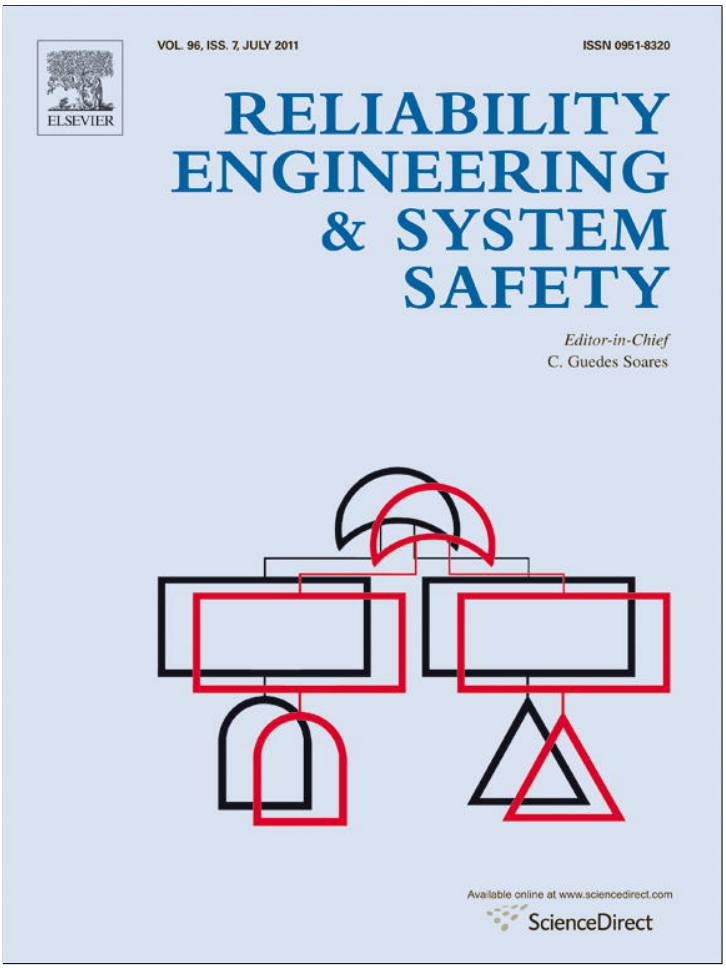

Reliability Engineering and System Safety 96 (2011) 739–747

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Explaining and predicting workplace accidents using data-mining techniques

T. Rivas [a,*], M. Paz [a], J.E. Martín [b], J.M. Matías [c], J.F. García [b], J. Taboada [a]

[a] Dpto. Ingeniería de los Recursos Naturales y Medio Ambiente, E.T.S.I. Minas, University of Vigo, Campus Lagoas, 36310 Vigo, Spain
[b] CIPP International, S.L. Parque Tecnológico de Asturias, Parcela 43, Oficina 11, 33428 Llanera, Spain
[c] Dpto. Estadística e Investigación Operativa, E.T.S.I. Minas, University of Vigo, Campus Lagoas, 36310 Vigo, Spain

ARTICLE INFO

Article history:
Received 19 January 2010
Received in revised form
28 February 2011
Accepted 2 March 2011
Available online 10 March 2011

Keywords:
Workplace accidents
Classification trees
Data mining
Bayesian networks
Support vector machines
Mine and construction safety

ABSTRACT

Current research into workplace risk is mainly conducted using conventional descriptive statistics, which, however, fail to properly identify cause-effect relationships and are unable to construct models that could predict accidents. The authors of the present study modelled incidents and accidents in two companies in the mining and construction sectors in order to identify the most important causes of accidents and develop predictive models. Data-mining techniques (decision rules, Bayesian networks, support vector machines and classification trees) were used to model accident and incident data compiled from the mining and construction sectors and obtained in interviews conducted soon after an incident/accident occurred. The results were compared with those for a classical statistical techniques (logistic regression), revealing the superiority of decision rules, classification trees and Bayesian networks in predicting and identifying the factors underlying accidents/incidents.

