2020

Applied machine learning in agro-manufacturing occupational incidents.

Fatemeh Davoudi Kakhki

Steven A. Freeman

Gretchen A. Mosher

Applied machine learning in agro-manufacturing occupational incidents.

Abstract

Commercial grain elevators are hazardous agro-manufacturing work environments where workers are prone to serious and life-threatening injuries. The aim of this study is to give insight into safety risks in grain handling facilities through information processing of workers' compensation data on agro-manufacturing occupational incidents within commercial grain elevators in the Midwest region of the United States between 2008 and 2016. The severity of occupational incidents is determined by total dollar amount incurred on medical, indemnity, and other expenses in workers' compensation claims. The most important factors that affect the cost escalation of occupational incidents are extracted using bootstrap partitioning method, and are applied as input for constructing two machine learning models: random forests decision trees, and naïve Bayes. Both models show high accuracy (87.64% and 92.78% respectively) in predicting that a future claim is classified as either low or medium, severity. The models contribute to identifying high injury risk groups, and prevalent incident causes, allowing a more research-based focused intervention effort in grain handling workplaces. In addition, the results are applicable in forecasting cost severity of future claims, and identifying factors that contribute to the escalation of claims costs.

Keywords

Machine Learning, Random Forest Decision Trees, Naïve Bayes, Agro-manufacturing Operations, Occupational Safety Analytics

Disciplines

Agriculture | Bioresource and Agricultural Engineering

Comments

This article is published as Kakhki, Fatemeh Davoudi, Steven A. Freeman, and Gretchen A. Mosher. "Applied machine learning in agro-manufacturing occupational Incidents." *Procedia Manufacturing* 48 (2020): 24-30. DOI: 10.1016/j.promfg.2020.05.016. Posted with permission.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 48 (2020) 24–30

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

48th SME North American Manufacturing Research Conference, NAMRC 48 (Cancelled due to COVID-19)

Applied Machine Learning in Agro-Manufacturing Occupational Incidents

Fatemeh Davoudi Kakhki[a,*], Steven A. Freeman[b], Gretchen A. Mosher[b]

[a] *Machine Learning & Safety Analytics Lab, Department of Technology, San Jose State University, San Jose, CA 95192, USA*
[b] *Department of Agricultural & Biosystems Engineering, Iowa State University, Ames, IA 50011, USA*

* Corresponding author. Tel.: +1-408-924-3195; *E-mail address:* fatemeh.davoudi@sjsu.edu

Abstract

Commercial grain elevators are hazardous agro-manufacturing work environments where workers are prone to serious and life-threatening injuries. The aim of this study is to give insight into safety risks in grain handling facilities through information processing of workers' compensation data on agro-manufacturing occupational incidents within commercial grain elevators in the Midwest region of the United States between 2008 and 2016. The severity of occupational incidents is determined by total dollar amount incurred on medical, indemnity, and other expenses in workers' compensation claims. The most important factors that affect the cost escalation of occupational incidents are extracted using bootstrap partitioning method, and are applied as input for constructing two machine learning models: random forests decision trees, and naïve Bayes. Both models show high accuracy (87.64% and 92.78% respectively) in predicting that a future claim is classified as either low or medium, severity. The models contribute to identifying high injury risk groups, and prevalent incident causes, allowing a more research-based focused intervention effort in grain handling workplaces. In addition, the results are applicable in forecasting cost severity of future claims, and identifying factors that contribute to the escalation of claims costs.

Keywords: Machine Learning; Random Forest Decision Trees; Naïve Bayes; Agro-manufacturing Operations; Occupational Safety Analytics.

1. Introduction

Each year, approximately 313 million occupational injuries occur globally [1]. Likewise, occupational injuries are a substantial and important part of the total U.S. injury burden [2]. Although injuries on the job are among the leading causes of death and disability around the world, their burden has historically been under-recognized [3]. Therefore, occupational injury analysis is necessary to provide insight about the factors affecting incident occurrence [4].

Agriculture-related industries are among the most dangerous working environments, and less is known about non-fatal injuries [5] . The challenge of lack of non-fatal injury data limits potential safety interventions and preventive measures in addressing safety risks in non-farm agricultural-related workplaces such as grain elevators [6], [7]. The grain handling industry in the U.S. is a hazardous work environment, with workers in these facilities constantly at risk of severe and life-threatening occupational injuries [6], [8] . In the U.S., the rapid expansion of the grain handling industry in recent years has resulted in an increase in occupational injuries and fatalities as compared to previous years [9], [10]. In addition, the frequency and costs of occupational injuries in grain handling industry are higher than other agribusinesses [10]–[14].

