Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

Quantitative Study of Traffic Accident Prediction Models: A Case Study of Virginia Accidents

Tahani Almanie
Department of Computer Science, George Mason University, Fairfax, VA, USA
Email: Talmanie@gmu.edu

----------ABSTRACT----------

Traffic accidents are a serious problem that threatens people's lives, health, and properties. Thus, decreasing traffic accidents is a crucial demand for public safety. This paper proposes two data mining models to predict accident risks based on the decision tree and the naive Bayes algorithms. The purpose of the classifiers is to predict the potential severity of a traffic accident based on a set of data attributes related to the weather factors, accident timing, and properties of the road. The models are developed using data on accidents in Virginia between 2016 and 2021. Several metrics are considered to measure the performance of each model such as accuracy, precision, recall, and F1-score. Furthermore, to statistically compare the performance of the prediction models, the study employs three quantitative analysis tools, approximate visual test, paired observations, and ANOVA. The experimental results revealed that the decision tree outperforms naive Bayes in terms of prediction accuracy.

Keywords - **Traffic Accidents, Severity Prediction, Quantitative Analysis, Decision Tree and Naive Bayes Algorithms.**

Date of Submission: Jan 18, 2023 Date of Acceptance: February 2, 2023

1. INTRODUCTION

The transportation system plays a pivotal role in people's lives and is one of the most important indicators of living standards. Traffic accidents are a significant problem that threatens people's lives, health, and properties [1]. According to the National Highway Traffic Safety Administration (NHTSA), approximately 43k traffic accident fatalities happened in 2021, which represents a 10.5 percent increase from fatalities in 2020 [2]. Additionally, the number of people killed in speeding accidents has increased by 5 percent. Further, police-reported and alcohol-related deaths have increased by 5 percent. Predictive models of traffic accidents could help in understanding the causes of accidents and reducing the number of accidents. In recent years, understanding the causes of traffic accidents and predicting and analyzing accidents have attracted many researchers. With the aim of reducing traffic accidents, this paper proposes two data mining models to predict accident risks. Moreover, it presents a thorough evaluation of the models using multiple performance metrics and quantitative analysis tools. This solution could help agencies predict the risk of future accidents and increase people's awareness of traffic accidents and their relative factors.

The paper is structured as follows. Section 2 presents the related study to this work. Section 3 provides a description of the problem and the importance of the study. Section 4 explains the proposed methodology, including the dataset description, data analysis, data preprocessing, model building, and statistical and quantitative analysis involved. Section 5 presents and analyzes the experimental results. Finally, the conclusion of this research is provided in Section 6.

2. RELATED WORK

Several studies are related to the analysis and prediction of traffic accidents. Researchers have developed various algorithms to predict traffic accidents using data mining and machine learning techniques. Below I review some of the work done in this field.

Banerjee et al. [3] presented a comparative study of several machine learning models used to predict traffic accident risk. The study concluded that random forests and classification tree algorithms were effective in analyzing accident factors and their correlation with the severity of the accident. Besides, they found that the K-means clustering can help identify locations that are more prone to accidents.

Thaduri et al. [4] proposed a convolutional neural network model (CNN) for traffic accident prediction using *India Accident* (2016-2018) dataset. The prediction was based on a set of traffic accident factors such as light, weather and traffic flow. According to their experimental results, the proposed CNN prediction model was more efficient than the traditional backpropagation (BP) algorithm and achieved high prediction accuracy.

In the work presented in [5], the author predicted the severity of traffic accidents using several machine learning techniques such as decision trees, Bayesian networks, artificial neural networks, regression models, and support vector machines (SVMs). The models were developed based on data from accidents in Seoul, Korea between 2009 and 2011. A comparative analysis of these machine-learning techniques was performed. The experimental results showed that SVM is better than others in terms of prediction accuracy.

Despite all the related work presented, none of them consider the use of statistical methods to conduct quantitative comparisons and statistical analysis of the performance of their models.

3. PROBLEM DESCRIPTION

Decreasing traffic accidents is a crucial demand for public safety. To address this problem, I propose a solution for predicting traffic accidents by utilizing two data mining techniques, decision tree, and naive Bayesian classifiers. I train the models using around 46k accident records that

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

occurred in Virginia between 2016 and 2021. The objective of the classifier is to predict the severity of a traffic accident (low, medium, high) based on a set of data attributes related to the weather, timing, and traffic conditions. To evaluate the performance of each model, I utilize several metrics such as precision, accuracy, recall, and F1 score. Furthermore, to statistically compare the prediction accuracy of the models and decide which is better, I employ three quantitative analysis tools, the approximate visual test, paired observations, and *one-factor* ANOVA at a 95% confidence level.