1. Introduction

A total of 922,253 workplace accidents resulting in lost workdays occurred in Spain in 2008; 194,248 occurred in the construction sector and 3255 occurred in the mineral extraction sector, representing falls of 23.3% and 14%, respectively, over 2007 [1]. The fall in 2008 compared to 2007 in terms of incidence (i.e., number of workplace accidents/population with social security accident insurance × 100,000) was 10.3%. This fall may be explained by the greater efforts of companies to ensure the safety of their workers by implementing preventive measures, whether on their own initiative or in response to legal obligations. Indeed, the fall may respond to Law 32/2006 governing construction sector subcontracting [2] and Royal Decree 1109/2007 deploying this law [3]. Subcontracting is a normal practice in Spain for construction and earth movement activities (in both the civil engineering and mining sectors), as it generally results in greater business efficiency. The new legislation was approved to ensure compliance with worker health and safety standards throughout the subcontracting chain. Royal Decree 1109/2007, passed in mid-2007 (and therefore in force when collecting data for this research) requires guarantees that ensure that any loss of control in the subcontracting regime does not result in health and safety risks for workers. The new legislation may be partly responsible for the fall in industrial accidents in the second half of 2007 and early 2008. The most recent statistics available, referring to the period April 2009–March 2010, reveal the same falling trend, both in the total number of accidents and in incidence: 4130.7 in 2009, a 18.5% lower than 2008.

Despite this fall, the issue of workplace safety continues to be a priority in social and economic policies and so requires in-depth studies that enable the causes of accidents to be accurately identified so that more effective measures and standards can be implemented.

Several statistical methods have been used in the workplace accident prevention field to process data. Analyses are usually descriptive, resulting in data in the form of historical summaries, percentages and indexes [4,5], or are based on linear models that evaluate the association between accidents and potential causes identified a priori ([6] and references therein). A priori, however, linear models excessively restrict how complex relationships between accidents and possible causes are modelled and may, in fact, fail to detect factors with a bearing on accidents if the relationship is non-linear.

The possibility for explaining and interpreting workplace accidents in terms of the entire range of possible causes is thus limited; this, in turn, conditions any working hypothesis aimed at predicting and reducing accidents. More sophisticated tools are thus needed that would enable full use to be made of information on accidents in terms of assessing dependence relationships between all the variables under consideration.

* Corresponding author. Tel.: +34 986811922; fax: +34 986811924.
E-mail addresses: trivas@uvigo.es (T. Rivas), mpaz.minas@gmail.com (M. Paz), jmartin@cippinternacional.com (J.E. Martín), jmmatias@uvigo.es (J.M. Matías), jgarcia@cippinternacional.com (J.F. García), jtaboada@uvigo.es (J. Taboada).

0951-8320/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2011.03.006

Data mining, which is an important discipline in fields such as medicine, engineering and finance, offers very positive results (see [7–9]). As for workplace risk management, studies have been conducted to assess the usefulness of such techniques in terms of their predictive power [10] and explanatory capacity [11]. Decision rules, for example, have been found to be particularly useful in identifying working conditions in the construction field that are associated with greater accident risk [12,13]. These studies, based on data from accident reports and interviews with workers and employers, confirm the advantages of data mining over conventional statistics in terms of the predictive function and the possibility of identifying interactions between variables with a bearing on accidents.

Our research aims to identify, from among a preselected group of methodologies, the data-mining techniques that extract the most useful information on workplace accidents from a database created from a survey of incidents/accidents in mining and civil construction companies, two sectors which, in Spain, head the list in terms of accidents involving lost workdays (data for 2008 from the Ministry of Labour and Immigration).

Our work is in a pilot phase aimed at evaluating the techniques available and contributing to the gradual development of a structured methodology for analysing workplace accidents that can eventually be safely applied to the design and planning of large-scale and far more costly studies. The potential of data-mining techniques not only derives from the possibility for processing large quantities of data but also from the following:

(1) their capacity to deal with large-dimension problems, which is necessary when endeavouring to identify relevant variables among a large number of potential factors;
(2) their flexibility in reproducing the data-generation structure, irrespective of complexity, thanks to a non-linear structure that is adaptable to the data (non-parametric philosophy);
(3) their great predictive and, in some cases, interpretative, potential.

In our research we evaluated Bayesian networks, decision rules, classification trees, logistic regression and support vector machines, with a view to ultimately reducing workplace accident rates by enabling preventive measures to be concentrated in an effective way in areas of greatest risk.

