ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

A Classification Model for Predicting the Injured Body Part in Construction Accidents in Korea

Jiseon Lim[1]*, Sungjin Choi[2], Sanghyeok Kang[3]

[1] *Department of Civil and Environmental Engineering, Incheon National University, Incheon, 22012, Korea,* E-mail address: tjs0531@naver.com
[2] *Domestic Biz. Team, DL E&C, Seoul, 03181, Korea,* E-mail address: kitt7406@gmail.com
[3] *Department of Civil and Environmental Engineering, Incheon National University, Incheon, 22012, Korea,* E-mail address: lifesine@inu.ac.kr

Abstract: It is difficult to predict industrial accidents in the construction industry because many accident factors, such as human-related factors and environment-related factors, affect the accidents. Many studies have analyzed the severity of injuries and types of accidents; however, there were few studies on the prediction of injured body parts. This study aims to develop a classification model to predict the part of the injured body based on accident-related factors. Construction accident cases from June 2018 to July 2021 provided by the Korea Construction Safety Management Integrated Information were collected through web crawling and then preprocessed. A naïve Bayes classifier, one of the supervised learning algorithms, was employed to construct a classification model of the injured body part, which has four categories: 1) torso, 2) upper extremity, 3) head, and 4) lower extremity. The predictor variables are accident type, type of work, facility type, injury source, and activity type. As a result, the average accuracy for each injured body part was 50.4%. The accuracy of the upper extremity and lower extremity was relatively higher than the cases of the torso and head. Unlike the other classifications, such as spam mail filtering, a naïve Bayes classifier does not provide a good classification performance in construction accidents. The reasons are discussed in the study. Based on the results of this study, more detailed guidelines for construction safety management can be provided, which help establish safety measures at the construction site.

Key words: Safety Management, Construction Occupational Injury, Injured Part of Body, Naïve Bayes Classification, Classification Model

1. INTRODUCTION

The construction industry is one of the most dangerous industrial sectors, and the frequency and intensity of accidents are higher than other industries. In 2015, the number of deaths in the construction industry accounted for 20% of the total deaths in the United States, which is more than any other industry. In addition, the fatality rate in the construction industry increased from 9.0 per 100,000 in 2011 to 9.9 in 2015 [1]. According to the 2019 Industrial Accident Report released by the Korea Occupational Safety and Health Agency, construction-related accidents were 27,024 (24.9%) out of 108,434 industrial accidents, the second-highest after manufacturing. The number of deaths in the construction industry was 517 (25.59%) out of 2,020, the highest among industries [2].

Various attempts have been made to reduce accidents in the construction industry; however, the number of accidents is not greatly reduced. This is because the construction industry has many accident factors, such as human-related factors and environment-related factors, making it difficult to predict industrial accidents [1]. Recently, many studies have been conducted to identify accident causes or to establish a predictive model for accidents [3,4,5,6,7]. Most of these studies focused on predicting death, injury, or accident occurrence, and few studies have been conducted on the injured part of the body.

This study aims to develop a classification model to predict the part of the body of the injured along with the construction site's accident-related factors. The naïve Bayes Classifier, one of the machine learning techniques, was employed to build a model. The accident factors covered in this paper are accident type, type of work, injury source, facility type, and activity type. This study used construction accident cases provided by the Construction Safety Management Integrated Information (CSI) operated by the Korea Authority of Land & Infrastructure Safety. We collected raw data on a number of web pages through the web crawling technique. The collected data were 5,717 construction accident cases from June 2018 to July 2021. After removing missing records and outliers, 5,116 accident data were finally analyzed.