4. METHODOLOGY USED

In this section, I start by presenting a detailed description of the dataset used in this project. Besides, to better understand the dataset and get more insight into the data, I present some statistical analysis using several graphic displays. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. After that, I explain how I prepared the dataset to be suitable for the intended study. Then, I introduce the selected data mining algorithms and describe with complete details of experimental settings how I constructed the models to achieve the study's purpose. In the following subsection, I provide the statistical and quantitative analysis tools that I have employed to compare the performance of the prediction models.

4.1 Dataset Description

The selected dataset for this project is titled *US-Accidents*, which is publicly available on the machine learning and data science community named *Kaggle* [6]. The dataset includes traffic accident data from 49 states of the US. Since February 2016, the data has been regularly gathered via a variety of data providers, including numerous APIs. It currently contains around 2.8 million records of traffic accidents that occurred from February 2016 to December 2021. The dataset includes 47 data attributes varied between accident location, timing, accident description, weather data, traffic signs, and the accident's severity. The attributes of the dataset are shown in Table 1 below. The selected sample for this project represents Virginia state with 46236 accident records.

Table 1. Attributes of US-Accidents Dataset [6]

Category	Attributes	Attribute Names
Identification	1	ID
Location	13	Start_Lat, Start_Lng, End_Lat, End_Lng, Number, Street, Side, City, County, State, Zipcode, Country, Airport_Code
Timing	8	Start_Time, End_Time, Timezone, Weather_Timestamp, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight
Description	2	Distance(mi), Description
Weather Properties	9	Temperature(F),Wind_Chill(F), Humidity(%),Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph),Precipitation(in), Weather_Condition
Traffic Sign	13	Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop
Traffic Condition	1	Severity

4.2 Dataset Analysis

In order to get a comprehensive overview of the selected data, I conducted a statistical analysis on the dataset's attribute values. Then I created multiple charts to visualize and better understand the data which could assist in performing the preprocessing of the dataset.

The first histogram, shown in Fig. 1, is derived from the *US-Accidents* dataset and displays the top 10 states with the most accidents. It is obvious that California comes in first place with a significant number of accidents, followed by Florida while Virginia comes in fifth place. In addition, Fig. 2 represents a histogram for the top 10 Virginia counties in number of accidents, with Fairfax County in first place.

Fig. 3 displays the distribution of the accidents' severity in Virginia. As shown, the majority of accidents are of low severity with 91%, whereas 7% of accidents are considered of high severity and only 2% of medium severity.

Fig. 4 demonstrates the temperature distribution for traffic accidents in Virginia. It can be observed that most of the accidents occur between 70 and 80 degrees Fahrenheit. Moreover, most of accidents happen at 40-60% humidity, followed by 90-100% humidity, as seen in Fig. 5.

Finally, Fig. 6 shows the traffic signs most associated with accident occurrence in Virginia. The presence of junctions clearly causes the most accidents, directly followed by the presence of a traffic signal.

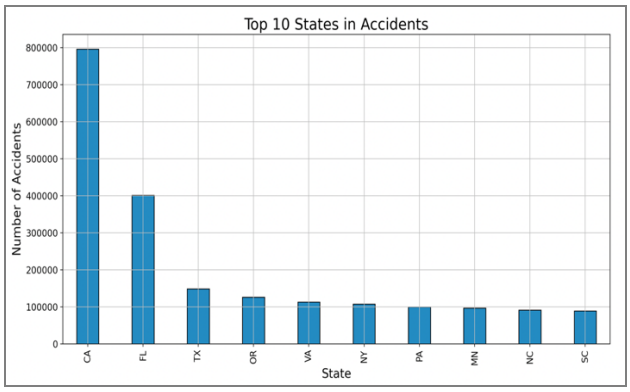

Fig. 1. Top 10 US states in the number of accidents.

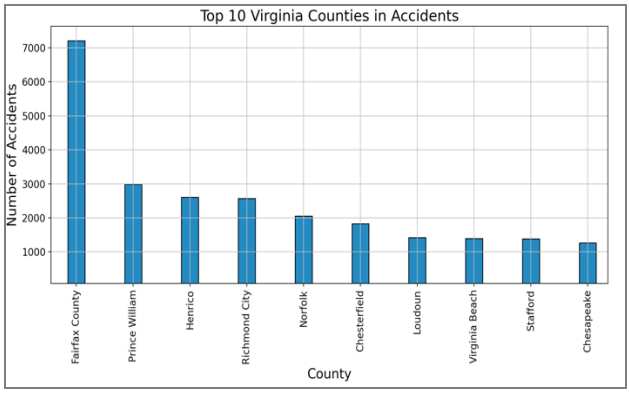

Fig. 2. Top 10 Virginia counties in the number of accidents.