2. Materials and methods

2.1. Information sources and data description

The information used in this study was obtained from a survey carried out among workers employed in two companies—one in the mining sector and the other in the civil engineering sector—belonging to the same group and with operations in the Aragón, Asturias and Valencia regions of Spain.

In the mining company, workers from four operations were interviewed: an opencast coalmine employing 48 workers, two opencast ceramic-quality clay pits employing 72 workers and a quarry employing 13 workers and producing quartzite for use as a filler for public works and aggregates. The public works company surveyed, which removes earth in road, motorway and railroad (including high-speed train railroad) construction works, normally employs about 50 workers when working at full capacity.

Delivered and circulated in these companies, between September 2007 and March 2008, was a questionnaire with 20 questions to be completed whenever an accident/incident occurred. The questions covered issues related to the circumstances of the accident/incident, the worker, the kind of activity, work conditions and compliance with regulations. An incident was defined as any unexpected deviation from work procedures that might have caused an accident, and an accident was defined as a deviation from working standards or procedures affecting the health or safety of a worker [14].

A total of 62 completed questionnaires, each corresponding to a single accident/incident at different work stations and in different operations, were returned; of these, 18 referred to accidents and 44 to incidents. The severity of the accidents was not recorded. Each of the fields in the survey represented a study variable and, in total, information was obtained on 17 variables that were categorized in three groups referring to the event, the worker and the company (in terms of overall job and risk management). The variables for each category were as follows (see Table 1):

2.1.1. Event (three variables)

- Time of day (HRD): first thing (FH), after lunch (AL), last 2 h of work (LH), overtime hours (OT), or other (OTH).
- Day of the week (DAY): Monday (M), Tuesday (TU), Wednesday (W), Thursday (TH), Friday (F), or Saturday (S).
- Month (MTH): September (S), October (O), November (N), December (D), January (J), February (F), or March (M).

2.1.2. Worker (nine variables)

- *Worker age* (*AGE*): Below 27 (s1_below_27), 27–32.5 years (s2_27_32c5), 32.5–40.5 years (s3_32c5_40c5), or over 40.5 years (s4_40c5_up).
- *Worker nationality* (*NAT*): Native-born Spanish (SP), Eastern European (Armenia, Bulgaria, Poland and Romania) (EE), African (Libya and Morocco) (AF), or Latin American (Chile, Colombia and Ecuador) (LA).
- *Job type* (*JOB*): Job that involved handling vehicles (VH), job that involved handling heavy machinery (MC), office job involving no use of machines (OF), other job (electricians, mechanics, etc.) (OTH), or unspecified job (NS).
- *Length of time in the company (TCO)*: Less than 6 months (s1_below_0c5), 6–12 months (s2_0c5_1), 1–1.5 years (s3_1_1c5), 1.5–2 years (s4_1c5_2), or more than 2 years (s5_2_up).
- *Length of time doing the specific job associated with the accident* (*TJB*): Less than 1 week (s1_below_1W), 1–4 weeks (s2_1W_1M), or more than 4 weeks (s3_1M_up).
- *Accident risk training* (*ATR*) *received*: General training and training specific to the post (applied and practical on-the-job training, in accordance with Law 31/1995 [15]) (GS), general theoretical training (GT), or no training (NT).
- *Type of employment contract* (*ECT*): Temporary employment for a specific purpose (SP), temporary employment (TE), or permanent employment (PE).
- *Job-associated risk awareness* (*RAW*): The worker perceived the risk (YES), or the worker did not perceive the risk (NO).
- *Personal factors* (*PFA*): Personal factors contributed to the accident (YES), or personal factors did not contribute to the accident (NO).

Note that these last two variables are subjective, reflecting, in turn, whether the worker was aware of any risk prior to the accident/incident, and the undeniable influence on accidents/incidents of factors such as worker diligence, stress, boredom, etc.

Table 1
Variables selected for study.

