Predicting Factors of Vehicular Accidents using Machine Learning Algorithm

Aklilu Elias Kurika¹, Irfan Ahmad Ganie², Yuliyanti Kadir³, Patrick D. Cerna⁴, Frice L. Desei⁵

¹ Lecturer, Department of Information Technology, Wolaita Sodo University, Ethiopia, akliluelias123@gmail.com
² Irfan Ahmad Ganie, Department of Electrical Engineering, Indian Institute of Technology Jodhpur, India, ganieirfan27@gmail.com
³ Yuliyanti Kadir, Assistant Professor, Department of Civil Engineering, Universitas Negeri Gorontalo, Indonesia
⁴ Professor, Technology, Engineering and Research, PSHS-CRC, Philippines, pcrerna@ieee.org
⁵ Assistant Professor, Department of Civil Engineering, Universitas Negeri Gorontalo, Indonesia

ABSTRACT

Vehicle traffic accident is one of the major agenda for the government in which special attention has been given to continuously reduce its occurrence and related risks. Wolaita zone is one of the major areas in which increased vehicle traffic accident occurs. Government and concerned bodies have given special attention to reduce accident rate in the country. By having this point as the motivating factor for study, this work tried to predict factors of vehicle accidents by using machine learning algorithms. We used unbalanced datasets with 1611 instances, which was seven years data from year 2012-2019. In order to analyze data and evaluate patterns of datasets, KDD process model was applied. The learning algorithms applied for experiments were J48 decision tree, Random forest tree, Rep tree, Naïve Bayes and Bayesian network classifiers. The experimental results, model evaluation and performance measurement shows that F-measure of J48 and Rep tree classifiers are comparatively similar i.e. 97.87% and 97.80% respectively and Random Forest tree performed less i.e. 90.9%. We identified the first experiment of J48 tree as the best model by performance and 23 best rules were generated from this experiment; best features were also identified. The most common victims, most commonly participated vehicles in accident and black spot areas for frequent accidents occurrences were identified. The findings of this study are significant for road and traffic authority and police commission for the revision and endorsement of the rules, regulations and standards related to traffic accidents; and therefore vehicle traffic accidents and related risks can be reduced generally in our country Ethiopia and specially at Wolaita Zone. We made accident data ready for further analysis in order to get most important patterns of datasets for any future researchers.

Key words: About four key words or phrases in alphabetical order, separated by commas.

1. INTRODUCTION

Road or vehicle traffic accident is a universal problem [1] and worldwide reports show that on average, more than four million peoples die because of many reasons in one year. Among this numbers, HIV AIDS and tuberculosis are the first and second cases for the deaths and vehicle traffic accident is the third known case for those dying on every day.

According to WHO and World Bank [2] in 2004, World Health Day, organized by the World Health Organization for the first time be devoted to Road Safety. Every year, according to the statistics, 1.2 million people are known to die in road accidents worldwide. The study conducted on Guardian [3] also shows that in the 2020, vehicle traffic accident will become the first factor that causes the death of human beings in the world. More than half the people killed in vehicle traffic crashes are young adults aged between 15 and 44 years often the breadwinners in a family. Furthermore, road traffic injuries cost low income and middle-income countries between 1% and 2% of their gross national product; more than the total development aid received by these countries WHO and World Bank [2]. A lot of researches were conducted on accidents from time to time in every parts of the world to reduce the accident rate and they used their own view on accident data according to their respective areas and country perspectives.

Even though plenty of researches were conducted, vehicle traffic accident increases rapidly and results in massive loss of humans’ life, materials damage and other equivalent losses. WHO and World Bank [2] show that worldwide, an estimated 1.2 million people are killed in road crashes each year and as many as 50 million are injured. Projections indicate that these figures will increase by about 65% over the next 20 years unless there is new commitment to prevention. The increased loses and related injuries cause various problems to the economic development of respective countries. According to the perspectives of different countries, there are different kinds of attributes and contributing cases of the traffic accidents. The accident risk factors are more over determined in the developed countries and some preventive measures have been taken to reduce the risk. But traffic accident risks, related material damages and life lose increases from time to time in developing countries. In Ethiopia, some researches has been conducted but the risk factors cannot be reduced from time to time. In the case of
In Ethiopia, Wolaita zone is one of the most commonly known areas in which traffic accidents and related injuries take place. By analyzing the factors with learning algorithms, the most contributing factors will be determined from traffic accident data which is obtained from WZPC. Other contributing factors other than these might also be obtained for increased traffic accidents. The methodologies used by various researchers are of various types. Akinbola et al., [10] and [11] machine learning algorithms to predict the factors of traffic accidents. Both of these authors used only decision tree; and Tibebè et al., [12] is all about machine learning algorithm but it is not for determining the causes of traffic accidents and Gupta and Baluni [7] also used classification and machine learning algorithms to determine traffic injury occurrences.