Despite the common applications of data mining in occupational incident analysis in different sectors, there is rare literature on evaluating the performance of decision trees and naïve Bayes algorithms in classifying and predicting the monetary severity levels of occupational incidents in grain elevators in the U.S. The aim of this study is to apply, validate and compare the performance of random forests decision trees and naïve Bayes in accurately classifying and predicting the financial severity outcomes for occupational incidents based on workers' compensation claims incurred amounts, using a data set with over 7,000 workers' compensation claims reported between 2008 and 2016. In addition, interpreting the results from models identifies high injury risk groups and prevalent causes of incidents, allowing more focused intervention efforts

10.1016/j.promfg.2020.05.016

in grain elevators. The results are applicable in forecasting costs severity of future claims and identifying factors that contribute to the escalation of claims costs. The overall research questions are:

- What are factors with highest contribution in predicting claims severity level of incidents in agro-manufacturing operations within grain elevators?
- What are the characteristics of various severity level claims in grain elevators' agro-manufacturing incidents?

2. Machine Learning in Occupational Incidents Analysis

Classification is an important data mining technique with broad applications in nearly every field [[15]]. Random forests decision trees and naïve Bayes are popular classification methods in various fields. To cite a few examples, they were successfully applied to classify and code injury narratives in workers' compensation claims [16]–[18]. They were applied to predict likelihood of future events in healthcare [16], [19], and analyze crash severity in transportation [20]–[22].

In occupational injury analysis, decision trees with different algorithms and naïve Bayes modeling are two commonly used and poplar methods [23]. Random forests decision trees were used to classify and predict the severity of injury in construction occupational incidents based on injured body parts and types of injury [24]. Compared to logistic regression method, decision trees and naïve Bayes methods showed superiority for accurate classification and prediction of causes of incidents in construction and mining industries [25]. Other examples include successful applications of random forests and naïve Bayes methods to classify and predict injury severity in mining industry [26], and develop predictive models of injury severity characteristics in construction industry [25], [27]–[29].

3. Materials and Methods

3.1. Data set and Data Processing

The data for this study were taken from a leading insurance company, specializing in agricultural commodities in the United States. According to this data, more than &78 million U.S. dollars was incurred in costs of occupational costs that occurred in commercial grain elevators and cooperatives over eight years, from 2008 to 2016.

3.1.1. Preliminary analysis of Data

Data for this analysis include three main injury types in grain elevators. Medical injuries have the highest frequency (5,942), followed by permanent partial disabilities (836). The least frequent injury type is temporary total or partial disability (629), out of all sample that includes 7,407 workers' compensation claims from 2008 to 2016 in agro-manufacturing operations occupational incidents within commercial grain handling facilities in the Midwest of the U.S. The records include the *type of injury*, *cause* and *cause groups of injury*, *nature* and *nature groups of injury*, *injured body part (s)* and *body part groups*, *claim status* (either closed or still open), the injured *age*, *tenure*, and *occupation class*. They also include the dollar amounts incurred on an incident with three groups of *medical* costs, *indemnity* costs, and *other* expenses. Table 1 shows the details of variables.

Table 1. Variables used in the Study Derived from the Original Dataset.

Variable Name	Variable Type	Description
Type of injury	categorical	if injury is medical or disability
cause	categorical	main cause of injury such as *slip, fall, trip, lifting, falling objects*, etc.,
Cause group	categorical	general category of main cause of injury such as *strain or injury by*, *struck or injured by*, etc.,
Nature	categorical	nature of injury such as contusion, concussion, carpel tunnel syndrome, etc.,
Nature group	categorical	general category of main nature of injury such as specific injuries, occupational diseases, cumulative injuries, etc.,
body part(s)	categorical	injured body part(s) such as knee, shoulders, lower back area, etc.,
Body part group	categorical	general category of injured body part(s) such as lower extremities, upper extremities, head, etc.,
Claim status	categorical	if the claim is open or closed
Occupation class	categorical	job category of injured worker such as grain milling, farm machinery operations, grain handling operations, etc.,
Age	numerical	age of injured worker
Tenure	numerical	experience (years) on the job for injured worker
Total incurred	numerical	monetary loss paid on a claim
Severity	categorical	financial severity of incidents as *low* or *medium*

A typical workers' compensation claim is monetary value, called "total incurred" amount, which is paid on medical costs, indemnity costs, and other expenses for an injured worker. The descriptive statistics per type of injury is shown in Table 2.