Variable	Values
(a) Event-related variables ($n=3$)	
Time of day (HRD)	First Thing (FH), After Lunch (AL), Last 2 hours of work (LH), Overtime hours (OH), Other (OTH)
Day of the week (DAY)	Monday (M), Tuesday (TU), Wednesday (W), Thursday (TH), Friday (F), Saturday (S)
Month (MTH)	September (S), October (O), November (N), December (D), January (J), February (F), March (M)
(b) Worker-related variables ($n=9$)	
Worker age (AGE)	Below 27 (s1_below_27) 27–32.5 (s2_27_32c5) 32.5–40.5 years (s3_32c5_40c5) Over 40.5 years (s4_40c5_up)
Worker nationality (NAT)	Native-born Spanish (SP) Eastern European (Armenia, Bulgaria, Poland and Romania) (EE) African (Libya and Morocco) (AF) Latin America (Chile, Colombia and Ecuador) (LA)
Job type (JOB)	Job that involve handling vehicles (VH) Job that involve handling heavy machinery (MC) Office job involving no use of machines (OF) Other job (electricians, mechanics, etc) (OTH) Unspecified job (NS)
Length of time in the company (TCO)	Less than 6 months (s1_below_0c5) 6–12 months (s2_0c5_1) 1–1.5 years (s3_1_1c5) 1.5–2 years (s4_1c5_2) More than 2 years (s5_2_up)
Length of time doing the job associated with the accident (TJB)	Less than 1 week (s1_below_1W) 1—4 weeks (s2_1W_1M) More than 4 weeks (s3_1M_up)
Accident risk training (ATR)	General training and job-specific training (GS) General theoretical training (GT) No training (NT)
Type of employment contract (ECT)	Temporary employment for a specific purpose (SP) Temporary employment (TE) Permanent employment (PE)
Job-associated risk awareness (RAW)	The worker perceived the risk (YES) The worker did not perceive the risk (NO)
Personal factors (PFA)	Personal factors contributed to the accident (YES) Personal factors did not contribute to the accident (NO)
(c) Company-related variables ($n=5$)	
Task duration in hours (TKH)	Less than 4 h (s1_below_4) 4–8 h (s2_4_8), Longer than 8 h (s3_8_up)
Company contractual status (CCS)	The employing company was subcontracted (SB) The employing company was the main contractor (MN)
Risk management and supervision at the accident site (MAS)	There was supervision and/or control (SC) There was a risk management policy but no supervision and/or control (RP) There was no risk management policy and so no supervision and/or control (NP)
Risk assessment in accordance with Royal Decree 39/1997 [11] (RKA)	Risk had been assessed (YES) Risk had not been assessed (NO)
Job-related protective measures (JBP)	Individual and collective protective equipment was provided (YES) Individual and collective protective equipment was not provided (NO)
(d) Prediction variable	
Event (EVT)	Accident (A), Incident (I)

2.1.3. Company (five variables)

- Task duration in hours (TKH): less than 4 h (s1_below_4), 4–8 h (s2_4_8), or longer than 8 h (s3_8_up). Note that this variable has a direct bearing on the use of protective measures.
- *Company contractual status* (*CCS*): The employing company was subcontracted (SB), or the employing company was the main contractor (MN).
- *Risk management and supervision at the accident site* (*MAS*): There was supervision and/or control (SC), there was a risk management policy but no supervision and/or control (RP), or there was no risk management policy and so no supervision and/or control (NP). Note that this variable reflects whether a company had health and safety personnel.
- *Risk assessment in accordance with Royal Decree 39/1997* [16] (*RKA*): Risk had been assessed (YES), or risk had not been assessed (NO).
- *Job-related protective measures* (*JBP*): Individual and collective protective equipment was provided (YES), or individual and collective protective equipment was not provided (NO).

Included also was a response variable called event (EVT), which could take an accident (A) or incident (I) value.

2.2. Methodology

The different data-mining techniques used the tools described below. In order to select the variables, determine their relevance and estimate (most of) the data-mining models, we used the Waikato Environment for Knowledge Analysis (WEKA) freeware developed by the University of Waikato [7]. We also used the Genie (Graphical Network Interface) software developed by the University of Pittsburgh to construct and train the Bayesian networks.

Our research was structured in two phases as follows: an initial phase in which the different methods were applied in order to select the most relevant variables, and a second phase that consisted of applying—using the most relevant variables—different data-mining techniques so as to determine their capacity for predicting an accident/incident and to assess their capacity for explaining the event.