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

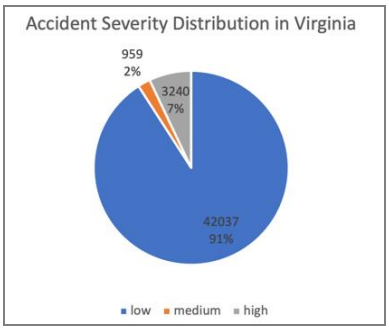

Fig.3. Distribution of accident severity in Virginia.

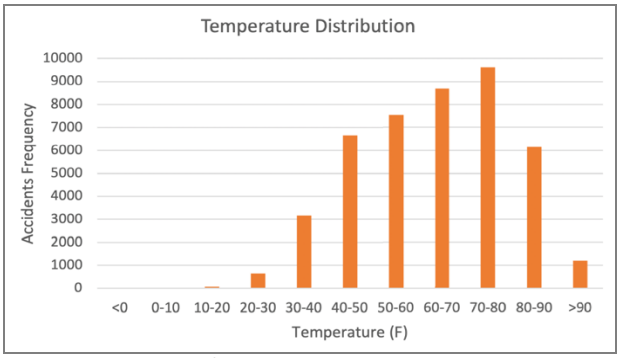

Fig.4. Distribution of the temperature (F) in Virginia accidents.

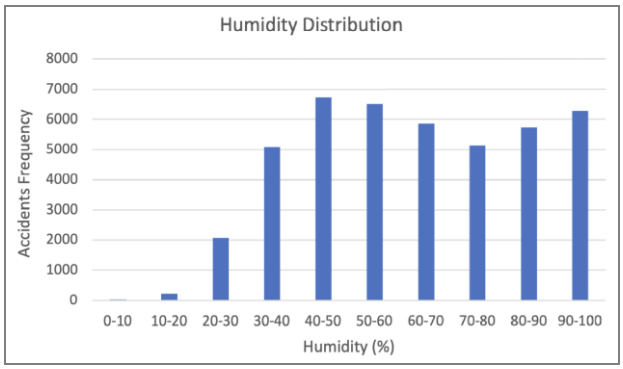

Fig.5. Distribution of the humidity percentage in Virginia accidents.

Fig.6. Distribution of the presence of traffic signs in Virginia accidents.

4.3 Data Preprocessing

Data preprocessing is a data mining task that involves preparing data and transforming it into a form suitable for mining and model building. The purpose of preprocessing is to reduce data size, handle missing data, remove outliers, and extract data features. It includes multiple techniques such as data cleaning, integration, transformation, and reduction [7].

The dataset contains multiple missing values in several attributes. I performed data cleaning through handling the missing data in a specific way. If the tuple contains three or more missing values, I remove the tuple. Otherwise, I fill the missing value with the attribute median or mode based on its data type. Accordingly, the missing values in the temperature attribute were filled with 65 degrees Fahrenheit, which is the temperature median value. 62 % is the median of the humidity attribute and was used to fill its missing values. The mode value "clear" was used to fill the missing values in the weather condition attribute. I applied the same process for the other attributes.

Next, I applied dimensionality reduction using attribute subset selection. The selected attributes are the relevant attributes for the mining purpose and all the irrelevant attributes have been removed from the dataset.

The data transformation was performed as well for two attributes of the dataset. For the *severity* attribute, I changed its distinct values from 2, 3 and 4 into three clear labels *low*, *medium* and *high*. Besides, to improve the accuracy of the model, I decided to reduce the diversity of the *weather condition* that contains 47 distinct values by mapping their values to fall within a smaller group. Thus, I minimized the *weather conditions* list by grouping them into only eight common weather conditions: *clear, cloudy, windy, fog, rain, snow and thunder.*

After the data cleaning phase, the number of instances in the dataset changed from 46236 to 44139. Additionally, after data reduction, the number of attributes changed from 47 to 11 attributes. The dataset is now ready for the mining phase. A detailed description of the dataset after the preprocessing phase is illustrated in Table 2.

4.4 Model Building

Predictive modeling is the process of using known and historical data to build, process, and validate a model that can be used to predict future results [8]. There are multiple types of predictive modeling such as classification, clustering, time series and outlier models [9]. For this study, I applied two classification models, a decision tree classifier and a naive Bayesian classifier, to the Virginia accidents dataset. The purpose of the classifiers is to predict the potential severity of a traffic accident (low, medium, high) based on a set of data attributes related to the weather factors such as temperature, humidity and visibility, accident timing, in addition to several properties of the road such as the existence of crossing or traffic signal. Therefore, the *severity* attribute is selected as the class label for both models. The other attributes are described in Table 2.