Table 2. Cost and frequency per type of injury.

Type of injury	Mean total incurred	Median total incurred	Frequency
Medical	$1,430.90	$389.38	5,942
Permanent partial disabilities	$34,363.63	$29,154.90	836
Temporary total or partial disability	$14,124.93	$6,608.46	629

The dependent variable in this study is the severity of the "total incurred" amounts. Therefore, the summation of medical, indemnity and other expenses are calculated to create a new categorical variable called *severity*. The new variable has two labels: low (L), and medium (M), for total incurred amounts of 0-10K, and 10-100K, respectively.

3.1.2. Feature Selection Analysis

Chi-square statistical test and bootstrap partitioning methods assess the contribution of the input variables in accurately predicting and classifying the output variable, which is severity levels in this study. The difference between chi-square statistical test and bootstrap partitioning for evaluating variable importance is that the former measures the effect of each input variable on the output individually (one-by-one analysis), while the latter considers the effect of a combination of the input variables on the response. Bootstrap partitioning method is a random selection of input variables. The random selection chooses different sets of inputs for each individual tree to reduce the collinearity effect. Multiple trees are made, each evaluating contribution of multiple inputs variables on response levels.

The contribution percentage of each predictor is determined via its chi-square statistics (χ^2). According to Fig. 1, the type of injury was the most important factor in determining the severity of an incident. The least important variable, yet statistically significant, was the nature group of injury. The results showed that the *claim status*, *injured body part*, *injury cause*, and *injury nature* were also predictors of the incident outcome.

Relying on the analysis results as shown in Fig. 1, the variables used in the study are *injury type*, *claim status*, *injured body part*, *injury cause*, and *injury nature*. These were selected out of all the variables in Table 1, as input/independent variables to predict the binary output/response, which is the severity of the occupational incidents in this data, based on the financial loss on the total incurred amount per claim.

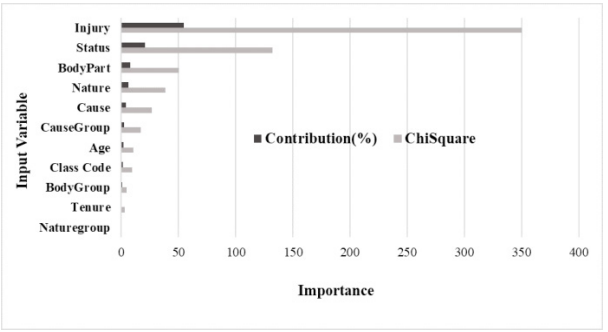

Fig. 1. Variable importance plot for bootstrap partitioning and chi-square

3.2. Applied Machine Learning Models

The same selected input variables are used to build NB and RF models. As mentioned earlier, both methods were selected because they are simple, fast, and effective for classification of categorical responses with multi-levels.

Data is partitioned into train, and test sets. The train set includes 70% of the data points (5,190 claims), and is used to fit the model of interest and estimate model parameters. The test set that includes 30% of data is used to assess the generalization error of the final model (2,217 claims). The performance of the model in the test set is the judgment criteria of how useful a predictive model is.

3.2.1. Naïve Bayes Modeling

One popular method for statistical inference such as classification in data processing is Bayes classifier principle [30]. Bayesian classifiers are based around the Bayes rule that uses conditional probabilities for classification of a categorical target variable based on the input variables [31]. The naïve Bayes (NB) classifier is one of the simplest and most popular Bayes classifiers [30]. NB modeling technique is adopted in many studies due to its simplicity, computational efficiency, and its high performance in classification tasks [32]. In addition, NB is one the fastest classifiers for prediction and classification purposes on large-scale data sets that can handle both categorical and continuous data [33]. Therefore, NB is proven to be a simple and effective classifier in text classification studies [21].

NB classifier assumes that variables are conditionally independent and, despite being a simplistic method, it reports high accurate performance in various classification tasks [34]. In other words, the NB algorithm reduces the complexity of Bayesian classifiers by making a conditional independence assumption that dramatically decreases the number of parameters to be estimated from the original *2(2n -1)* to just *2n* when modeling *P(X|Y)*. *X* is the independent variable, *Y* is the categorical response variable, and *n* is the number of independent variables used in the analysis [35], [36]. NB models usually have high accuracy with both binary and multi-class categorical response variables [37]. The NB model is used to assign a probability to each class of the response variables, and the class with the largest estimated probability is then chosen as the predicted class [38].