2.2.1. Variable selection

A first phase was aimed at identifying the most relevant of the 17 variables in terms of explaining the response variable. The WEKA software offers a number of variable selection methods for studying the relevance of database variables in relation to a response.

A total of three tests using cross-validation were performed with 10, 5 and 3 groups. The WEKA attributes selection function (select attributes) grouped 31 different methods in two distinct categories based on either of the following:

- The frequency with which each variable was selected in each cross-validation training round. For this group we used:
  - CfsSubSetEval, which selects variables according to the level of correlation with the response variable.
  - WrapperSubSetEval, which selects variables on the basis of the improvement in predictive capacity brought about by the incorporation of each variable in 19 different data-mining models.
- An indicator of the merit of each variable that produced a ranking of variables in terms of relevance. A total of six such selection models were used.

The evaluation options, classifiers and search methods used in some of the models are listed in Table 2.

2.2.2. Model application and predictive and explanatory capacity analysis

Once the variables were ordered according to weight-relevance for the event prediction variable, 17 different data-mining models were applied, resulting in a percentage of correctly classified cases and a confusion matrix for each.

With a view to testing whether the number of variables had a bearing on model predictive and explanatory capacities, the models were tested in three different ways: by including the seven most relevant causal variables obtained in the first phase of the study along with the prediction variable; by introducing the first 10 causal variables along with the prediction variable; and finally, by introducing all the variables.

The data-mining techniques used to model the influence of the explanatory variables on the response variable (event) were decision rules, classification trees, Bayesian networks, support vector machines and logistic regression (see, in addition to the specific references provided below [7–9]).

- *Logistic regression*: Logistic regression [17] is a generalised linear regression technique which, rather than directly modelling the response variable in terms of the covariables, models the logarithm of the odds ratio, which represents the probability of occurrence of the class of interest (accident) versus the probability of the occurrence of another class (incident). Logistic regression is estimated using the maximum likelihood method.
- *Decision rules*: Decision rules produce rules of the form "if A then B" ($A=>B$), where the antecedent is a combination of values for the explanatory variables and the consequent is a value for the response variable (in our case, an event in the form of a workplace accident/incident). The goodness of rules obtained from the data were evaluated as follows:
  - *Coverage or support*: The proportion of data instances in which the rule is satisfied. Thus, the coverage of $A=>B$ is the proportion of data instances that verify both A and B. It can be written as *coverage* $(A=>B)=P(A\cap B)$, where the probability is understood to refer to the data sample.
  - *Confidence or precision*: The percentage of correct predictions of a rule when it can be applied. Thus, the confidence that $A=>B$ is the proportion of data instances that satisfy both A and B in the set of instances that satisfy A. This number coincides with coverage divided by the number of instances that satisfy A. It can be written as *confidence* $(A=>B)=P(B|A)=P(A\cap B)/P(A)$.

  For our research we used the following WEKA algorithms:
  - *OneR algorithm* [18]: This algorithm, equivalent to a single-level classification tree, tests all the possible attribute-value pairs and selects that producing the least error.
  - *PART algorithm* [19]: Based on partial C4.5 decision trees [22], this algorithm combines splitting and covering operations.
- *Classification trees*: Classification trees [20] are statistical classification techniques that can be graphically represented as diagrams. There are different kinds of trees, but they are all generally trained by progressively dividing the data into groups—on the basis of some condition in regard to one of the explanatory variables—in such a way that each group is as similar as possible in terms of the response variable. Each group obtained in the previous stage is divided again, with a view to enhancing similarity, using a new condition based on an explanatory variable, and so on successively until some stop criterion is satisfied. The classification trees used in our research were WEKA-implemented algorithms as follows:
  - *ID3 algorithm* [21]: This antecedent of the C4.5 algorithm, based on the divide-and-conquer method, can be applied with discrete variables and binary-type objective variables. There is no pruning and division is by entropy.

Table 2
WEKA variable selection methods, classifier types and search methods.