In this section, I define the two classification algorithms and then describe the implementation details of building the models to achieve the research objective.

Decision Tree Classifier: A decision tree is one of the predictive modeling techniques used in data mining and machine learning. It is a flow-chart tree structure, where each internal node denotes a test on an attribute, each branch represents a test outcome, and each leaf node holds a class label. To make a prediction, the attribute values of a new instance are tested against the decision tree using a path from the root to a leaf node that holds the class prediction for that instance [10].

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

Table 2. Dataset of Virginia Accidents after Preprocessing Phase [6]

Attribute	Description	Type	Mean	Median	Mode
Side	Indicates the relative side (right / left) of the street	Nominal	-	-	R
Temperature	Indicates the temperature (in Fahrenheit)	Numeric	63.6	65	-
Humidity	Indicates the humidity (in percentage)	Numeric	63.5	62	-
Visibility	Indicates the visibility (in miles)	Numeric	9.4	10	-
Weather Condition	Indicates the weather condition (clear, cloudy, windy, fog, rain, snow, or thunder)	Nominal	-	-	Clear
Crossing	Indicates the presence of crossing in the accident location	Boolean	-	-	False
Junction	Indicates the presence of junction in the accident location	Boolean	-	-	False
Stop	Indicates the presence of stop sign in the accident location	Boolean	-	-	False
Traffic_Signal	Indicates the presence of traffic signal in the accident location	Boolean	-	-	False
Sunrise_Sunset	Indicates the time of day (day or night) according to sunrise/sunset.	Nominal	-	-	Day
Severity	Indicates the severity of the accident as low, medium, or high	Nominal	-	-	Low

I built the decision tree model for the prepared dataset using Python programming language. I used Scikit–Learn, a software library that provides a set of data mining tools for Python including decision tree induction [11]. Since Scikit–Learn package requires attribute values to be in a numeric format, I converted the dataset nominal values into numbers [12]. For example, the *Day* and *Night* values of the *Sunrise_Sunset* attribute were converted into 1 and 0, respectively. Also, I used Pandas, a software library written in Python for data analysis and manipulation [13].

I split the dataset into an 80% training set and a 20% test set. I passed 100 as the random-state parameter, which is a random number parameter that enables you to get the same training test split every time you run the code [14]. The decision tree graph was created using the Graphviz package, an open-source graph visualization software for representing diagrams from structural data [15].

To measure the quality of the split for the decision tree induction, I applied the information gain measure using the entropy function. Besides, the maximum depth of the tree was set to be four and three for the minimum samples in each leaf. To evaluate the performance of the constructed model, I used the accuracy score of the classifier along with generating a confusion matrix and a classification report that shows the related metrics.

Naïve Bayesian Classifier: The naive Bayes algorithm is a simple probabilistic classifier that computes a set of probabilities by finding the frequency and combinations of values in the given dataset. The algorithm applies Bayes's theorem and assumes that all attributes are independent of the value of the class attribute. However, this assumption is scarcely valid in real-world applications; thus, it is named as Naïve. Despite that, the algorithm tends to be fast in a variety of classification problems [16].

This model was constructed using Scikit–Learn and Pandas software libraries as well. I chose Gaussian Naive Bayes, which applies the Gaussian Naive Bayes algorithm for classification where the probability of features is assumed to be Gaussian [17].

Similarly, I divided the dataset into an 80% training set and a 20% test set, and I set the random state parameter to 100. I also used the classifier's accuracy score, the generated confusion matrix, and the classification report to evaluate the model's performance.

4.5 Statistical and Quantitative Analysis

After building the two models, we need to compare their performance according to the obtained accuracy scores to decide which model is better for the study's objective. In this section, I clarify in detail the statistical approaches that I applied for the comparisons.

The first and simplest approach to compare two alternatives is called the "Approximate Visual Test". In this approach, we compute the confidence intervals for the two proportions or two sets of measurements and visually check the overlap [18]. Since the accuracy of the model represents the ratio of correct predictions to the total number of predictions, we can use it for comparing proportions. To apply this test, I found the accuracy performance of each model using 100 as the random state parameter. Then I computed the confidence interval for the accuracy of each model at a 95% confidence level using (1).

$$\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \quad (1)$$

Where p is the sample proportion, n is the sample size, α is the significance level, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. After that, I visually checked whether the confidence intervals for the two proportions being compared overlapped. In order to say that one model is better than the other using this test, we should not have overlapping intervals.

