International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 72-75, 2019, ISSN:-2319–8656

Kenya Road Accidents Cause Classification Using Bayesian Networks

Raphael Ngigi Wanjiku
School of Computing and Information Technology
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya

Abstract: In Kenya, the number of fatalities from road accidents rise year after year due to various causes. However, these numbers differ year after year and it is very difficult to identify the causation making analysis and management of anti-accident public campaigns difficult. With the use of Bayesian networks, the causal analysis can be probabilistically estimated giving a better analysis and therefore better measures in addressing the underlying causes. This paper utilises data from the Kenya National Transport Safety Authority website which is pre-processed and prepared for use in a Bayesian network model. Thereafter a Bayesian network model is built using 70% of the dataset as the training data and 30% as testing data. The model is developed with the aid of the Weka software utilising a sample of 120 instances from the prepared data with 401 instances. Furthermore, to validate the model, a Naïve Bayes model is developed with the same dataset. The Bayesian network model results in 69.125% accuracy which is lower compared to those given by the naïve Bayes model with 72.5% accuracy possibly due to the fact that Naïve Bayes algorithm performs well even with small amounts of data. Also, from the results, the model identifies that most of the accidents are driver related with 63.8% on the Bayesian network and 78.2% on the Naïve Bayes model and therefore more need to be done in addressing the driver causes. However, more variables need to be introduced in the dataset by the transport agency.

Keywords: Bayesian network, National Transport Safety Authority, normalization, Naïve Bayes, Matatu.

1. INTRODUCTION

The number of people who die on the Kenyan roads is worrying with the year 2018 recording 2965 deaths [6] at the scene of crash. However, many other people die from subsequent effects of the accidents and therefore proper analysis of the causes could alleviate proper campaigns to help reduce the fatalities.

In Kenya, accidents are recorded by the National Transport and Safety Authority which keep daily and monthly reports. The accidents are tabulated in Excel spreadsheets with simple analysis of variations in the monthly reports. This data is then shared among the stake-holders, mostly the Kenya Police Service in ensuring motorists follow the Traffic Act [1]. Despite all the efforts being made, these accidents and especially 2019 have increased by 13% compared to the year 2018. Looking at the previously available data, these accident patterns vary over the years and alternative ways of addressing the causes through technology could be quite beneficial-Bayesian networks in artificial intelligence.

A Bayesian network refer to a probabilistic graphical model that effectively deals with various uncertainty problems. [2] Bayesian Networks are normally used to detect causal relationships for example over speeding and carelessness causing road accidents. They have been used in fault diagnosis [3], Customer satisfaction in public transport management [4], and prediction of criminal cases [5].

Bayesian networks aim at modeling conditional dependence showing causation among variables [7]. They are built from a probability distribution since they work with probabilities. The network is made of nodes and arcs connecting the nodes. Each arc has an arrow that point from a parent variable and the network utilises conditional probability. Given four nodes A, B, C and D as shown in the figure one below, A is considered a parent of A due to the direct causal relationship while D is C is independent of A since there is no direct connection between the two nodes. Each of the node for example B has a conditional probability distribution P (B | Parent(B)).

$$P(D) = P\left(Yi \middle| parent(Yi)\right) \quad (1)$$

$$P(D) = \frac{P(parent(Yi)|Yi)P(Yi)}{P(parent(Yi))} \quad (2)$$

meaning calculating the probability of B given the probability of A (since A is the parent of B).

A Bayesian network utilises a joint probability distribution and the conditional probability. A joint probability distribution of the variables [8] shown in figure 1 is P (A, B, C, D).

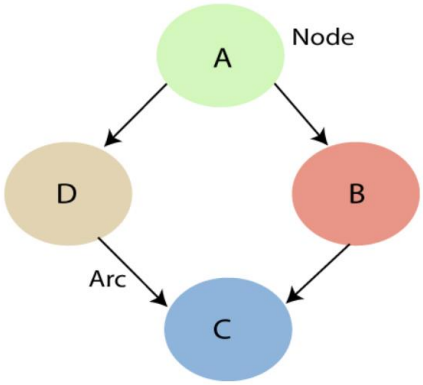

Figure 1. A classification-type decision tree output for this prediction domain.

2. RELATED WORK

There are several literatures works on the utilization of Bayesian networks in evaluating road safety. In a study by Mujalli [10], various Bayes classifiers were used to identify the factors that affect the severity of an accident using Jordan accidents data for the years 2009-2011.They used balanced databases which improved the performance of the classification to the original classification[11]. They were also able to identify other factors that lead to serious injuries or fatalities in accidents for example, the number of vehicles involved, speed limits and the type of accident.

In a different study conducted by De Ona and Mujalli [10], Bayesian networks were used to model accidents classification in Spain (moderate, severe and fatal) and the study concluded that the number of probability variables can be reduced and still maintain the accuracy of the network.