2. LITERATURE REVIEW

Several studies confirmed that injured body part is affected by many accident factors. Kang et al. (2021) confirmed that the injured body part is one of the main variables determining the number of workdays lost in Korea [8]. Choi (2015) analyzed 143 occupational injury reports in construction project on a highway in the Midwestern United States and found that the injured body parts vary by age and injury type [9]. Sugama and Ohnishi (2015) investigated the occupational accidents while using stepladders in Japan and revealed that the most commonly injured body parts were the lower limbs (34.7%) and upper limbs (21.4%) [10]. Chi and Han (2013) empirically and statistically analyzed 9,358 accidents in the U.S. construction industry between 2002 and 2011. The study investigated relationships between accidents and the injured body part, one of the injury elements by accident type [11]. Amiri et al. (2016) analyzed construction accident data from 2007 to 2011 in Iran and found that falls and falling objects-related accidents are associated with the time of the accident, place of accident, body part affected, and lost workdays. The study showed that the frequency of injury to the head, back, spine, and limbs was more [12]. Halvani et al. (2012) showed the relationships between the type of occupation and the injured part of the body were statistically significant in the construction industry in Iran [13]. Most studies tried to understand the relationship between the accident factor and the injured body part. Predicting the injured body part of a victim considering multiple accident factors would greatly help safety management. This study establishes a predictive model that classifies the injured body part with several accident factors as predictor variables.

3. METHOD

This study used construction accident cases provided by the Construction Safety Management Integrated Information (CSI) operated by the Korea Authority of Land & Infrastructure Safety. We used a 'Python program' to crawl the website for raw data collection. The collected data is converted into a database in a standardized form to be utilized for data analysis. Correlation analysis between the injured body part and accident-related factors and the chi-square test is performed to determine the associations between variables, and a classification model is developed using a naive Bayes classifier.

3.1. chi-square test

The chi-square test is suitable for analyzing categorical data, and it verifies the significance of observed and expected frequencies. The chi-square test is used to compare the distributions of individual groups. The independence test determines whether there is a dependency between two characteristics of the data. [4] In this study, workers' injured body parts were used as outcome variables, and accident type, type of work, injury source, facility type, and activity type were used as predictive variables. The chi-square test results between variables were compared, and the correlation between variables was determined based on the significance level of 0.05.

3.2. Naïve Bayes classification

A naïve Bayes classifier is one of the supervised learning algorithms traditionally used to classify spam emails, and it is relatively simple and shows good classification performance. Basically, this classifier uses Bayes' rule in Equation (1) to find the posterior probability.

$$P(C_i|x_1,\cdots\cdots,x_P) = \frac{P(x_1,\cdots\cdots,x_P|C_i)P(C_i)}{P(x_1,\cdots\cdots,x_P)} \tag{1}$$

where $P(C_i|x_1,\cdots\cdots,x_P)$ is the posterior probability of class C_i given predictors of x1, x2, …, xp, $P(C_i)$ is the prior probability of class, $P(x_1,\cdots\cdots,x_P|C_i)$ is the likelihood, which is the probability of predictor given class, and $P(x_1,\cdots\cdots,x_P)$ is the prior probability of the predictors [14].

The exact Bayesian classification is technically impractical since we have many evidence variables (predictors) in our dataset. When the number of predictors is too many, the records that we want to classify would not have an exact match. The naïve assumption introduces that the variables are independent given the class. So we can calculate the naïve Bayes probability of class 1 given attributes of x1, x2, …, xp as Equation 2.

$$P_{nb}(C_1|x_1,\cdots\cdots,x_P) = \frac{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots\cdots P(x_P|C_1)]}{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots\cdots P(x_P|C_1)]+\cdots\cdots+P(C_m)[P(x_1|C_m)P(x_2|C_m)\cdots\cdots P(x_P|C_m)]} \tag{2}$$

We can convert the Bayes equation into a simpler and naïve one by assuming the conditional independence between variables. Even though assuming independence between variables sounds ‘naïve,’ the naïve Bayes algorithm performs well in many classification tasks [15].

A naïve Bayes classifier contributes little to reveal the relationship or importance between factors. However, the classifier is superior to existing statistical models in improving the prediction accuracy and is used in many fields such as traffic, medicine, and character classification [14,16,17,18,19,20]. A naïve Bayes classifier is a classification or prediction method suited for categorical predictor variables. Therefore, a naïve Bayes classifier was used to establish a classification model in this study.

