

PAPER • OPEN ACCESS

## Traffic Accident Severity Prediction Using Naive Bayes Algorithm - A Case Study of Semarang Toll Road

To cite this article: W Budiawan *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **598** 012089

View the [article online](#) for updates and enhancements.



**IOP | ebooks™**

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the **collection** - download the first chapter of every title for free.

# Traffic Accident Severity Prediction Using Naive Bayes Algorithm- A Case Study of Semarang Toll Road

W Budiawan<sup>1,2</sup>, S Saptadi<sup>1</sup>, Sriyanto<sup>1</sup>, C Tjioe<sup>1</sup> and T Phommachak<sup>2</sup>

<sup>1</sup>Department of Industrial Engineering, Faculty of Engineering, Diponegoro University, Jl. Prof. H. Soedarto, SH. Semarang 50275, Indonesia

<sup>2</sup>Student at Department of Architectural and Civil Engineering, Toyohashi University of Technology, Japan

wiwikbudiawan@ft.undip.ac.id

**Abstract.** A traffic accident was one of the leading cause of death in Indonesia. Toll Road is one of the places where traffic accidents occur. In 2007-2017 there were 501 accidents at Semarang Toll Road. Accident in Semarang Toll Road has a variety of severity. The most severe case is death. A traffic accident can lead to death. One of the ways to decrease the number of the accident was decreased the severity of the accident. This achieved by making a prediction model. The prediction model can predict the severity of the accident based on the attribute affecting the severity of the accident. In this research, Days, Type of Road, Weather, Condition of Road, Time of the accident, Sex of Driver, and Type of Vehicle were chosen as attributes to make prediction model of accident severity. Naive Bayes algorithm was used to make the model which can predict accident severity. The result was an accident prediction model with an accuracy of 39.49% to predict accident severity and the probability of an accident.

## 1. Introduction

Accidents are defined as an unplanned and controlled event that can be caused by humans, situations, environmental factors, or combinations of these things [1]. The causes of traffic accidents were grouped into four elements, namely human, vehicle, road, and environment [2]. Environmental factors were weather conditions (foggy and rainy). Weather conditions had a significant impact on vehicle performance, driver's visibility, driver behavior, travel demand, traffic flow characteristics, and traffic safety [3]. A traffic accident is one of the leading cause of death in Indonesia. The amount of traffic accident in Semarang Toll Road were 501 from 2007 to 2017. A traffic accident has a level of severity in which the highest was death. Based on this, any attempt to increase safety in Toll Road need to be done, and one of the ways to do this is by decreasing the level of severity in an accident.

In their research, [4] said that many factors are leading to how an accident happens namely environmental factor like weather condition, type of vehicle, driver behavior and characteristic factor like an age of driver and sex type of driver. These factors have a role in determining the severity level of the accident. In other research, [5] said that by applying the data mining technique to make a prediction model in these traffic accident data, it could help decision maker to make a decision related to the safety of a driver. So in accordance with this, we can agree that safety in Toll Road can be increased by making a prediction model of accident severity.

According to [6], *data mining* was a process to gain pieces of information from a group of data which help in making a decision. Data mining consists of Classification, Clustering, Estimation, and



Association. In their research on making a prediction model of accident severity, [7] applied classification technique of data mining all of which are Naive Bayes, ID3, and Random Tree. Another research in the making of a prediction model of accident severity is seen in the research of [8] in which they use classification technique like Naive Bayes, Decision Tree, and Random Tree. So based on this, a prediction model is made by classification technique, specifically using the Naive Bayes algorithm technique.

## 2. Methods

This research done on Semarang Toll Road. The data taken from PT Jasa Marga Semarang (Indonesia Highway Corporation) which consists of traffic accident data from 2007 to 2017 and traffic data from 2007 to 2017. Research has a quantitative approach because research has the aim to identify dangerous area from the probability of an accident happening for each area in Semarang Toll Road. Then to identify the accident severity as a result of traffic accident attributes, classification is made from which the result can be used to predict accident severity of traffic accident. Lastly, the classification model's accuracy will be measured.

