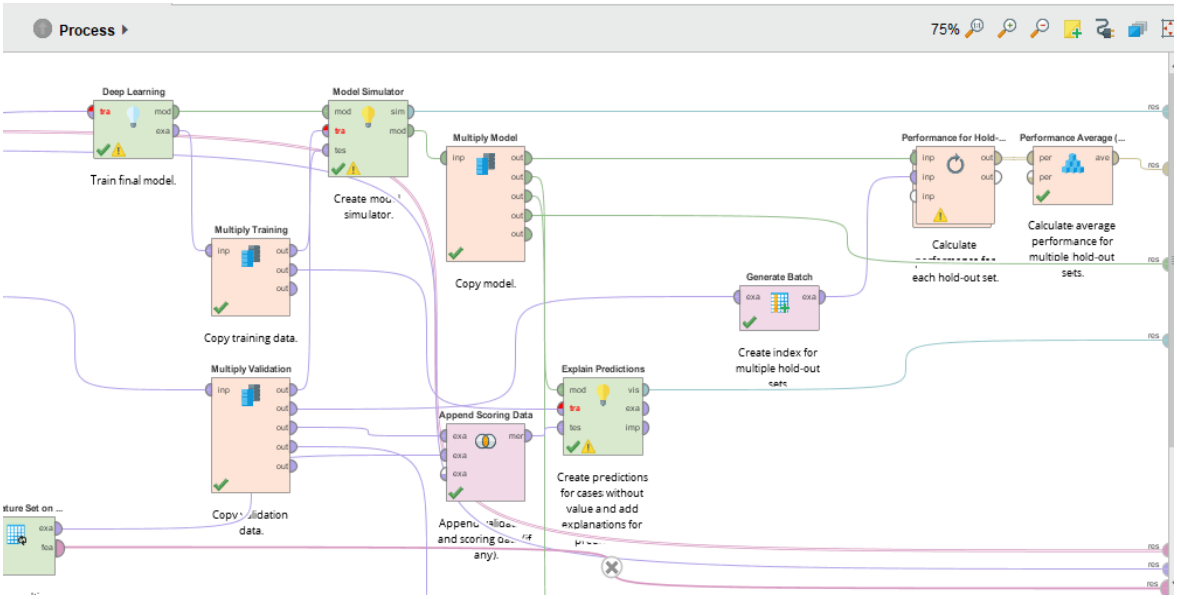

Fig. 5. Process of Deep Learning classifier in RapidMiner

All of the Deep Learning classifier performance was observed. The accuracy is high for 69.22 percent. The precision performance was evaluated at 72.52 percent. Recall of the model was recorded as 87.01 percent, and the *f*- measure in this model is 79.10 percent.

3.4. Performance Comparison

In this project, a traffic violation dataset has been testing the efficiency of the proposed algorithms. In this research work, three methods are chosen to be compared and evaluated. These methods are Naïve Bayes, Gradient Boosted Trees, and Deep Learning. As shown in Table 4, Naïve Bayes got an accuracy of 65.75 percent, while Gradient Boosted Trees accuracy got 69.59 percent, and Deep Learning accuracy was 69.22 percent. The results are shown in Fig. 6.

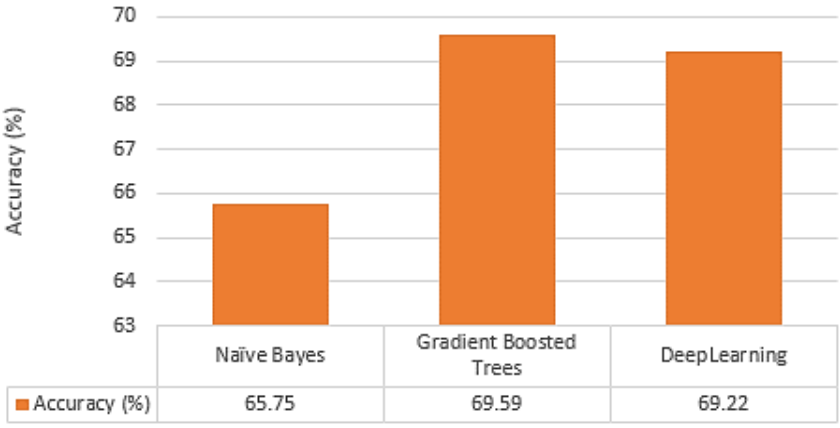

Fig. 6. Comparison between Naïve Bayes, Gradient Boosted Trees, and Deep Learning

Based on Fig. 6, the prediction accuracies of all three predictive models on the dataset were 65.75 percent for Naïve Bayes, 69.59 percent for Gradient Boosted Trees, and 69.22 percent for Deep Learning. Hence, Gradient Boosted Trees based on performance is an efficient classifier with an accuracy of 69.59 percent.