4. DATA

This study used the construction accident cases reported to CSI operated by the Korea Authority of Land & infrastructure Safety. We web-crawled the accident cases webpages for data collection; 5717 cases from June 2018 to July 2021 were collected. The cases contain accident details, such as the date and time of the accident, weather condition, facility type, accident type, availability of protective (protection) measures, accident location, number of fatalities, number of injured. In addition, the information about the injured body part was extracted from the accident narratives, accident cause, and damage details to make a new variable. We have created a dataset of 5,116 construction accident cases through data preprocessing. The injured body part, the outcome variable, has four categories: 1) torso, 2) upper extremity, 3) head and 4) lower extremity. The predictor variables are accident type, type of work, injury source, facility type, and activity type. Table 1 decribes the variables used in this study with their major categories.

Table 8. Variable Description

Variable		Category
Outcome variable	Injured Body Part	Upper extremity (1,973), Lower extremity (1,625), Torso (915), Head (603)
Predictor variable	Accident Type (10)	Slip down (1,226), Fall (988), Hit (976), Caught in between (577), Amputation/cutting (422), Struck by (293), Pressed/overturned (73), Stabbing (65), Unclassified (32), Others (464)
	Type of work (29)	Reinforced concrete building (1,454), Temporary construction (774), Dismantling and demolition (356), Reinforced concrete civil (242), Plumbing (237), Earthwork (213), Steel frame work (182), Building earthwork (181), Water storage construction (173), Bridge construction (131), Plastering work (131), Tile and masonry (125), Building ancillary construction (116), Tunnel construction (105), Road and pavement (97)
	Injury source (95)	Formwrok (622), Materials (575), Tools (439), Scaffolding (312), Steel bar (221), Work platform (174), Earth Retaining Wall (152), Steel frame (111), System shore (106), Excavator (105), Steel pipe shore (100), Ladder (76), Slab (57), Driving pile and Extract pile machine (47), Cranes (42)
	Activity Type (36)	Installation (950), Dismatling and Demolition (705), Moving (443), Transport (422), Assembling (352), Rearrangement (313), Formwork and Carpentry (211), Cutting (174), Concrete pouring (174), Finishing (163), Preparation (125), Lifting (91), Painting (73), Connecting(65), Pipe laying (57)
	Facility Type (62)	Apartment (1,458), Education and Research facility (366), Business facility (354), Roadway (318), Neighborhood infrastructure (277), Factory (268), Sewerage (181), Cultural facility (131), Waterworks (127), Site development (116), Road Bridge (113), Accommodation (76), Plant (73), Warehouse (70), House (70)

5. RESULTS

5.1. Correlation between predictor variables and outcome variable

This study first conducted the correlation analysis between the workers' injured body parts and accident factors and tested it using the chi-square test to determine whether the correlation analysis is statistically significant. The research hypothesis is that 'Injured body parts are independent of accident factors such as accident type, type of work, injury source, facility type, and activity type.' In this study, the chi-square test was conducted based on the significance level α=0.05, meaning a correlation between the injured part and the accident factors. As shown in Table 2, when the significance level α=0.05 for the accident type, work type, cause, and activity type, the p-value is smaller than the significance level, so the null hypothesis is rejected, and the alternative hypothesis is adopted. In other words, it can be seen that the injured body part is correlated with the accident type, work type, cause of death, and activity type. In the case of facility type, the value of the p-value was higher than the significance level, so there was no significant correlation with the injured body part. However, since this study aims to increase the prediction performance of classification, and it was determined that the facility also contributes to improving the prediction performance, the facility type was also included in the predictor variables.