3.2.2. Random Forest Decision Trees

A decision tree (DT) is a commonly used methodology for building classification systems based on multiple covariates for the development of a predictive model for a target variable [39]. DTs are among the most popular predictive analytics techniques among practitioners due to being relatively straightforward to build and understand, as well as handling both nominal and continuous inputs [40]. Other advantages of classification via decision trees include the support for multi-level and nonlinear classification capability [41]. DTs do not need any assumptions regarding the distribution or the independence of the attributes. They do not involve any transformation of the variables, and are statistically useful methods for regression (with numerical response) and classification problems with categorical dependent variables [42]. DT algorithms simply split the dataset hierarchically and can be applied as a replacement for logistic or multiple regression [43]. According to Jones & Sall [44], in classification trees, where the response variable is categorical, the decision criteria for choosing the best split is the likelihood ratio *chi*-square. The node splitting is based on the *LogWorth* statistics which is defined as [*-log10 (P-value)*]; where the *P*-value is calculated so that it takes into account the number of different ways that splits can happen. The calculation includes an unadjusted *P*-value, which supports input variables with many levels, and the Bonferroni *P*-value, which favors input variables with small number of levels. The optimal split is the one that maximizes the *LogWorth* statistics.

The random forests (RF) method is a useful DT algorithm for classification and prediction problems [45]. The RF method has a set of characteristics that makes it advantageous [46]. As a powerful data driven method, the RF is non-parametric, has high predictive accuracy, and determines variable importance, which contributes to better understanding of the individual role of each input factor [47]. RF decision trees consist of a collection of arbitrary simple trees used to determine the outcome. RF decision trees are random since a subset of the input and output variables is used to build each individual tree, and also each split within each tree is created based on a subset of input variables, not all [48]. Building a large number of trees, the overall prediction of the forest is the average prediction of all individual trees. In classification, the ensembles of simple trees vote for the most popular class while in regression problems, the responses are averaged to obtain an estimate of the dependent variable. Applying the RF method is expected to significantly improve the prediction accuracy in classification problems [49], [50].

4. Results

The ML models were used to classify the binary severity response using the selected input variables. In this section, the performance of the ML models on the training, testing, and overall data sets is discussed. The quantitative measures of model performance were gained from the confusion matrices, which included the frequency of the binary response in actual and predicted classes. The model performance metrics are also explained. A discussion of the information gained from NB and RF models regarding the factors influential on predicting the injury severity outcomes, completes this section.

4.1. Model Performance Values

To compare classification models, various performance metrics gained from a confusion matrix are used typically. The confusion matrix for a binary classifier is shown in Table 3.

Table 3. Confusion matrix for binary classification.

Actual class	Predicted class	
	L (Negative)	M (Positive)
L (Negative)	TN	FP
M (Positive)	FN	TP

The confusion matrix, which has the form of a contingency table, shows how the observations are spread over actual classes (rows) and predicted classes. In this study, the binary confusion matrix for each ML model was used for calculating the model performance quantitative measures using the following equations:

$$Recall = TP/(FN + TP)$$

$$Specificity = TN/(TN + FP)$$

$$Precision = TP/(TP + FP)$$

$$F\text{-}score = 2(Precision*Recall)/(Precision + Recall)$$

$$Overall\ accuracy = (TN + TP)/Totall$$

4.2. Analysis and Model Evaluation

Data was split into training set (70%) that includes 5,190 incidents and testing set (30%) that has 2,217 incidents records. Assigning data points to the training and testing data sets was done using stratified re-sampling. The models that were built using the training data were then used on the testing data to evaluate their performance. Table 4 includes the results of models in classifying low severity and medium severity injuries in actual versus predicted relevant categories.

Table 4. Confusion matrix for both models (train vs test data).

Model	Actual class	Predicted (train)		Predicted (test)	
		L	M	L	M
NB	L	4,242	176	1,825	67
	M	161	611	93	232
RF	L	4,412	6	18,90	2
	M	647	125	272	53

Results from Table 4 were used to calculate the numerical values for recall, specificity, precision, *F*-score, overall accuracy, and overall error (misclassification) metrics per model. The main purpose of comparing model performance is determining the accuracy differences among all model types to choose the best model [50]. The prediction results on test data sets are presented in Table 5.