Table 1 shows that all three-model performance was compared. Gradient Boosted Trees is the highest in performance accuracy, the second-highest is Deep Learning, and the lowest is Naïve Bayes. Naïve Bayes is at the highest percentage at precision performances, 77.24 percent, while Deep Learning is 72.52 percent, and the lowest percentage is Gradient Boosted Tree with 70.92 percent. Recall of the model for Gradient Boosted Trees was the higher percentage which is 91.92 percent, and the second-highest is Deep Learning with 87.01 percent and Naïve Bayes with 69.01 percent. For *f*-measure, Gradient Boosted Trees was the highest percentage which is 80.06 percent, while Deep Learning was 79.10 percent and 72.89 percent for the Naïve Bayes model.

Table 1. Performance of Classifier

	Naïve Bayes	Gradient Boosted Tree	Deep Learning
Accuracy	65.75% (+/- 3.08%)	69.59% (+/- 1.09%)	69.22% (+/- 0.59%)
Precision	77.24% (+/- 3.30%)	70.92% (+/- 1.35%)	72.52% (+/- 0.72%)
Recall	69.01% (+/- 2.83%)	91.92% (+/- 0.90%)	87.01% (+/- 1.40%)
f-Measure	72.89% (+/- 2.93%)	80.06% (+/- 1.00%)	79.10% (+/- 0.69%)

According to the findings, the Gradient Boosted Trees and Deep Learning algorithms have good accuracy and recall, but have low precision. As the predicted labels are inaccurate when compared to the training labels, this situation can arise for a variety of reasons. In general, a low precision indicates that the results contain a higher proportion of false positives. With regard to accuracy and precision, Naïve Bayes was shown to have low accuracy and precision but has high recall. Naïve Bayes is a picky classifier that does not process all of the findings and only performs well on high-precision datasets, as

demonstrated in the following example. The challenge, however, is that as the sample data size grows, improving the recall rate becomes more difficult since precision diminishes.

4. Conclusion

Using three classification techniques, including gradient boosted trees, Naïve Bayes, and deep learning, the classification of traffic violation kinds has been successfully accomplished. In performance assessment, Gradient Boosted Trees scores the highest accuracy of 69.59%, Deep Learning scores the second-highest accuracy of 69.22%, and the Naïve Bayes scores the lowest accuracy of 65.75%. In the future, the results of this comparative classification experiment could serve as a standard or as a baseline for the development of classification or prediction models for traffic infractions. Towards the end of the research project, a visualization method might be implemented to illustrate the intensity of activity violations across geographical areas and accident-prone areas. The dataset also plays an essential role in this research work, and finding an excellent dataset is not easy. Therefore, we will move forward to find a high-quality dataset to improve prediction performance in the future. The prediction results can also be further improved through different classifiers such as Support Vector Machines and other Deep Learning models to get better performance and accuracy.

Acknowledgment

This research is supported by Universiti Tun Hussein Onn Malaysia (UTHM) under Tier 1 Grant Scheme Vot H101.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] F. Kamanga, V. Smercina, B. G. Brents, D. Okamura, and V. Fuentes, "Costs and Consequences of Traffic Fines and Fees: A Case Study of Open Warrants in Las Vegas, Nevada," *Soc. Sci.*, vol. 10, no. 11, p. 440, Nov. 2021, doi: 10.3390/socsci10110440.

[2] A. J. Khattak, N. Ahmad, B. Wali, and E. Dumbaugh, "A taxonomy of driving errors and violations: Evidence from the naturalistic driving study," *Accid. Anal. Prev.*, vol. 151, p. 105873, Mar. 2021, doi: 10.1016/j.aap.2020.105873.

[3] N. A. S. Zaidi, A. Mustapha, S. A. Mostafa, and M. N. Razali, "A Classification Approach for Crime Prediction," Khalaf M., Al-Jumeily D., Lisitsa A. Appl. Comput. to Support Ind. Innov. Technol. ACRIT 2019. *Commun. Comput. Inf. Sci.* vol 1174. Springer, Cham., pp. 68–78, 2020, doi: 10.1007/978-3-030-38752-5_6.

[4] R. Factor, "An empirical analysis of the characteristics of drivers who are ticketed for traffic offences," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 53, pp. 1–13, Feb. 2018, doi: 10.1016/j.trf.2017.12.001.

[5] B. Jiang *et al.*, "Transport and public health in China: the road to a healthy future," *Lancet*, vol. 390, no. 10104, pp. 1781–1791, Oct. 2017, doi: 10.1016/S0140-6736(17)31958-X.

[6] A. M. Pérez-Marín and M. Guillen, "Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations," *Accid. Anal. Prev.*, vol. 123, pp. 99–106, Feb. 2019, doi: 10.1016/j.aap.2018.11.005.

[7] S. Thapa and J. Lee, "Data Mining Techniques on Traffic Violations," *Dep. Electr. Comput. Eng. Univ. Bridg. CT*, 2016. Available: Google Scholar.

[8] X. Guo, "Traffic Flow Forecasting Model Based on Data Mining," *Proc. 2016 Int. Conf. Educ. Manag. Comput. Soc.*, pp. 1043–1046, 2016, doi: 10.2991/emcs-16.2016.257.

[9] R. Factor, "Reducing traffic violations in minority localities: Designing a traffic enforcement program through a public participation process," *Accid. Anal. Prev.*, vol. 121, pp. 71–81, Dec. 2018, doi: 10.1016/j.aap.2018.09.005.