2. RELATED WORKS

Studies [5] and [4] are related to the locations of accident related factors; accordingly the road features are one of the contributing factors of traffic accidents. But the types of road features are not clearly specified in these studies.

Studies performed by authors [6] and [7] are comparative analysis in the performance measurement and accuracy of algorithms. The first author compared six algorithms (classification and regression tree, Random Forest, ID3, Functional trees, Naïve Bayes and J48) algorithms to determine the accidents severity level. It reveals that Naive Bayes value and J48 techniques value are approximately same in accuracy. The second one the comparative study on machine learning algorithms; the comparison has been made for decision tree and neural networks to determine factors of increased traffic injury. It comes up with that the decision trees are better than neural networks in performance.

Studies conducted by researchers explained in references [6], [7], [8] identified the factors of traffic accidents; their findings show that the causality factors are un-adapted speech, in-attention, behavior of passengers, roadway features, demographic features, environmental characters, technical characters, speed, age, gender, younger aged drivers, alcohol, less control, wrong overtaking and tire blow. These factors were identified in various areas as the contributing factors for the accidents. But it is impossible to blindly take control over all these characteristics to be considered in particular area. So accident factor analysis is needed to identify the most commonly contributing factors that hold a lion share of the commonly known determinant attributes. Some of the factors are common in one area and some other factors become common in other areas. While [9] used social media data as the primary data for predicting the causes of accident; secondary data is not suitable for analysis.
3.2 Data Selection
In order to get data for prediction, applicable data was selected from 12 districts and three city administrations of Wolaita Zone. The case study is limited to Wolaita zone only. This is because we wanted to define the scope of our study.

3.3 Data Preprocessing:
In this step the data cleaning, data reduction and data transformation has been made to prepare the best quality datasets for further analysis. The original data was obtained from Wolaita Zone police commission (PC) but, it has a lot of drawbacks such as spelling errors, unreadable data, misspelt attributes names, unknown values for some attributes and irrelevant personal representations of some terms. Some terms were inconsistent and considered to be outliers. We removed irrelevant attributes from the original Data. In this step we made the cleaning process of data before loading it to WEKA.

3.4 Data Transformation:
The original data was recorded in word processor while some data were in spreadsheet. The researcher transformed it to a .svc format which the weka workbench can read and supported.

Be aware of the different meanings of the homophones “affect” (usually a verb) and “effect” (usually a noun), “complement” and “compliment,” “discreet” and “discrete,” “principal” (e.g., “principal investigator”) and “principle” (e.g., “principle of measurement”). Do not confuse “imply” and “infer.”

Prefixes such as “non,” “sub,” “micro,” “multi,” and “ultra” are not independent words; they should be joined to the words they modify, usually without a hyphen. There is no period after the “et” in the Latin abbreviation “et al.” (it is also italicized). The abbreviation “i.e.,” means “that is,” and the abbreviation “e.g.,” means “for example” (these abbreviations are not italicized).

An excellent style manual and source of information for science writers is [9].

4. EXPERIMENTATION

4.1 Most Prone Accident Vehicles
From the total 31 kinds of vehicles participated in accidents, we have identified 7 kinds vehicles as the most commonly participated. They account 75.34% and remaining 24 vehicles participation is only 24.66%. So we can conclude that if these vehicles were given separate road in cities specially Sodo-City (>25%) traffic accident can be possibly reduced.

4.2 Most Common Victims of Accidents
The above diagram shows that the most common victims of accidents are pedestrians (40.16%) and passengers (19.93%). Derives are less victims. So we can conclude that car traffic accident most commonly affects pedestrians and passengers in our case study. Males (53.8%) are most commonly affected by car traffic accidents compared to females (19.6%); which are opposite to study by [22] that revealed majority of participants as females in accidents. 18.75% of victims were aged between 1-18, 30.54% were aged between 19-30 and 18.56% were aged between 31-50.

As it is known, the most productive human power is aged between 18 and 50. Therefore traffic accident affects the most productive classes of humans as we can conclude from the above result.
### 4.3 Most Common Black Spot Areas

We have selected 19 places with frequent accident occurrences from the above five Woredas. We selected areas with \( \geq 15 \) accidents within 7 years. From the total accidents occurred, these places account 521 (32.34\%) accidents. So concerned bodies has to give attention to these areas.

**Figure 4:** Most Common Black Spot Areas

From 15 different areas shown above, the first five (Sodo-city, Damot-Gale, Humbo, Sodo-Zuria and Boditi-City) account a lot accidents i.e. 73.37\% of total accidents. The remaining 10 districts account only 26.63\%. Each of them accounts > 5\% accident occurrences from the total one, so we selected the black spot areas for frequent accidents occurrences from these five Woredas.