Recall or sensitivity shows the effectiveness of a classifier in identifying positive labels, which is M (medium severity) in Table 3. Specificity shows how effectively a classifier recognizes negative labels, which is L (low severity) in Table 2. Precision evaluates class agreement of the data labels with the positive labels defined by the classifier. F-score is a weighted average of the recall and precision. Overall accuracy shows how often the classifier is correct in overall while overall error rate shows how often the classifier is wrong in overall, which equals to 1 minus the overall accuracy rate [51–53].

Both RF and NB models have high prediction accuracy. The RF accuracy rate in test sets is 87.64%. The NB model outperforms the RF by almost 5%, with overall accuracy of 92.78% in test set. Based on the overall performance of both models, 87-93% of the claims are accurately classified in various severity levels based on the *type*, *cause*, and *nature of the injury*, *injured body part*, and *claim status*. Regarding per severity level prediction, both models show high performance in classifying low severity incidents (over 95%), while the NB outperforms the RF in predicting medium severity classes. Table 4 shows the results for per severity level classification and prediction accuracy for NB and RF models. The accuracy rates are gained based on the frequency of correct classification per class from the confusion matrix, which shows the number of correct and incorrect classified cases under a specific label [51].

Positive and negative classes in this study were considered as M, and L respectively. This was used in interpreting recall and specificity values. Recall value showed the models'

performance in classifying the M cases while specificity revealed the models' ability in classifying the L cases correctly.

Table 5. Model performance comparison on test data.

Model	Recall	Specificity	Precision	F-score	Overall accuracy	Overall error
NB	71.38 %	96.45 %	77.59 %	0.7435	92.78 %	7.22%
RF	16.30 %	99.98 %	96.36 %	0.2788	87.64 %	12.36 %

All models were capable of classifying L injuries with high accuracy between 87.64%% and 92.78%. This was expected due to the high frequency of L cases in the original data set.

Another metric used in this study was *F*-score. To evaluate the performance of a classifier, the *F*-score is one of the most useful measures since it is the harmonic mean of precision and recall. Overall, NB classifiers showed a higher *F*-score (0.74) compared to RF with values of 0.28. Considering *F*-score as a weighted measure of performance between recall and precision, the NB classifier showed a much higher performance in predicting the severity class of occupational incidents in this study.

4.3. Applications in Safety

Considering the overall model performance, both NB and RF models have high prediction accuracy, and provide useful information about the significant factors distinguishing severity of occupational incidents in grain elevators. The results from the NB model shows that, on average, the most significant variable in prediction of the injury severity level is the type of injury followed by injury nature, injured body part (s), injury cause, and claim status. The same variables, in the same order, are statistically important predictors of low and medium severity levels.

When the causes of incidents include struck or strained, electrical shock, crash of rail vehicle, temperature extremes, abnormal air pressure, twisting, and lifting objects, most cases are high severity. In the NB model, the distinguishing factor between various severity levels is the type of injury. Medical injuries are generally predicted to end with low severity while temporary total or partial disabilities are predicted as medium severity. Permanent partial disabilities are predicted as medium severity in closed claims with 52% chance. Those incidents caused by caught in/between, collision with fixed objects, steam or hot fluid, welding operations, object handling, cut and puncture in foot, thumb, upper back area, wrists, and fingers are also mainly predicted to have either low or medium severity. For wrists and fingers, in particular, the medium severity claims have the open status, and those injuries with closed status are predicted as low severity.

The results from the RF model agree with those of the NB when estimating the probability of low and medium severity incidents. The injury type is the most significant factor in determining the severity level (64.4%). Medical injuries with open status have 72% chance of turning as low, and 12% chance of turning as medium severity. However, closed medical injuries are all predicted as low severity. Similarly, temporary total or partial disabilities are predicted as either low or medium based on the claim status. Since the injury type and claim status together explain 78% of the variation in the severity levels, the chance of incidents being low or medium severity are not highly affected by cause, nature, or body part(s), according to the RF model. Furthermore, Table 2 shows the results of tabulating the injury cause groups versus the predicted severity classes from the NB model. The injury causes groups with highest chance of low severity among all groups are *cut, puncture, scrape, miscellaneous causes*, and *heat or cold exposures* with probability over 90%. The highest chance for a medium severity injury is when it is caused by *fall, slip, or trip Injury*, or *motor vehicle* (23.4% and 18.9%).

Table 2. NB model predicted injury cause groups per severity class.