Table 9. Injured body part by predictor variables

	Torso	Upper extremity	Head	Lower extremity	Total
Total	915(17.9%)	1973(38.6%)	603(11.8%)	1625(31.8%)	5116(100%)
Accident type (x^2=1502.433, P-value < 0.001)					
Slip down	300(24.5%)	355(29.0%)	66(5.4%)	505(41.2%)	1226
Falling	312(31.6%)	194(19.6%)	141(14.3%)	341(34.5%)	988
Hit	104(10.7%)	302(30.9%)	267(27.4%)	303(31.0%)	976
Caught in between	13(2.3%)	469(81.3%)	6(1.0%)	89(15.4%)	577
Amputation/cutting	2(0.5%)	354(83.9%)	9(2.1%)	57(13.5%)	422
Type of work (x^2=130.453, P-value = 0.001)					
Reinforced concrete_building	278(19.1%)	572(39.3%)	165(11.3%)	439(30.2%)	1454
Temporary construction	158(20.4%)	275(35.5%)	89(11.5%)	252(32.6%)	774
Dismantling and demolition construction	60(16.9%)	137(38.5%)	52(14.6%)	107(30.1%)	356
Reinforced concrete civil	43(17.8%)	100(41.3%)	30(12.4%)	69(28.5%)	242
Plumbing	37(15.6%)	86(36.3%)	28(11.8%)	86(36.3%)	237
Injury source (x^2=806.327, P-value < 0.001)					
Formwrok	130(20.9%)	235(37.8%)	63(10.1%)	194(31.2%)	622
Materials	76(13.2%)	227(39.5%)	80(13.9%)	192(33.4%)	575
Tools	15(3.4%)	332(75.6%)	40(9.1%)	52(11.8%)	439
Scaffolding	80(25.6%)	90(28.8%)	50(16.0%)	92(29.5%)	312
Steel bar	34(15.4%)	94(42.5%)	25(11.3%)	68(30.8%)	221
Activity type (x^2=312.739, P-value < 0.001)					
Installation	200(21.1%)	378(39.8%)	108(11.4%)	264(27.8%)	950
Dismatling and demolition work	118(16.7%)	264(37.4%)	115(16.3%)	208(29.5%)	705
Moving	98(22.1%)	119(26.9%)	31(7.0%)	195(44.0%)	443
Transport	85(20.1%)	146(34.6%)	37(8.8%)	154(36.5%)	422
Assembling	53(15.1%)	162(46.0%)	48(13.6%)	89(25.3%)	352
Facility Type (x^2=191.589, P-value = 0.317)					
Apartment	246(16.9%)	567(38.9%)	159(10.9%)	486(33.3%)	1458
Education and research facility	74(20.2%)	145(39.6%)	41(11.2%)	106(29.0%)	366
Business facility	54(15.3%)	139(39.3%)	48(13.6%)	113(31.9%)	354
Roadway	57(17.9%)	115(36.2%)	41(12.9%)	105(33.0%)	318
Neighborhood infrastructure	50(18.1%)	102(36.8%)	36(13.0%)	89(32.1%)	277

5.2. Classification Accurary

This study used the Naïve Bayes classifier to develop a predictive model for injured body parts by accident factors. A naïve Bayes-based classification model was constructed with accident type, type of work, facility type, injury source, and activity type as predictor variables and injured body part as an outcome variable. As shown in Table 3, the model's accuracy varied by class. Using the naive Bayes classifier, the accuracy of predicting the injured part with the arm by the predictor was the highest at 61.1%, and the accuracy of predicting the injury with the head was the lowest at 25.0%. Therefore, this study showed an average accuracy of 50.4%. For the upper extremity and lower extremity with a relatively large number of records, the accuracy was high, while in the case of the torso and head with a small number of records, the accuracy was relatively low.