The second approach is "Paired Observations", also known as "Paired Samples t Test", which is a commonly used statistical tool for comparing two alternatives [19]. The objective of the test is to indicate whether there is statistical evidence that the mean difference between paired observations differs significantly from zero [18]. First, we need to get the confidence interval at an identified confidence level for the mean of the differences of the paired observations using (2).

$$\left(\bar{x} - t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}}, \bar{x} + t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}}\right) \quad (2)$$

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

Where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, α is the significance level, and $t_{[1-\alpha/2, n-1]}$ is the critical value of the t distribution with $n-1$ degrees of freedom. Then if that interval contains zero, we can say that the measured differences are not statistically significant at the specified confidence level. Otherwise, we can state whether one model is better than the other one.

To apply this test, I measured the accuracy score of each model using ten different values for the random state parameter. Then I computed the confidence interval for the mean of the differences of the paired measurements at a 95% confidence level. The selected random state values were (10, 100, 200, 330, 750, 2200, 8000, 9000, 20800, and 42000).

The third and more efficient approach is "ANOVA" which stands for analysis of variance and using ANOVA to compare different alternatives is called a *one-way classification* or a *one-factor* ANOVA [18]. It separates the total variation in a set of measurements into a variation due to the effects of alternatives and a variation due to errors or random factors within each alternative. Then the statistical *F-test*, which is based on the *F* distribution is used to test whether the two variances are significantly different at a specific confidence level. If we find a statistically significant difference between the alternatives, the method of contrasts can help indicate which alternative is better. First, we find the confidence interval for contrasts at a specified confidence level using (3).

$$c \mp t_{1-\alpha/2;k(n-1)} s_c \quad (3)$$

$$s_c = \sqrt{\frac{\sum_{j=1}^{k} (w_j^2 s_e^2)}{kn}} \quad (4)$$

Where the contrast c is $\sum_{j=1}^{k} w_j \alpha_j$ a linear combination of the alternative effects α_j, and w_j are the alternative weights which are chosen in a way that their sum equals zero, $t_{[1-\alpha/2; k(n-1)]}$ is the critical value of the t distribution with $k(n-1)$ degrees of freedom, k is the number of alternatives, n is the number of experiments, and α is the significance level. S_c is the variance of c given by (4) where S^2_e is the mean square error. After that, if the interval contains zero, then there is no statistically significant difference between the alternatives included in the contrast at the defined confidence level. Otherwise, we can say which model is better.

I applied this test to compare the two models using the same set of ten accuracy measurements that I utilized in the previous statistical test. Moreover, I applied the method of contrasts to indicate which model has better performance at a 95% confidence level.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, I discuss in detail the experimental results obtained from the constructed models. I start with an evaluation of the models' performance metrics. Then I provide the results of the conducted comparison of the models through a detailed statistical analysis.

5.1 Models Evaluation

The first result obtained from building the decision tree is the induced tree structure shown in Fig. 7. The attribute *Sunrise_Sunset* was chosen as the root of the tree, which implies that it is the best splitting attribute for the dataset.

Both models were built on a total of 35311 instances as training set and 8828 instances as a testing set. The decision model achieved an initial prediction accuracy of 90% whereas the Naive Bayes classifier reported less prediction accuracy with 82%. Tables 3 and 4 report the confusion matrix of applying each model which shows the total number of correct and misclassified instances in each class label.

Further, other metrics were used to measure the quality of predictions. Precision represents the percentage of correct positive predictions out of the total number of positive predictions [20]. Recall that provides the percentage of correct positive predictions out of the total number of actual positive predictions. Finally, the weighted harmonic mean of precision and recall is reported using the F1-score.

Overall, the decision tree model outperforms naive Bayes in terms of overall accuracy and predicting the "low" class. Yet, it performs poorly in predicting the "high" class compared to the other model. Tables 5 and 6 display the classification report for each model which summarizes the overall performance for each class based on the different metrics. Fig. 8 displays a chart for each metric to visualize the overall performance of the two models according to the other metrics.