Injury Cause Group	L	M
Heat or cold exposures	91.9	6.36
Caught in, under, or between	84.7	13.7
Cut, puncture, scrape	99.2	0.84
Fall, slip, or trip injury	73.1	23.4
Miscellaneous causes	92.0	7.31
Motor vehicle	75.2	18.9
Rubbed or abraded by	84.6	15.4
Strain or injury by	77.4	22.0
Striking against or stepping on	92.0	7.25
Struck or injured by	89.8	8.31

5. Conclusion

The initial intent of this study was to identify what factors from workers' compensation data, as the source of injury details, affect the financial severity of the incidents that occurred in agro-manufacturing operations within grain elevators in the Midwest region of the United States. The analyses show that the type of injury, injured body parts, as well as injury cause and nature are statistically significant predictors of claim severity. Claim status is an important factor in escalation of the claim costs, as well. Although the naïve Bayes model shows a better performance considering overall and per severity class accuracy, the random forests decision trees are still reliable in accurate classification and prediction of future claims severity with the overall accuracy rate of 85%.

While this study is not specifically designed for suggesting interventions, its methods and approach may contribute to the efforts of safety practitioners by providing quantitative research-based information about the dominant and important safety risk factors in agro-manufacturing operations. The results can be applied in identifying the risk factors of incidents that escalate the incident costs, allowing research-based focused intervention efforts and strategy planning based on empirical data analysis. Integrating the extracted information from empirical data analyses with the knowledge of safety professionals and practitioners and the training and education

of employees are expected to decrease the rate and alleviate the severity outcomes of occupational incidents in grain elevators, and other industries.