Other works conducted by Zou [2] similarly develop models that address the severity of accidents. Further work by Deublein [8] had shown how Bayesian networks can be used in the prediction of road accidents with case study of Austrian rural motorway network using multivariate regression analysis.

This paper primarily addresses the relationships that exist among the vehicles involved in accidents (car, bus, cycle, tuk-tuk, lorry, other vehicles), time of accident occurrence, gender casualty, pedestrians involved and the Kenyan county most likely to happen.

3. METHODOLOGY

3.1 Accidents database

The study used the NTSA data released every month. The objective was to determine the causes of the accidents and the other related variables. There were difficulties in obtaining complete records hence the 401 instances used do not exhaustively report all the accidents that have occurred with fatalities in Kenya. There were other errors noted in the recording of the data including double entries and mistyped data.

In the study, twenty (20) variables were used to determine the relationships that exist between the various classification of fatalities: Period parameters (year, time),Kenyan county, vehicle involved (car, bus, cycle, tuk-tuk, lorry, matatu, other vehicle, unknown vehicle), cause, gender(male, female), NTSA code, driver, victim involved (passenger, pedestrian, cyclist) and number of casualties.

3.2 Bayesian network modelling

Out of the 401 provided instances, the software selected a sample of 120 instances as the dataset and dividing it into two sets: training data (70%) and the rest (30%) as the test data. The modeling process was aided by the Weka software, a data mining tool developed by the University of Waikato, New Zealand.

The twenty (20) variables were further reduced to fifteen (15) variables and three classification classes were developed: driver, pedestrian and vehicle. The values of the variables were coded into discrete values for computational purposes as shown in the figure below.

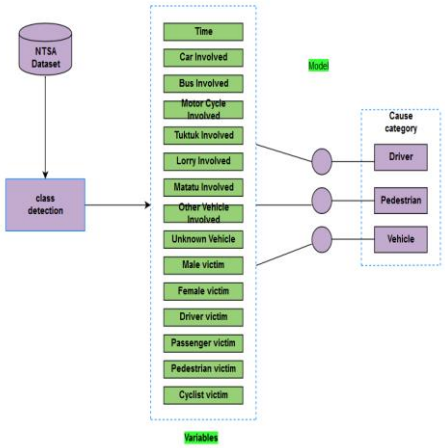

Figure 2. A Bayesian network model on the NTSA data.

The raw data was rescaled to make it suitable for modeling on the Weka software through normalization of the attribute values. Normalization refers to the process of changing the numeric columns in the dataset to use a common scale without distorting differences in the ranges of values or losing information [9]. In this case, all the attributes used have been rescaled to be in the range of 0 and 1.

4. EXPERIMENTS AND RESULTS ANALYSIS

A total of 120 sampled instances gave an accuracy of 69.17% representing 83 correctly classified instances and 30.83% representing 37 instances as wrongly classified which is shown in the confusion matrix in table 1.

Table 1. Confusion matrix on the cause classes

Cause class	Vehicle	Driver	Pedestrian
Vehicle	32	5	0
Driver	29	51	0
Pedestrian	2	1	0

From table 2 below, the precision of the model using the Bayesian net-work shows that it could not classify the accidents resulting from pedestrians' faults while giving a precision of 50.8% for the accidents resulting from vehicle related issues and 63.8% for the driver related accidents.

Table 2. Accuracy of the Bayesian network

Cause class	TP Rate	FP Rate	Precision	F-Measure	PRC Area
Vehicle	0.865	0.375	0.508	0.640	0.551
Driver	0.638	0.895	0.638	0.745	0.838
Pedestrian	0.000	0.000	0.000	0.000	0.025
Weighted Average	0.692	0.215	0.000	0.000	0.729

Table 3. Confusion matrix on the cause classes using the Naïve Bayes network

Cause class	Vehicle	Driver	Pedestrian
Vehicle	18	17	2
Driver	11	68	1
Pedestrian	0	2	1

Table 4. Accuracy of the Naïve Bayes network

Cause class	TP Rate	FP Rate	Precision	F-Measure	PRC Area
Vehicle	0.486	0.133	0.621	0.545	0.705
Driver	0.850	0.475	0.782	0.850	0.861
Pedestrian	0.333	0.026	0.333	0.286	0.140
Weighted Average	0.692	0.215	0.000	0.000	0.729