Table 3. Decision Tree Confusion Matrix

		Predicted		
Actual	class	low	medium	high
	low	7928	3	1
	medium	197	1	0
	high	698	0	0

Table 4. Naive Bayes Confusion Matrix

		Predicted		
Actual	class	low	medium	high
	low	7208	542	182
	medium	165	32	1
	high	634	30	34

Table 5. Decision Tree Classification Report

	Precision	Recall	F1-Score	Support
low	0.90	1.00	0.95	7932
med	0.25	0.01	0.01	198
high	0.00	0.00	0.00	698
accuracy			0.90	8828
macro avg	0.38	0.33	0.32	8828
weighted avg	0.81	0.90	0.85	8828

Table 6. Naive Bayes Classification Report

	Precision	Recall	F1-Score	Support
low	0.90	0.91	0.90	7932
med	0.05	0.16	0.08	198
high	0.16	0.05	0.07	698
accuracy			0.82	8828
macro avg	0.37	0.37	0.35	8828
weighted avg	0.82	0.82	0.82	8828

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

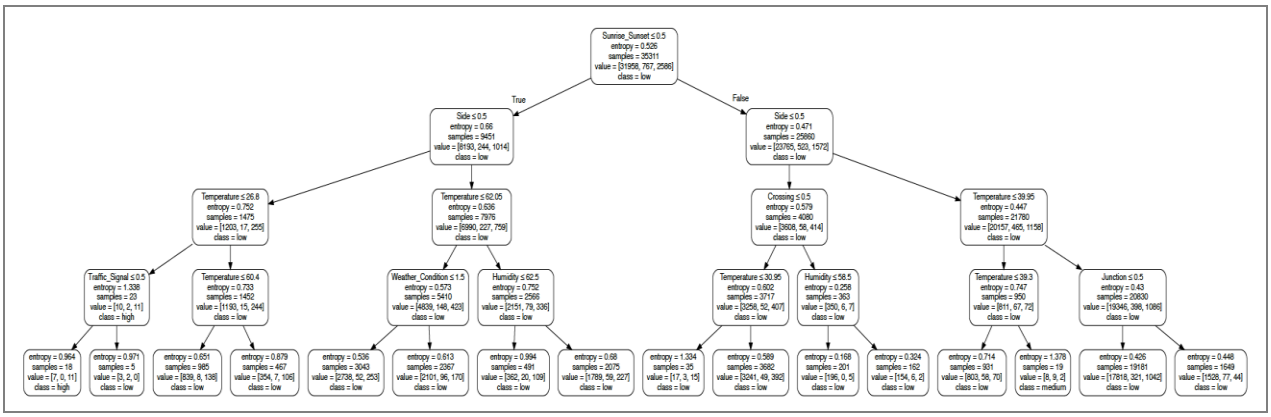

Fig. 7. The induced tree structure from building the decision tree model.

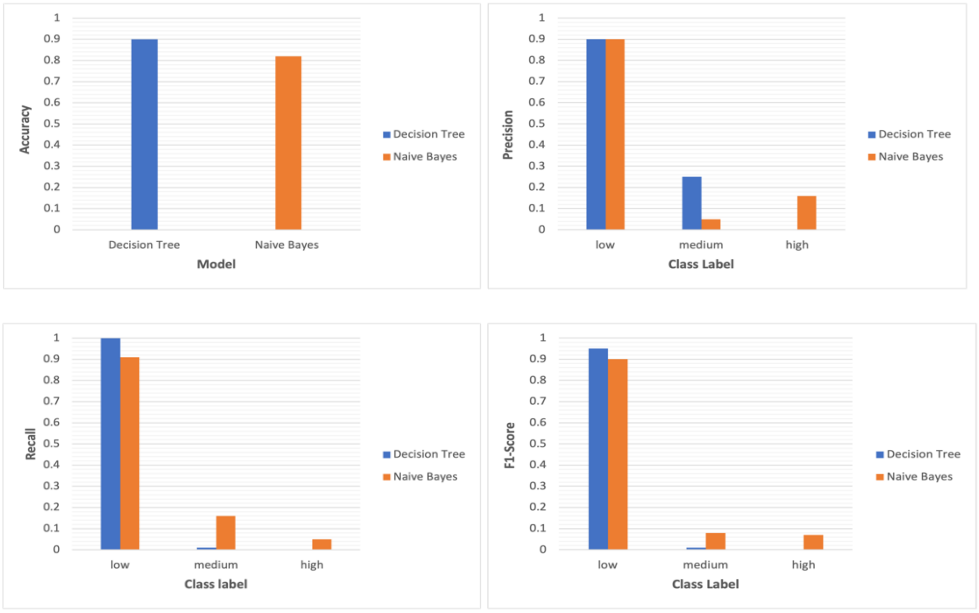

Fig. 8. The overall performance of the two models based on different metrics: accuracy, precision, recall, and F1-score.