### 4.4 Determinant Cases of Accidents

The Most Determinant Cases and causality condition of Accidents are: Lack of attention (65.49\%), over speed (10.62\%), Prohibiting Priority (10.37\%), lack of experience (6.33\%) and technic failure (3.54\%). The causality condition of accidents is mostly crossing the road (32.96\%) straight crash (28.80\%), roll down (16.70\%), side to side crash (8.57\%) and walking on the road (5.90\%).

**Figure 5:** Determinant Cases of Accidents

As we can see from the above experimental results and below diagram, J48 and Rep tree classifiers are comparatively similar by their accuracy. We computed average Precision and Recall of J48 and Rep tree and selected the J48 decision tree algorithm as a better than Rep tree. 1st Expt J48 tree Precision = 98\% and Recall = 97.75\%.(FM= 97.87\%) 1st Expt. Rep tree Precision = 97.70\% and Recall = 97.90\%.(FM= 97.80\%). The first experimental results of J48 decision tree, includes more features than exp.2 and 3 even though the number of leaves and size of tree generated are more. So we selected it as a working model and generated 23 best rules from this particular experiment.

### Table 1: Summary of Experimental Results

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Models</th>
<th>NL</th>
<th>ST</th>
<th>TP Rate</th>
<th>PP Rate</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp.1</td>
<td>Trees,J48 -C 0.5M-4 Testmode-10-fold Dataset=Unbalanced Attributes=All</td>
<td>141</td>
<td>145</td>
<td>0.994</td>
<td>0.030</td>
<td>0.964</td>
<td>99.45%</td>
</tr>
<tr>
<td>Exp.2</td>
<td>Trees,J48 -C 0.5M-4 Testmode-Split=66% Dataset=Unbalanced Attributes=All</td>
<td>4</td>
<td>5</td>
<td>0.989</td>
<td>0.015</td>
<td>0.989</td>
<td>99.96%</td>
</tr>
<tr>
<td>Exp.3</td>
<td>Trees,J48 -C 0.5M-4 Testmode-Split=99% Dataset=Unbalanced Attributes=All</td>
<td>4</td>
<td>5</td>
<td>0.989</td>
<td>0.0014</td>
<td>0.989</td>
<td>99.93%</td>
</tr>
<tr>
<td>Exp.4</td>
<td>RandomForest -P 100-I 100-nm-slots 1-K 0-M 1.6-V 0.001-S 1 Testmode-10-fold Dataset=Unbalanced Attributes=All</td>
<td>-</td>
<td>-</td>
<td>0.921</td>
<td>0.237</td>
<td>0.926</td>
<td>92.12%</td>
</tr>
<tr>
<td>Exp.5</td>
<td>RandomForest -P 100-I 100-nm-slots 1-K 0-M 1.6-V 0.001-S 1 Testmode-Split=66% Dataset=Unbalanced Attributes=All</td>
<td>-</td>
<td>-</td>
<td>0.905</td>
<td>0.325</td>
<td>0.914</td>
<td>90.51%</td>
</tr>
<tr>
<td>Exp.6</td>
<td>RandomForest -P 100-I 100-nm-slots 1-K 0-M 1.6-V 0.001-S 1 Testmode-Split=99% Dataset=Unbalanced Attributes=All</td>
<td>-</td>
<td>-</td>
<td>0.901</td>
<td>0.361</td>
<td>0.912</td>
<td>90.66%</td>
</tr>
<tr>
<td>Exp.7</td>
<td>trees,REPTree-M 2-V 0.001-N 3.5-L 1-1:0.0 Testmode-10-fold Attributes=All Dataset=Unbalanced</td>
<td>4</td>
<td>5</td>
<td>0.994</td>
<td>0.026</td>
<td>0.984</td>
<td>99.38%</td>
</tr>
<tr>
<td>Exp.8</td>
<td>trees,REPTree-M 2-V 0.001-N 3.5-L 1-1:0.0 Testmode-10-fold Attributes=All Dataset=Unbalanced</td>
<td>4</td>
<td>5</td>
<td>0.989</td>
<td>0.015</td>
<td>0.989</td>
<td>99.90%</td>
</tr>
<tr>
<td>Exp.9</td>
<td>trees,REPTree-M 2-V 0.001-N 3.5-L 1-1:1.0 Testmode-Split=90% Attributes=All Dataset=Unbalanced</td>
<td>4</td>
<td>5</td>
<td>0.991</td>
<td>0.003</td>
<td>0.991</td>
<td>99.07%</td>
</tr>
<tr>
<td>Exp.10</td>
<td>Bayes,NaiveBayes-output-deluge-info Testmode-Split=90% Attributes=All Dataset=Unbalanced</td>
<td>-</td>
<td>-</td>
<td>0.946</td>
<td>0.068</td>
<td>0.948</td>
<td>94.60%</td>
</tr>
<tr>
<td>Exp.11</td>
<td>Bayes,NaiveBayes-output-deluge-info Testmode-split=66% Attributes=All Dataset=Unbalanced</td>
<td>-</td>
<td>-</td>
<td>0.954</td>
<td>0.066</td>
<td>0.956</td>
<td>95.43%</td>
</tr>
<tr>
<td>Exp.12</td>
<td>Bayes,NaiveBayes-output-deluge-info Testmode-split=90% Attributes=All Dataset=Unbalanced</td>
<td>-</td>
<td>-</td>
<td>0.969</td>
<td>0.027</td>
<td>0.971</td>
<td>96.89%</td>
</tr>
<tr>
<td>Exp.13</td>
<td>Weka Classifiers-bayes.net Testmode-10-Fold Attributes=All Dataset=Unbalanced</td>
<td>-</td>
<td>-</td>
<td>0.942</td>
<td>0.061</td>
<td>0.946</td>
<td>94.16%</td>
</tr>
<tr>
<td>Exp.14</td>
<td>Weka Classifiers-bayes.net Testmode-split=90% Attributes=All Dataset=Unbalanced</td>
<td>-</td>
<td>-</td>
<td>0.954</td>
<td>0.048</td>
<td>0.957</td>
<td>95.44%</td>
</tr>
<tr>
<td>Exp.15</td>
<td>Weka Classifiers-bayes.net Testmode-split=90% Attributes=All Dataset=Unbalanced</td>
<td>-</td>
<td>-</td>
<td>0.969</td>
<td>0.010</td>
<td>0.972</td>
<td>96.89%</td>
</tr>
</tbody>
</table>
Figure 6: Diagrammatical representations of selected experiments