[10] N. Boyko, P. Mykhailyshyn, and Y. Kryvenchuk, "Use a cluster approach to organize and analyze data inside the cloud," *ECONTECHMOD An Int. Q. J. Econ. Technol. Model. Process.*, vol. 7, 2018. Available: Google Scholar.

[11] J. R. Ingram, "The Effect of Neighborhood Characteristics on Traffic Citation Practices of the Police," *Police Q.*, vol. 10, no. 4, pp. 371–393, Dec. 2007, doi: 10.1177/1098611107306995.

[12] K. S. Hlaing and Y. M. K. K. Thaw, "Applications, Techniques and Trends of Data Mining and Knowledge Discovery Database," *Int. J. Trend Sci. Res. Dev.*, vol. 3, no. 5, pp. 1604–1606, 2019, [Online]. Available: https://www.ijtsrd.com/papers/ijtsrd26733.pdf.

[13] A. Azevedo, "Data Mining and Knowledge Discovery in Databases," *Adv. Methodol. Technol. Netw. Archit. Mob. Comput. Data Anal.*, pp. 502–514, 2019, doi: 10.4018/978-1-5225-7598-6.ch037.

[14] M. A. O'Reilly, W. Johnston, C. Buckley, D. Whelan, and B. Caulfield, "The influence of feature selection methods on exercise classification with inertial measurement units," *2017 IEEE 14th Int. Conf. Wearable Implant. Body Sens. Networks*, pp. 193–196, May 2017, doi: 10.1109/BSN.2017.7936039.

[15] J. Li *et al.*, "Feature Selection," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Jan. 2018, doi: 10.1145/3136625.

[16] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data Cleaning," *Proc. 2016 Int. Conf. Manag. Data*, pp. 2201–2206, Jun. 2016, doi: 10.1145/2882903.2912574.

[17] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," *2016 6th Int. Conf. - Cloud Syst. Big Data Eng.*, pp. 300–305, Jan. 2016, doi: 10.1109/CONFLUENCE.2016.7508132.

[18] D. Leslie, "Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector," *SSRN Electron. J.*, 2019, doi: 10.2139/ssrn.3403301.

[19] A. Tiron-Tudor and D. Deliu, "Big Data's Disruptive Effect on Job Profiles: Management Accountants' Case Study," *J. Risk Financ. Manag.*, vol. 14, no. 8, p. 376, Aug. 2021, doi: 10.3390/jrfm14080376.

[20] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools," *Int. J. Inf. Technol. Comput. Sci.*, vol. 9, no. 3, pp. 50–61, Mar. 2017, doi: 10.5815/ijitcs.2017.03.06.

[21] O. Adeniji, "Business to consumers (B2C): the effect of machine learning application in telecom customer churn management," Dublin Business School, 2020. Available: Google Scholar.

[22] A. S. Gran, "Automatic machine learning applied to time series forecasting for novice users in small to medium-sized businesses: a review of how companies accumulate and use data along with an interface for data preparation as well as easy and powerful prediction analysis capable of providing valuable insight," 2019. Available: Google Scholar.

[23] T. Hastie, J. Friedman, and R. Tibshirani, "Model Assessment and Selection," *Elem. Stat. Learn. Springer Ser. Stat. Springer, New York, NY.*, pp. 193–224, 2001, doi: 10.1007/978-0-387-21606-5_7.

[24] K. Lan, D. Wang, S. Fong, L. Liu, K. K. L. Wong, and N. Dey, "A Survey of Data Mining and Deep Learning in Bioinformatics," *J. Med. Syst.*, vol. 42, no. 8, p. 139, Aug. 2018, doi: 10.1007/s10916-018-1003-9.

[25] P. Gaur, "Neural networks in data mining," *Int. J. Electron. Comput. Sci. Eng.*, vol. 1, no. 3, pp. 1449-1453, 2012. Available: Google Scholar.

[26] P. S. Patel and S. Desai, "A comparative study on data mining tools," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 4, no. 2, 2015. Available: Google Scholar.

[27] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," 2021, doi: 10.1016/j.jksuci.2021.06.002.

[28] A. Benussi *et al.*, "Classification accuracy of TMS for the diagnosis of mild cognitive impairment," *Brain Stimul.*, 2021. doi: 10.1016/j.brs.2021.01.004.

[29] S. N. M. M. Nafi, A. Mustapha, S. A. Mostafa, S. H. Khaleefah, and M. N. Razali, "Experimenting Two Machine Learning Methods in Classifying River Water Quality," *Khalaf M., Al-Jumeily D., Lisitsa A. Appl. Comput. to Support Ind. Innov. Technol. ACRIT 2019. Commun. Comput. Inf. Sci. vol 1174. Springer, Cham.*, pp. 213–222, 2020, doi: 10.1007/978-3-030-38752-5_17.

[30] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *J. Inform.*, vol. 15, no. 1, p. 45, Feb. 2021, doi: 10.26555/jifo.v15i1.a20111.