5.2 Statistical Results of Comparing Models

This section explains statistically the results obtained from comparing the performance of the two constructed models using the three predefined statistical approaches. The goal is to explore which model is better for the study's objective. The comparison of the models is conducted based on their obtained accuracy score.

I started by applying the *approximate visual test*. Here I used the obtained accuracy score of each model when passing 100 as the random state parameter. First, I derived the confidence interval for the accuracy score of each model at a 95% confidence level using (1). In this case, p in the formula represents the accuracy ratio (0.9 for the decision tree, 0.82 for naive Bayes), n is the size of the test set (8828 total number of instances in the test set), α is the significance level (0.05), and $z_{1-\alpha/2}$ is (1.96). The confidence interval for the accuracy performance at 95% confidence level is [0.894, 0.906] for the decision tree model, and [0.812, 0.828] for the naive Bayes model.

Next, I visually checked whether the confidence intervals for the two proportions being compared overlapped. As shown in Fig. 9, we do not have overlapping intervals. Thus, we can say that using the *approximate visual test*, the decision tree model statistically performs better than the naive Bayes model with a 95% confidence level.

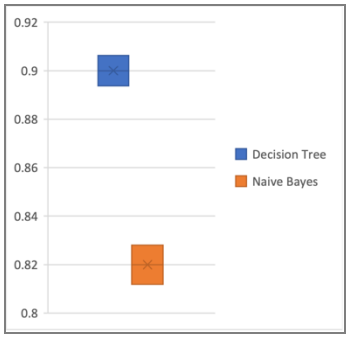

Fig. 9. Approximate Visual Test.

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

Following, I employed the *paired observations test* on a set of ten accuracy measurements generated by changing the values of the random state parameter for both models. Table 7 reports the observed accuracies for the two models. using formula (2), I found the confidence interval for the mean of the differences of the paired measurements at 95% confidence level. In this case, $\bar{x}$ is the sample mean (0.205), s is the sample standard deviation (0.261), n is the sample size (10 measurements), α is the significance level (0.05), and the value of $t_{[1-\alpha/2,\, n-1]}$ is (2.262). The resulted confidence interval for the accuracy performance at 95% confidence level is [0.018, 0.392]. Since the interval does not contain zero, we can state that using the *paired observations test*, the decision tree model statistically outperforms the naive Bayes model with 95% confidence level.

To get a more accurate comparison of the models, I employed the *one-factor* ANOVA technique using the same set of accuracy measurements in Table 7. The results of applying ANOVA analysis at the significance level of 0.05 are reported in Table 8. Each group represents a model and the summary of their measurements count, sum, average, and variance are mentioned in the first table. On the other hand, the second table displays for each source of variation (between/ within) groups, the sums of squares *SS*, the degrees of freedom *df*, the mean square values *MS*, in addition to the computed *F* statistic, and the tabulated (critical) *F* value. By applying the statistical *F-test*, we observe that the computed *F* statistic (6.242) is greater than the critical *F* value (4.414). Therefore, we can confirm that with 95% confidence, the differences among alternatives are statistically significant.

Next, by applying the method of contrasts using (3), I computed the confidence interval for contrasts at the significance level of 0.05. In this case, the computed contrast c is (0.205), the variance S_c is (0.058), and $t_{[1-\alpha/2;\, k\,(n-1)]}$ is (2.101). The resulted confidence interval for the contrast is [0.083, 0.327]. Since the interval does not contain zero, we can conclude that the decision tree model is statistically better the naive Bayes model with 95% confidence level.

Table 7. Paired Observations Test

Measurement	Model A Decision Tree	Model B Naive Bayes	A-B
1	0.91	0.83	0.08
2	0.9	0.82	0.08
3	0.91	0.22	0.69
4	0.9	0.83	0.07
5	0.91	0.82	0.09
6	0.91	0.83	0.08
7	0.91	0.81	0.1
8	0.91	0.2	0.71
9	0.91	0.83	0.08
10	0.9	0.83	0.07
Sample Mean			0.205
Sample Standard Deviation			0.261

Table 8. One-Factor ANOVA Test

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Model A	10	9.07	0.907	2.3E-05
Model B	10	7.02	0.702	6.7E-02

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*F crit*
Between Groups	0.210	1	0.210	6.242	4.414
Within Groups	0.606	18	0.034		
Total	0.816	19			

6. CONCLUSION

This paper introduced a quantitative analysis of traffic accident prediction models based on the decision tree and naive Bayes algorithms for a dataset of accidents in Virginia, US. It started with a detailed description of the selected dataset, and to get more insight into the data, it presented a set of visual statistical analyses for the data attributes. The preprocessing phase of the dataset was explained through data cleaning, reduction, and transformation. Then, it provided the implementation details of building the models. Several metrics such as precision, accuracy, recall, and F1 score were used to measure the performance of each model. In addition, the study applied three quantitative analysis tools, the approximate visual test, paired observations, and ANOVA to statistically compare the performance of the prediction models. Experimental and statistical results showed that the prediction accuracy of the decision tree outperforms naive Bayes with 95% confidence. The constructed models could be used to assist agencies in predicting the risk of future accidents and to raise people's awareness regarding traffic accidents and their related factors.