Below are some of the best rules generated:

1. If Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident = Morning/Evening Then Fatal in Accident: Yes.

2. If Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident = Night and Number of Victims > 2: Then Fatal in Accident: Yes.

3. If Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident = Afternoon and Type of Crashes = Vehicle With Pedestrian: Then Fatal in Accident: No.

4. If Severity of Accident = Slight and Edu/n Level = Primary and Settlement of Road = Upward and Type of Causality Vehicle = Motor Cycle, ISUZU, ISUZU-Autobus, Minibus Then Fatal in Accident: Yes.

5. If Severity of Accident = Slight and Edu/n Level = Primary and Settlement of Road = Upward and Type of Crashed Vehicle!= Motor Cycle Then Fatal in Accident: No.

4.5 Performance Measurement of Learning Algorithms

In the experiment evaluation part, we have identified that J48 and Rep tree are comparatively similar and better than the remaining three classifiers. So we have used selected the first and third experiments for each classifiers and measured performance of their classifiers accuracy as follows.

Figure 7: Confusion Matrix

Since the dataset we have was unbalanced, taking accuracy of the model to decide one model as best model is misleading. In such cases, it is advisable to take precision and recall for deciding whether one model is better than the other or not. In our cases, four of the experiments listed above have comparatively similar precision and recall values. But the 1st and 7th experiments were computed by 10 fold cross validation and the rest were computed by 90% split value for training and testing the model. So model with good predictive accuracy can be obtained by experiments performed with 10 fold cross validation tests according to expert judgments. Then we ignored the rest experiments with 90% split tests and accepted experiments with cross validation tests. Experiment 1st (98%) average precision and (97.75%) average recall for two class labels and 7th experiment (97.70%) average precision and (97.90%) average recall were selected to determine the best model with good predictive accuracy for fatal and non-fatal accident occurrences.

The above result shows that J48 Tree and Rep tree are significantly best by performance than all other classifiers with the given dataset. Naïve Bayes and Bayesian network classifiers are significantly good by their performance and the rest two algorithms (Random forest and Random tree) classifiers are poor by performance when compared to other classifiers with the given dataset.

5. CONCLUSION

In this study, machine Learning approaches have been applied for data analysis and prediction of car traffic accident datasets to explore important features and pattern relationships to car traffic accident occurrences. We addressed various statements of problems and objectives to determine determinant factors of car traffic accidents. We identified 7 most commonly participating vehicles, 20 areas for frequent accident occurrences, pedestrians and passengers as the most common victims and J48 and Rep tree as best algorithms by performance and model accuracy. 23 best rules were generated from the selected model for accident occurrences, results have been discussed and finally some points have been recommended for the future researchers.

Based on the outcomes of this study, the following points were recommended for the future researchers. Comparatively better results might be obtained if they try accident predictions with techniques like support vector machine, multilayer perceptron and artificial neural networks. Add some unconsidered attributes to datasets and relate cases to behavior of drivers like amount of alcohol taken and mental normality of drivers to get better results. Try with deep learning with large amount of instances to get better result and integrate it with knowledge base to know cases for accident occurrences to use is as an expert system.
REFERENCES