For the attributes of traffic accident taken from [7] research, it is then adjusted to the real data from PT Jasa Marga the result can be seen in Table 1.

**Table 1.** The attribute of a traffic accident

Attributes	Data Type	Description
<b>Day</b>	Text	Monday/ Tuesday/ Wednesday/ Thursday/ Friday/ Saturday/ Sunday
<b>Type of Road</b>	Text	Straight flat/ Straight descend/ Straight ascend/ Curve flat/ Curve descend/ Curve ascend
<b>Weather</b>	Text	Sunny/ Cloudy/ Foggy/ Dusty/ Smoky/ Drizzly/ Rainny
<b>Road surface</b>	Text	Dry/ Wet/ Sandy
<b>Time of accident</b>	Text	Morning/ Noon/ Afternoon/ Night
<b>Sex type of driver</b>	Text	Man/ Woman
<b>Type of Vehicle</b>	Text	Sedan/ Jeep/ Pick Up/ Minibus/ Bus/ Truck
<b>Accident Severity</b>	Text	Material damage/ Minor injuries/ Major injuries/ Fatal

From the assembled data, a process consists of Data Selection, Data Cleaning and Data Transformation will be done. Data Selection has an aim to select attributes suites for research. Data Cleaning has an aim to delete incomplete data. Lastly, Data Transformation has an aim to make data easier to be made to become a prediction model.

Division of data will be applied after data has gone through the above processes. Data will be divided by 60% and 40% ratio. 60% of the data will become Data Training to make prediction model and 40% of the data will become Data Testing to measure the model accuracy. After that prediction model will be made by calculating traffic accident probability based on the number of a traffic accident in Semarang Toll Road for a period of 20017 to 2017 for each 1km. Section A and B from Semarang Toll Road will be merged into one Area named Area 1 and Section C will be renamed Area 2.

Prediction model will predict accident severity by relating traffic accident attributes and accident severity by using Naive Bates classifier. The result at the very end will be a prediction model which predict the probability of traffic accident in a given area and accident severity is given the traffic accident attributes.

Based on 6 Naive Bayes classifier is a classification method using probability. Naive Bayes algorithm predicts probability in the future based on past data. Probability is stated as follows:

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \quad (1)$$

*In which:*

$P(H|X)$  = Probability of  $Y$  (outcome) if  $X$  is known

$P(H)$  = Probability of outcome known from historical data

$P(X)$  = Evidence Probability

### 3. Result and Discussion

Calculation of probability of traffic accident is done for each 1km area in Semarang Toll Road. The probability of traffic accident is a result of some a traffic accident for the period of 2007 to 2017 divided by the number of traffic from 2007 to 2017. The result can be seen in Table 2.

**Table 2.** Traffic accident probability

Area	Kilometre	Probability	Percent	Area	Kilometre	Probability	Percent
<b>1</b>	0km-1km	4.00E-08	3.593	<b>2</b>	0km-1km	4.44E-08	3.992
	1km-2km	2.22E-08	1.996		1km-2km	7.77E-08	6.986
	2km-3km	2.66E-08	2.395		2km-3km	6.00E-08	5.389
	3km-4km	4.00E-08	3.593		3km-4km	3.33E-08	2.994
	4km-5km	4.89E-08	4.391		4km-5km	5.11E-08	4.591
	5km-6km	6.00E-08	5.389		5km-6km	2.66E-08	2.395
	6km-7km	2.22E-08	1.996		6km-7km	1.78E-08	1.597
	7km-8km	1.78E-08	1.597		7km-8km	1.78E-08	1.597
	8km-9km	2.00E-08	1.796		8km-9km	2.22E-09	0.2
	9km-10km	1.47E-07	13.174		9km-10km	2.00E-08	1.796
	10km-11km	1.07E-07	9.581		10km-11km	7.99E-08	7.186
	11km-12km	6.22E-08	5.589				
	12km-13km	4.66E-08	4.192				
	13km-14km	2.22E-08	1.996				