References

1. Wang, H., Chen, G., Wang, Z., Zheng, X.: Socioeconomic inequalities and occupational injury disability in China: A population-based survey. Int. J. Environ. Res. Public Health. (2015). https://doi.org/10.3390/ijerph120606006.
2. Smith, G.S., Sorock, G.S., Wellman, H.M., Courtney, T.K., Pransky, G.S.: Blurring the distinctions between on and off the job injuries: Similarities and differences in circumstances, (2006). https://doi.org/10.1136/ip.2006.011676.
3. Marucci-Wellman, H.R., Courtney, T.K., Corns, H.L., Sorock, G.S., Webster, B.S., Wasiak, R., Noy, Y.I., Matz, S., Leamon, T.B.: The direct cost burden of 13 years of disabling workplace injuries in the U.S. (1998-2010): Findings from the Liberty Mutual Workplace Safety Index. J. Safety Res. (2015). https://doi.org/10.1016/j.jsr.2015.07.002.
4. Carrillo-Castrillo, J.A., Pérez-Mira, V., Pardo-Ferreira, M. del C., Rubio-Romero, J.C.: Analysis of Required Investigations of Work-Related Musculoskeletal Disorders in Spain. Int. J. Environ. Res. Public Health. (2019). https://doi.org/10.3390/ijerph16101682.
5. Scott, E., Bell, E., Hirabayashi, L., Krupa, N., Jenkins, P.: Trends in Nonfatal Agricultural Injury in Maine and New Hampshire: Results From a Low-Cost Passive Surveillance System. J. Agromedicine. (2017). https://doi.org/10.1080/1059924X.2017.1282908.
6. Issa, S.F., Cheng, Y.H., Field, W.E.: Summary of agricultural confined-space related cases: 1964-2013. J. Agric. Saf. Health. (2016). https://doi.org/10.13031/jash.22.10955.
7. Ramaswamy, S.K., Mosher, G.A.: Using workers' compensation claims data to characterize occupational injuries in the commercial grain elevator industry. J. Agric. Saf. Health. (2017). https://doi.org/10.13031/jash.12196.
8. Davoudi Kakhki, F., Freeman, S.A., Mosher, G.A.: Use of Neural Networks to Identify Safety Prevention Priorities in Agro-Manufacturing Operations within Commercial Grain Elevators. Appl. Sci. 1–16 (2019).
9. Riedel, S.M., Field, W.E.: Summary of Over 800 Grain Storage and Handling-related Entrapments and Suffocations Documented in the U.S. between 1970 and 2010. In: Efficient and safe production processes in sustainable agriculture and forestry XXXIV CIOSTA CIGR V Conference (2011).
10. Davoudi Kakhki, F., Freeman, S.A., Mosher, G.A.: Use of Logistic Regression to Identify Factors Influencing the Post-Incident State of Occupational Injuries in Agribusiness Operations. Appl. Sci. 9, 3449 (2019). https://doi.org/10.3390/app9173449.
11. Geng, Y., Dee Jepsen, S.: Current grain storage and safety practices of Ohio cash grain operators. J. Agric. Saf. Health. (2018). https://doi.org/10.13031/jash.12574.
12. Davoudi Kakhki, F., A. Freeman, S., A. Mosher, G.: Segmentation of Severe Occupational Incidents in Agribusiness Industries Using Latent Class Clustering. Appl. Sci. 9, 3641 (2019). https://doi.org/10.3390/app9183641.
13. Mosher, G.A., Keren, N., Freeman, S.A., Hurburgh, C.R.: Development of a safety decision-making scenario to measure worker safety in agriculture. J. Agric. Saf. Health. (2014). https://doi.org/10.13031/jash.20.10358.
14. Davoudi Kakhki, F., Freeman, S., Mosher, G.: Analyzing Large Workers' Compensation Claims Using Generalized Linear Models and Monte Carlo Simulation. Safety. 4, 57 (2018). https://doi.org/10.3390/safety4040057.
15. Sherekarar, Patil, T.: Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. Int. J. Comput. Sci. Appl. (2013). https://doi.org/10.1108/IJCST-09-2015-0100.
16. Chen, L., Vallmuur, K., Nayak, R.: Injury narrative text classification using factorization model. BMC Med. Inform. Decis. Mak. (2015). https://doi.org/10.1186/1472-6947-15-S1-S5.
17. Marucci-Wellman, H.R., Lehto, M.R., Corns, H.L.: A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accid. Anal. Prev. (2015). https://doi.org/10.1016/j.aap.2015.06.014.
18. Marucci-Wellman, H.R., Corns, H.L., Lehto, M.R.: Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. Accid. Anal. Prev. (2017). https://doi.org/10.1016/j.aap.2016.10.014.
19. Teimouri, M., Farzadfar, F., Alamdari, M.S., Hashemi-Meshkini, A., Alamdari, A., Rezaei-Darzi, E., Varmaghani, M., Zeynalabedini, A.: Detecting diseases in medical prescriptions using data mining tools and combining techniques. Iran. J. Pharm. Res. (2016).
20. Alikhani, M., Nedaie, A., Ahmadvand, A.: Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran. Saf. Sci. (2013). https://doi.org/10.1016/j.ssci.2013.06.008.
21. Liu, B., Blasch, E., Chen, Y., Shen, D., Chen, G.: Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier. In: Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013 (2013). https://doi.org/10.1109/BigData.2013.6691740.
22. Delen, D., Tomak, L., Topuz, K., Eryarsoy, E.: Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. J. Transp. Heal. (2017). https://doi.org/10.1016/j.jth.2017.01.009.
23. Davoudi Kakhki, F., Freeman, S.A., Mosher, G.A.: Evaluating machine learning performance in predicting injury severity in agribusiness industries. Saf. Sci. 117, 257–262 (2019). https://doi.org/10.1016/j.ssci.2019.04.026.
24. Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D.: Application of machine learning to construction injury prediction. Autom. Constr. 69, 102–114 (2016). https://doi.org/10.1016/j.autcon.2016.05.016.
25. Rivas, T., Paz, M., Martín, J.E., Matías, J.M., García, J.F., Taboada, J.: Explaining and predicting workplace accidents using data-mining techniques. Reliab. Eng. Syst. Saf. (2011). https://doi.org/10.1016/j.ress.2011.03.006.
26. Sanmiquel, L., Rossell, J.M., Vintró, C.: Study of Spanish mining accidents using data mining techniques. Saf. Sci. (2015). https://doi.org/10.1016/j.ssci.2015.01.016.
27. Chokor, A., Naganathan, H., Chong, W.K., Asmar, M. El: Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning. In: Procedia Engineering (2016). https://doi.org/10.1016/j.proeng.2016.04.200.
28. Chen, H., Luo, X.: Severity Prediction Models of Falling Risk for Workers at Height. In: Procedia Engineering (2016). https://doi.org/10.1016/j.proeng.2016.11.642.
29. Yi, W., Chan, A.P.C., Wang, X., Wang, J.: Development of an early-warning system for site work in hot and humid environments: A case study. Autom. Constr. (2016). https://doi.org/10.1016/j.autcon.2015.11.003.
30. Mujalli, R.O., López, G., Garach, L.: Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 88, 37–51 (2016). https://doi.org/10.1016/j.aap.2015.12.003.
31. Troussas, C., Virvou, M., Espinosa, K.J., Llaguno, K., Caro, J.: Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning. In: IISA 2013 - 4th International Conference on Information, Intelligence, Systems and Applications (2013). https://doi.org/10.1109/IISA.2013.6623713.
32. Kwon, O.H., Rhee, W., Yoon, Y.: Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. (2015). https://doi.org/10.1016/j.aap.2014.11.005.
33. Kumar Bhowmik, T.: Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. Intel. Artif. (2015). https://doi.org/10.4114/intartif.vol18iss56pp14-30.
34. Moreira, M.W.L., Rodrigues, J.J.P.C., Oliveira, A.M.B., Saleem, K., Neto, A.: Performance evaluation of predictive classifiers for pregnancy care. In: 2016 IEEE Global Communications Conference, GLOBECOM 2016 - Proceedings (2016). https://doi.org/10.1109/GLOCOM.2016.7842136.
35. Ng, A.Y., Jordan, M.I.: On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes. In: Advances in Neural Information Processing Systems (2002).
36. Mitchell, T.M.: CHAPTER 1 GENERATIVE AND DISCRIMINATIVE CLASSIFIERS : NAIVE BAYES AND LOGISTIC REGRESSION Learning Classifiers based on Bayes Rule. Mach. Learn. (2010). https://doi.org/10.1093/bioinformatics/btq112.
37. Marucci-Wellman, H.R., Corns, H.L., Lehto, M.R.: Classifying injury narratives of large administrative databases for surveillance-A practical approach combining machine learning ensembles and