ACKNOWLEDGMENT

This work is done using the US Accidents dataset. I would like to thank Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath for making this useful dataset publicly available [6].

REFERENCES

[1] M. Gaber, A. Mohamed Wahaballa, A. Mahmoud Othman, and A. Diab, "Traffic accidents prediction model using Fuzzy Logic: Aswan desert road case study," *JES. Journal of Engineering Sciences*, vol. 45, no. 1, pp. 28–44, 2017.

[2] NHTSA, "Newly released estimates show traffic fatalities reached a 16-year high in 2021," *NHTSA*, 17-May-2022. [Online]. Available: https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities. [Accessed: 04-Dec-2022].

[3] K. Banerjee, V. Bali, A. Sharma, D. Aggarwal, A. Yadav, A. Shukla, and P. Srivastav, "Traffic accident risk prediction using machine learning," *2022 International Mobile and Embedded Technology Conference (MECON)*, 2022.

[4] A. Thaduri, V. Polepally, and S. Vodithala, "Traffic accident prediction based on CNN model," *2021 5th*

Volume: 14 Issue: 05 Pages: 5582 - 5589 (2023) ISSN: 0975-0290

International Conference on Intelligent Computing and Control Systems (ICICCS), 2021.

[5] S.-L. Lee, "Assessing the severity level of road traffic accidents based on machine learning techniques," *Advanced Science Letters*, vol. 22, no. 10, pp. 3115–3119, 2016.

[6] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A countrywide traffic accident dataset," *arXiv.org*, 12-Jun-2019. [Online]. Available: https://arxiv.org/abs/1906.05409. [Accessed: 04-Dec-2022].

[7] S. Alasadi and W. Bhaya, "Review of Data Preprocessing Techniques in Data Mining," *Journal of Engineering and Applied Sciences*, 2017.

[8] J. Frankenfield, "Reading into Predictive modeling," *Investopedia*, 21-Sep-2021. [Online]. Available: https://www.investopedia.com/terms/p/predictive-modeling.asp. [Accessed: 04-Dec-2022].

[9] I. Editorial Team, "10 predictive modeling types (with benefits and uses)," *Indeed*, 16-Nov-2021. [Online]. Available: https://www.indeed.com/career-advice/career-development/predictive-modeling-types. [Accessed: 04-Dec-2022].

[10] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in Data Mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094–2097, 2016.

[11] "1. supervised learning," *scikit*. [Online]. Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. [Accessed: 04-Dec-2022].

[12] "1.10. decision trees," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart. [Accessed: 04-Dec-2022].

[13] "Pandas," *pandas*. [Online]. Available: https://pandas.pydata.org/. [Accessed: 04-Dec-2022].

[14] M. Galarnyk, "Understanding train test split," *Built In*, 28-Jul-2022. [Online]. Available: https://builtin.com/data-science/train-test-split. [Accessed: 04-Dec-2022].

[15] *Graphviz*. [Online]. Available: https://graphviz.org/. [Accessed: 04-Dec-2022].

[16] M. M. Saritas and A. Yasar, "Performance analysis of ann and naive Bayes classification algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019.

[17] "1.9. naive Bayes," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html. [Accessed: 04-Dec-2022].

[18] D. J. Lilja, *Measuring Computer Performance: A practitioner's guide*. Cambridge, UK: Cambridge University Press, 2005.

[19] "SPSS tutorials: Paired samples T test," *LibGuides*. [Online]. Available: https://libguides.library.kent.edu/spss/pairedsamplesttest. [Accessed: 04-Dec-2022].

[20] Zach, "How to interpret the classification report in sklearn (with example)," *Statology*, 09-May-2022. [Online]. Available: https://www.statology.org/sklearn-classification-report/. [Accessed: 04-Dec-2022].

Author Biography

Tahani Almanie received the M.S. degree in computer science from University of Colorado Boulder, CO, USA, in 2016. She is currently working toward the Ph.D. degree in computer science with George Mason University, Fairfax, VA, USA. She is a Lecturer at Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her research interests include data mining, predictive modeling, recommender systems, and decision optimization.