Prediction can be made by multiplying each probability of each attribute of the traffic accident, so based on this probability of each attribute need to be calculated. This can be done by calculating the frequency of each attribute. Table 3 is the frequency of each type of accident severity. Table 4 is frequency of each type of Time of accident attributes. Table 5 is the frequency of each type of Day of the accident attributes. Table 6 is the frequency of each type of vehicle attributes. Table 7 is the frequency of each type of gender of the driver attributes. Table 8 is the frequency of each type of weather attributes. Table 9 is the frequency of each type of Road attributes. Table 10 is the frequency of each type of Road Surface attributes

**Table 3.** The frequency of accident severity

Accident Severity	Amount
<b>Material Damage</b>	78
<b>Minor injuries</b>	125
<b>Major injuries</b>	79
<b>Fatal</b>	11
<b>Total</b>	293

As we can see from Table 3 above, the number of frequency of the accident severity will become the denominator for calculating the probability of each attribute. The following table of Table 3 until table

10 we can see that the total number of each attribute is the same number of the frequency of the accident severity in Table 3.

**Table 4.** Frequency of Weather

Time	Material damage	Minor injuries	Major injuries	Fatal
<b>Day</b>	17	20	13	3
<b>Noon</b>	27	35	13	1
<b>Afternoon</b>	10	14	16	1
<b>Night</b>	24	56	37	6
Total	78	125	79	11

**Table 5.** Frequency of Day

Day	Material damage	Minor injuries	Major injuries	Fatal
<b>Monday</b>	11	15	10	1
<b>Tuesday</b>	10	22	11	2
<b>Wednesday</b>	13	9	11	3
<b>Thursday</b>	18	20	7	2
<b>Friday</b>	12	19	12	1
<b>Saturday</b>	8	24	16	1
<b>Sunday</b>	6	16	12	1
Total	78	125	79	11

**Table 6.** The frequency of Type of Vehicle

Vehicle	Material damage	Minor injuries	Major injuries	Fatal
<b>Sedan</b>	18	16	4	0
<b>Jeep</b>	3	2	1	0
<b>Pick Up</b>	4	14	6	0
<b>Minibus</b>	18	34	13	2
<b>Bus</b>	9	4	11	1
<b>Truck</b>	26	55	44	8
Total	78	125	79	11

**Table 7.** The frequency of Type of Sex

Gender	Material damage	Minor injuries	Major injuries	Fatal
<b>Man</b>	73	121	78	11
<b>Woman</b>	5	4	1	0
Total	78	125	79	11

**Table 8.** The Frequency of Weather

Weather	Material damage	Minor injuries	Major injuries	Fatal
<b>Sunny</b>	57	87	63	9
<b>Cloudy</b>	5	15	3	1
<b>Foggy</b>	0	0	0	0
<b>Dusty</b>	0	0	0	0
<b>Smoky</b>	0	0	0	0
<b>Drizzly</b>	7	5	10	1
<b>Rainy</b>	9	18	3	0
Total	78	125	79	11

**Table 9.** The frequency of Type of Road

Type of Road	Material damage	Minor injuries	Major injuries	Fatal
<b>Straight flat</b>	30	54	32	5
<b>Straight descend</b>	17	40	24	3
<b>Straight ascend</b>	9	10	9	1
<b>Curve flat</b>	6	10	2	0
<b>Curve descend</b>	11	7	9	2
<b>Curve ascend</b>	5	4	3	0
Total	78	125	79	11

**Table 10.** The frequency of Road Surface

Road Surface	Material damage	Minor injuries	Major injuries	Fatal
<b>Dry</b>	61	99	66	10
<b>Wet</b>	17	25	13	1
<b>Sandy</b>	0	1	0	0
Total	78	125	79	11

Naive Bayes probability is calculated by Dividing the number of frequency of attribute and total frequency. As an example based on Table 4, the probability is calculated by dividing the number of case Time=Pagi and Severity= Material damage divided by the number of case of Severity= Material damage which can be written as follows

$$P(\text{Time} = \text{Day} | \text{Material damage}) = \frac{17}{78} = 0.217949$$

So, with the same calculation as above for the rest of the attributes, the result can be seen in Table 11.