Table 3. Confusion matrix

Actual class	Torso	Upper extremity	Head	Lower extremity	Actual Total
Torso	341 (37.3%)	159 (17.4%)	39 (4.3%)	376 (41.1%)	915 (17.9%)
Upper extremity	189 (9.6%)	1206 (61.1%)	90 (4.6%)	488 (24.7%)	1,973 (38.6%)
Head	106 (17.6%)	161 (26.7%)	151 (25.0%)	185 (30.7%)	603 (11.8%)
Lower extremity	262 (16.1%)	384 (23.6%)	98 (6.0%)	881 (54.2%)	1,625 (31.7%)
Predicted Total	898 (17.6%)	1,910 (37.3%)	378 (7.4%)	1,930 (37.7%)	5,116 (100%)

A naïve Bayes algorithm is a classification technique based on Bayes' Theorem, assuming independence among predictor variables. A naïve bayes classifier assumes that a variable in a class is not correlated to the presence of any other variable. The highly correlated variables such as facility type, type of work, activity type are voted twice in the model, leading to overinflating importance. To improve the accuracy of the model presented in this study, correlated variables need to be removed from the model.

6. DISCUSSION

This study is important in the following aspects. First, few studies focused on the prediction of the injured body part in construction accidents in Korea until recently. Previously, there were studies on predicting the severity of injury (death, severe injury, minor injury) and the accident type (fall, hit, slip down); however, no studies attempted to predict the injured body part. Second, using the results of this study, more detailed guidelines for construction safety management can be provided. The classification model for predicting the injured body part by environmental risk factors can help establish safety measures at the construction site. Information on which body parts are frequently injured by accident type, type of work, injury source, facility type, and activity type enables construction managers to manage in a more detailed manner. In addition, it will contribute to minimizing injuries if laborers are familiar with this information in advance.

The strengths of the naïve Bayes classifier are its simplicity, computational efficiency, and good classification performance. A sufficient amount of data in machine learning is essential in improving classification performance. Although approximately 5,000 cases were used in this study, more cases are likely to be needed to obtain satisfactory classification performance. Compared to the number of categories of variables, the number of cases appears insufficient. One of the variables used in this study has approximately 100 categories.

The data used in the analysis do not include information on the victims, which is a human-related factor. For privacy reasons, it is known that information about the victims is not disclosed. Therefore, this study only focused on the environmental risk factors rather than the human risk factors. A more accurate classification model can be built if the information on the victim, such as age, gender, occupation, experience, specialty, is included in the predictor variables.

There are many various data mining techniques. Techniques to be applied are different depending on the type of data, and accordingly, prediction and classification performance also varies. This study applied the naïve Bayes algorithm with excellent classification performance in a model with categorical variables. In future research, it is necessary to establish a predictive model with other data mining techniques and compare their performance.

7. CONCLUSION

This study developed a classification model to predict the injured body part of victims along with the construction site's accident-related factors. The naïve Bayes classifier, one of the machine learning algorithms, was employed to build a model. The injured body part, the outcome variables, has four categories: 1) torso, 2) upper extremity, 3) head, and 4) lower extremity. The predictor variables are accident type, type of work, injury source, facility type, and activity type. As a result of the study, the model's accuracy was different for each injured body part, and the average accuracy was 50.4%. The naïve Bayes classifier shows better classification performance as the number of cases increases. Although approximately 5,000 cases were used in this study, the classification accuracy was relatively low due to insufficient cases and categories of variables. Also, the data used in the analysis did not include human-related factors, so only the environmental-related factors were considered.

This study has practical significance by providing a classification model to predict the injured body part of construction accidents based on environmental factors on a construction site. The study results can help establish safety measures at the construction site by providing more detailed guidelines to injury characteristics related to the body part by accident type, type of work, injury source, facility type, and activity type, enabling construction managers to manage in a more detailed manner.

ACKNOWLEGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (No. 2019R1C1C1009979) and Incheon National University Research Grant in 2019.