Table 5. Model variables means and standard deviations

Variable	Vehicle		Driver		Pedestrian	
	Mean	Std	Mean	Std	Mean	Std
Time	0.5874	0.2737	0.5403	0.2962	0.5865	0.3092
Car	0.4436	0.4968	0.4048	0.4908	0.2500	0.4330
Bus	0.0752	0.2637	0.1071	0.3093	0.2500	0.4330
Motor cycle	0.2556	0.4362	0.2341	0.4235	0.0000	0.1667
Tuk-tuk	0.0977	0.2970	0.0476	0.2130	0.0000	0.1667
Lorry	0.2707	0.4443	0.1905	0.3927	0.1875	0.3903
Matatu	0.1654	0.3716	0.1310	0.3373	0.1250	0.3307
Other vehicle	0.0752	0.2637	0.0317	0.1753	0.1250	0.3307
Unknown vehicle	0.0301	0.1708	0.2183	0.4131	0.1250	0.3307
Male	0.2462	0.1371	0.2054	0.1032	0.2188	0.0827
Female	0.0739	0.0186	0.0344	0.0706	0.0208	0.0551
Driver	0.2331	0.4228	0.0437	0.2043	0.0000	0.1667
Passenger	0.3910	0.4880	0.1548	0.3617	0.5000	0.5000
Pedestrian	0.1128	0.3163	0.5675	0.4954	0.5000	0.5000
Cyclist	0.3233	0.4677	0.2421	0.4283	0.0000	0.1667

From table 4, the Naïve Bayes network performed a better classification of test data giving a 33.3% for accidents caused by pedestrians and 78.2% for the driver caused accidents performing better compared to the Bayesian network which gave a 63.8% for the same sampled dataset.

The same data modeled using a Naïve Bayes gives an accuracy of 72.5% (87 instances) wrongly classifying 27.5% (33 instances).

5. CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORKS

In this paper, a model showing causation of Kenyan accidents fatalities selected in the years 2017 to 2019.The methodology utilises a Bayesian model whose performance(precision) is compared with the naïve Bayes model.

From the results, a naïve Bayes gives a better precision for the classification giving a 72.5% compared to 69.125% given by the Bayesian network. This is possibly due to the fact that it requires less data compared to the Bayesian network with the selected sample of 120 incidents[12].

For future work, more data is needed to supplement the selected sample: the available data was scanty and not compatible for utilization in machine learning from the source Agency's website. This led to a lot of preprocessing in order for it to give the obtained results. Furthermore, there is need for addition of more variables during collection of the data since most of the variables obtained were general causations of fatalities and for better usable models to be achieved there is need to look into other factors for example, the terrain where the accidents occurred, road conditions and weather conditions.

6. REFERENCES

[1] Act Title: Traffic.2012. Kenya Law-Laws of Kenya. Retrieved from http://kenyalaw.org:8181/exist/kenyalex/actview.xql?actid=CAP.%20403.Accessed on 20th December,2019.

[2] Cai, B., Liur, Y., Liu, Z., Chang, Y and Jiang, R. 2020. Application of Bayesian Networks in Reliability Evaluation. Bayesian Networks for Reliability Engineering. Springer, Singapore.

[3] Cai, B., Huang, L. and Xie, M. 2017. Bayesian networks in fault diagnosis. IEEE Trans. Ind. Inf. 13(5), pp.2227–2240.

[4] Chakraborty, S., Mengersen, K. and Fidge, C. 2016. A Bayesian Network-based customer satisfaction model: a tool for management decisions in railway transport. Decis. Anal. 3, 4 doi:10.1186/s40165-016-0021-2.

[5] Chao, W., Xin, L., Zhunchen, H., Yakun and M., Wenjia.2019. Interpretable Charge Prediction for Criminal Cases with Dynamic Rationale Attention. Journal of Artificial Intelligence Research. 66. 743-764.10.1613/jair.1.11377.

[6] Daily Nation.2019. WHO: Kenya road deaths four times higher than NTSA reported. Retrieved from https://www.nation.co.ke/news/Kenya-road-deaths-

grossly-underreported–WHO/1056-4893792-ve7d07z/index.html. Accessed on 16th December,2019.

[7] Devin, S.2018. Introduction to Bayesian Networks. Retrieved from https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e. Accessed on 19th December,2019.

[8] Deublein, M., Schubert, M., Adey, T., Kohler, M. and Faber, H. 2013..Prediction of road accidents: A Bayesian hierarchical approach. Accident Analysis and Prevention.

[9] Microsoft. 2019. Normaliza Data. Microsoft. Retrieved fromhttps://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data.Accessed on 19th December,2019.

[10] Mujalli, R., O., G., López and L., Garach. 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accident Analysis and Prevention v. 88, p. 37-51.

[11] Nils, J., S. 2018. Artificial Intelligence-Bayes Network. Retrieved from https://www.norwegiancreations.com/

2018/09/artificial-intelligence-bayes-network/. Accessed on 19th December,2019.

[12] Richante.2014. What is the difference between a Bayesian network and a naïve Bayes classifier? Retrieved from https://stackoverflow.com/questions/12298150/what-is-the-difference-between-a-bayesian-network-and-a-naive-bayes-classifier.Accessed on 20th December,2019.