#### 4. Discussion

A measure of the Naive Bayes model can be calculated by predicting data test and then comparing the results of prediction and real data. For example one of the data from Data Testing can be seen in Table 12

**Table 11.** Naive Bayes probability

Attributes	Parameters	Material damage	Minor injuries	Major injuries	Fatal
<b>Time</b>	Day	0.217949	0.16	0.164557	0.272727
<b>Time</b>	Noon	0.346154	0.28	0.164557	0.090909
<b>Time</b>	Afternoon	0.128205	0.112	0.202532	0.090909
<b>Time</b>	Night	0.307692	0.448	0.468354	0.545455
<b>Day</b>	Monday	0.141026	0.12	0.126582	0.090909
<b>Day</b>	Tuesday	0.128205	0.176	0.139241	0.181818
<b>Day</b>	Wednesday	0.166667	0.072	0.139241	0.272727
<b>Day</b>	Thursday	0.230769	0.16	0.088608	0.181818
<b>Day</b>	Friday	0.153846	0.152	0.151899	0.090909
<b>Day</b>	Saturday	0.102564	0.192	0.202532	0.090909
<b>Day</b>	Sunday	0.076923	0.128	0.151899	0.090909
<b>Vehicle</b>	Sedan	0.230769	0.128	0.050633	0
<b>Vehicle</b>	Jeep	0.038462	0.016	0.012658	0
<b>Vehicle</b>	Pick Up	0.051282	0.112	0.075949	0
<b>Vehicle</b>	Minibus	0.230769	0.272	0.164557	0.181818
<b>Vehicle</b>	Bus	0.115385	0.032	0.139241	0.090909
<b>Vehicle</b>	Truck	0.333333	0.44	0.556962	0.727273
<b>Sex Type</b>	Man	0.935897	0.968	0.987342	1
<b>Sex Type</b>	Woman	0.064103	0.032	0.012658	0
<b>Weather</b>	Sunny	0.730769	0.696	0.797468	0.818182
<b>Weather</b>	Cloudy	0.064103	0.12	0.037975	0.090909
<b>Weather</b>	Drizzly	0.089744	0.04	0.126582	0.090909
<b>Weather</b>	Rainny	0.115385	0.144	0.037975	0
<b>Type of Road</b>	Straight flat	0.384615	0.432	0.405063	0.454545
<b>Type of Road</b>	Straight descend	0.217949	0.32	0.303797	0.272727
<b>Type of Road</b>	Straight ascend	0.115385	0.08	0.113924	0.090909
<b>Type of Road</b>	Curve flat	0.076923	0.08	0.025316	0
<b>Type of Road</b>	Curve descend	0.141026	0.056	0.113924	0.181818
<b>Type of Road</b>	Curve ascend	0.064103	0.032	0.037975	0
<b>Road Surface</b>	Dry	0.782051	0.792	0.835443	0.909091
<b>Road Surface</b>	Wet	0.217949	0.2	0.164557	0.090909
<b>Road Surface</b>	Sandy	0	0.008	0	0

**Table 12.** Example of Data Testing

Time	Day	Vehicle	Gender	Weather	Road Type	Road surface	Severity
Noon	Thursday	Truck	Man	Sunny	Straight flat	Dry	Minor injury

Based on Table 11 calculation of the probability of Material damage, Minor injury, Major injury, and Fatal is as follows:

a) Time

Can be seen that if Time= Noon probability of Material Damage is 0.307692, Minor injury is 0.448, Major injury is 0.468354, and Fatal is 0.545455

b) Day

Can be seen that if Day=Thursday probability of Material Damage is 0.230769, Minor injury is 0.16 Major injury is 0.088608, and Fatal is 0.181818

c) Vehicle

Can be seen that if Vehicle=Truck probability of Material Damage is 0.333333, Minor injury is 0.44, Major injury is 0.556962, and Fatal is 0.727273

d) Sex Type

Can be seen that if Sex=Man probability of Material Damage is 0.935897, Minor injury is 0.968, Major injury is 0.987342, and Fatal is 1

e) Weather

Can be seen that if Weather=Sunny probability of Material Damage is 0.730769, Minor injury is 0.696, Major injury is 0.797468, and Fatal is 0.818182

f) Type of Road

Can be seen that if Type of Road=Straight flat probability of Material Damage is 0.384615, Minor injury is 0.432, Major injury is 0.405063, and Fatal is 0.454545

g) Road Surface

Can be seen that if Type of Road Surface=Dry probability of Material Damage is 0.782051, Minor injury is 0.792, Major injury is 0.835443, and Fatal is 0.909091