REFERENCES

[1] The Center for Construction Research and Training(CPWR), “THE CONSTRUCTION CHART BOOK The U.S. Construction Industry and Its Workers Sixth Edition.”, 2018

[2] Ministry of Employment and Labor, “2019 industrial accident status analysis.”, 2020 (in Korean)

[3] Choi. J, Gu. B, Chin. S, Lee. J, “Machine learning predictive model based on national data for fatal accidents of construction workers”, Automation in Construction, vol. 110, 2020

[4] Lee J., Yoon Y., Oh T., Park S., Ryu S., “A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry”, Applied Sciences, vol. 10, no. 21, pp. 1-23, 2020

[5] Mistikoglu G., Gerek IH., Erdis E., Mumtaz Usmen P.E., Cakan H., Kazan E.E., “Decision tree analysis of construction fall accidents involving roofers”, vol. 42, no. 4, pp. 2256-2263, 2015

[6] Cho Y., Kim Y., Shin Y., “Prediction Model of Construction Safety Accidents using Decision Tree Technique.”, Journal of the Korea Institute of Building Construction, vol. 17, no. 3, pp. 295-303, 2017 (in Korean)

[7] Ayhan B.U., Tokdemir O.B., “Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial Neural Networks”, Journal of Construction Engineering and Management, vol. 146, no. 3, pp.

[8] Kang K., Choi J., Ryu H., “Analysis of the Feature Importance of Occupational Accidents Occurring at Construction Sites on the Severity of Lost Workdays.”, Journal of Korea Institute of Building Construction, vol. 21, no. 2, pp. 165-174, 2021 (in Korean)

[9] Choi S., “Aging Workers and Trade-Related Injuries in the US Construction Industry”, Safety and Health at Work, vol.6, no.2, pp. 151-155, 2015

[10] Sugama A., Ohnishi A., “Occupational Accidents Due to Stepladders in Japan: Analysis of Industry and Injured Characteristics”, Procedia Manufacturing, vol. 3, pp. 6632-6638, 2015

[11] Chi S., Han S., “Analyses of systems theory for construction accident prevention with specific reference to OSHA accident reports”, International Journal of Project Management, vol. 31, no. 7, pp. 1027-1041, 2013

[12] Amiri M., Ardeshir A., Fazel Zarandi M.H., Soltanaghaei E., “Pattern extraction for high-risk accidents in the construction industry: a data-mining approach”, International Journal of Injury Control and Safety Promotion, vol. 23, no. 3, pp. 264-276, 2016

[13] Halvani G.H., Jafarinodoushan R., Mirmohammadi S.J., Mehrparvar A.H., “A survey on occupational accidents among construction industry workers in Yazd city: Applying Time Series 2006-2011”, Journal of Occupational Health and Epidemiology, vol. 1, no. 1, pp. 1-8, 2012

[14] Amra I.A.A., Maghari A.Y.A., “Students performance prediction using KNN and Naïve Bayesian”, International Conference on Information Technology, 2017

[15] Shmueli G., Bruce P.C., Yahav I., Patel NR., Lichtendahl K.C.Jr., Data mining for business analytics: concepts, techniques, and applications in R, Wiley, New Jersey, U. S., 2018.

[16] Sharma M., Kumar N., Kumar P., “Badminton match outcome prediction model using Naïve Bayes and Feature Weighting technique”, Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 8441-8455, 2021

[17] Subbalakshmi G., Ramesh K., Rao M.C., “Decision Support in Heart Disease Prediction System using Naïve Bayes”, Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 170-176, 2011

[18] Zhang H., Ma J.X., Liu C.T., Ren J.X., Ding L., “Development and evaluation of in silico prediction model for drug-induced respiratory toxicity by using naïve Bayes classifier method”, Food and chemical toxicology, vol. 121, pp. 593-603, 2018

[19] Yang L., Fu B., Li Y., Liu Y., Huang W., Feng S., Xiao L., Sun L., Deng L., Zheng X., Ye F., Bu H., “Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a Naïve Bayes algorithm”, Computer Methods and Programs in Biomedicinet, vol. 192, 2020

[20] Hazra A., Mandal SK., Gupta A., “Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms”, International Journal of Computer Applications, vol. 145, no. 2, pp.39-45, 2016