human review. Accid. Anal. Prev. (2017). https://doi.org/10.1016/j.aap.2016.10.014.
38. Marucci-Wellman, H.R., Lehto, M.R., Corns, H.L.: A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accid. Anal. Prev. (2015). https://doi.org/10.1016/j.aap.2015.06.014.
39. Song, Y.Y., Lu, Y.: Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry. (2015). https://doi.org/10.11919/j.issn.1002-0829.215044.
40. Abbott, D.: Applied predictive analytics: Principles and techniques for the professional data analyst. (2014). https://doi.org/10.1002/ejoc.201200111.
41. Cui, Z., Chen, W., He, Y., Chen, Y.: Optimal Action Extraction for Random Forests and Boosted Trees. Presented at the (2015). https://doi.org/10.1145/2783258.2783281.
42. Mistikoglu, G., Gerek, I.H., Erdis, E., Mumtaz Usmen, P.E., Cakan, H., Kazan, E.E.: Decision tree analysis of construction fall accidents involving roofers. Expert Syst. Appl. (2015). https://doi.org/10.1016/j.eswa.2014.10.009.
43. Voznika, F., Viana, L.: Data mining classification. Springer. (2001).
44. Jones, B., Sall, J.: JMP statistical discovery software. Wiley Interdiscip. Rev. Comput. Stat. (2011). https://doi.org/10.1002/wics.162.
45. Bharathidason, S., Jothi Venkataeswaran, C.: Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees. Int. J. Comput. Appl. (2014). https://doi.org/10.5120/17749-8829.
46. Goldstein, B.A., Polley, E.C., Briggs, F.B.S.: Random forests for genetic association studies, (2011). https://doi.org/10.2202/1544-6115.1691.
47. Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., Ribeiro, L.: Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Sci. Total Environ. (2014). https://doi.org/10.1016/j.scitotenv.2014.01.001.
48. Grömping, U.: Variable importance assessment in regression: Linear regression versus random forest. Am. Stat. (2009). https://doi.org/10.1198/tast.2009.08199.
49. Ilersic, A.R., Griffin, J.I.: Statistics: Methods and Applications. J. R. Stat. Soc. Ser. A. (2006). https://doi.org/10.2307/2982585.
50. Oztekin, A., Al-Ebbini, L., Sevkli, Z., Delen, D.: A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. Eur. J. Oper. Res. (2018). https://doi.org/10.1016/j.ejor.2017.09.034.
51. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. (2009). https://doi.org/10.1016/j.ipm.2009.03.002.
52. Guns, R., Lioma, C., Larsen, B.: The tipping point: F-score as a function of the number of retrieved items. Inf. Process. Manag. (2012). https://doi.org/10.1016/j.ipm.2012.02.009.
53. Shreve, J., Schneider, H., Soysal, O.: A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. Decis. Support Syst. (2011). https://doi.org/10.1016/j.dss.2011.08.001.