Based on the above numbers of probability, results of prediction can be found by multiplying each attribute for each severity. The example is as follows:

$$P(\text{Severity} = \text{Material damage} | X) \\ = 0.262 \times 0.307692 \times 0.230769 \times 0.333333 \times 0.935897 \times 0.730769 \times 0.384615 \times 0.782051 \\ P(\text{Severity} = \text{Material damage} | X) = 0.001826$$

The calculation is also used to calculate the probability severity of Minor injury, Major injury and Fatal. The result can be seen in Table 13

**Table 13.** Naive Bayes probability of Severity

Material damage	Minor injury	Major injury	Fatal
0.001826	0.00909	0.000356	0

Based on the above number confidence of each severity can be calculated as follows:



$$\text{Confidence (Severity=material damage)} = \frac{0.001826}{0.001826+0.00909+0.000356+0} = 0.501$$

The confidence of all Severity can be calculated using the same formula. The highest confidence is used as a prediction. The result can be seen in Table 14

**Table 14.** Example result of prediction

Confidence(1)	Confidence(2)	Confidence(3)	Confidence(4)	Prediction
0.501219	0.39978	0.09899	0	Minor injury

Data in Table 14 shows that the real data is the same as prediction. This is done to the other Data Testing. The number of right prediction result is used as model accuracy. The result of that can be seen in Table 15

**Table 15.** Naive Bayes accuracy

	true 1.0	true 2.0	true 3.0	true 4.0	class precision
<b>pred. 1.0</b>	12	24	4	0	30.00%
<b>pred. 2.0</b>	26	43	26	6	42.57%
<b>pred. 3.0</b>	14	16	22	2	40.74%
<b>pred. 4.0</b>	0	0	0	0	0.00%
class recall	23.08%	51.81%	42.31%	0.00%	39.49%

Based on Table 15 the accuracy of Naive Bayes classifier to predict accident severity is 39.49%.

## 5. Conclusion

Dangerous area is identified by calculating the probability of traffic accident in which the most dangerous area is at Section B on 9km to 10km area with a ratio of an accident happening at 13.17365269% from 2007 to 2017 compared to other areas. Attributes that affect traffic accident is Time, Day, Vehicle type, Sex Type, Weather, Type of Road, and Road Surface. Prediction Model can predict accident severity based on traffic attributes which are Time, Day, Vehicle type, Sex Type, Weather, Type of Road, and Road Surface although with an accuracy of 39.49%. At the same time model can predict the probability of traffic accident in each 1km of the area in Semarang Toll Road.

## References

- [1] Colling D A 1990 *Industrial safety: management and technology* (Prentice Hall).
- [2] Warpani S P 2002 *Traffic and Mass Transportation Management (in Bahasa)* (Penerbit ITB).
- [3] Mohamed S A, Mohamed K and Al-Harhi H A 2017 *Transp. Res. Proc.* **25** 2098–107.
- [4] Al-Radaideh Q A and Daoud E J 2018 *Int. Journal of Neural Networks and Advanced Applications.* **5** 1–12.
- [5] Tesema T B, Abraham A and Grosan C 2005 *Int. Journal of Simulation: Systems, Science & Technology.* **6** 80–94.
- [6] Sinwar D and Kaushik R 2014 *Www.Ijrasnet.Com.* **2** 270–4.
- [7] Khera D and Singh W 2015 *Int. Journal of Computer Applications; National Conf. on Advances in Computing Communication and Application.* **ACCA 2015** 1–7
- [8] Shanthi R S and Geetha Ramani 2012 *Lect. Notes. Eng. Comp.* **1**