Evaluation method	Classifier/options	Search method
Attribute selection methods		
WrapperSubSetEval	BayesNet, algorithm K2—P 1	Greedy
	BayesNet, algorithm K2—P 3	BestFirst
	BayesNet, algorithm HC—P 1	BestFirst
	Naive Bayes, predetermined values	Genetic
	Logistic, predetermined values	BestFirst
	Tree J48, predetermined values	BestFirst
CfsSubSetEval	Predetermined values	Exhaustive
	Predetermined values	Rank
Relevance ranking methods		
ChiSquareAttributeEval	Predetermined values	Ranker
OneRAttributeEval	Predetermined values	Ranker
GainRatioAtrributeEval	Predetermined values	Ranker

- *J48 algorithm* [22]: This advanced version of the C4.5 algorithm implemented in WEKA works with discrete and continuous variables and generates n-ary trees. It uses the gain criterion for division and applies a pruning process.
- *LMT algorithm* [23]: This hybrid induction tree and linear logistic regression algorithm produces binary or n-ary trees and logistic regression models in the leaves. Post-pruning is based on estimating error complexity.

- *Bayesian networks*: Bayesian networks [24] are directed acyclic graphs used for descriptive and predictive purposes. Their node-and-arc network structure provides information on independence/dependence relationships (depicted by arcs) and variables (depicted by nodes). For our research we used K2, hill-climber and TAN implemented in WEKA as the network training algorithms, with different constraints on the number of parents. We also used a particular case of the Bayesian networks, called naive Bayes, with a structure of just two levels and a single parent (the response variable) pointing to all the covariables. The networks were trained by means of a greedy search of the space of possible structures, with the best network chosen on the basis of a specific goodness-of-fit criterion for the selected algorithm.
  - *K2 algorithm* [25]: This algorithm, which uses a greedy search mechanism, starts from the simplest possible network, which, in each successive iteration, is modified by the addition of new parents producing greater benefits according to a pre-established criterion.
  - *Hill-climber algorithm*: This algorithm, which is a discrete version of the gradient descent (ascent) algorithm (see, for example, [26]), implements a local search in each iteration. The algorithm starts with an initial network and determines a nearest-neighbour graph that improves the network by including, eliminating or inverting an arc in the graph. The process is repeated until there is no neighbour that improves the current solution.
  - *TAN algorithm* [27]: This algorithm first constructs the attributes tree structure and then adds the class variable following the naive structure.
- *Support vector machines*: Support vector machines (SVM) [28] implement a linear classification rule that maximises the distance between classes (the solution margin). They do this in a larger dimension space that is the outcome of suitably transforming the input space variables, resulting in a non-linear frontier between the classes in the original space. SVMs are trained by resolving a quadratic programme with linear constraints that has a single solution. The algorithm used in our research was the SMO algorithm, which efficiently solves the quadratic programme for the SVMs [28] by dividing the quadratic optimisation problem into a series of small problems that can be resolved analytically.

3. Results

3.1. Variable selection

The different variable evaluation methods produced fairly similar results. With a view to constructing a single ranking of variables by order of importance, the cross-validation results for the frequency-based selection methods were translated into a ranking based on the most frequently used variables. Finally, the means were calculated for the rankings obtained by each of the methods, resulting in the ranking shown in Table 3.

The variables with higher rankings (i.e., greater weight) were associated with the kind of work done by the worker and with aspects associated with the type of employment. The eight variables with the highest weights included worker age and risk training.

Table 3
Ranking based on the means calculated for the results obtained by the different relevance evaluation methods for the response variable. Ranking was based on the ranking obtained by each method, whereas selection considered the number of times each variable was selected in each cross-validation iteration.

Ranking	Variable
1	TKH—Task duration in hours
2	CCS—Company contractual status
3	TJB—Length of time doing the job
4	JOB—Job type
5	AGE—Worker age
6	HRD—Time of day
7	ART—Accident risk training
8	RKA—Risk assessment
9	MAS—Risk management and supervision
10	PFA—Personal factors
11	DAY—Day of the week
12	TCO—Length of time in the company
13	RAW—Job-associated risk awareness
14	ECT—Type of employment contract
15	MTH—Month
16	NAT—Worker nationality
17	JBP—Job-related protective measures

Table 4
Success rates and confusion matrices for the models for the variables producing the best results.

Model	Variables (*n*)	Success (%)	Confusion Matrix
BayesNet—K2—1 parent	7 & 10	85.48	[108,143]
BayesNet—K2—3 parents	7	88.71	[126,143]
BayesNet—K2—8 parents	7	88.71	[126,143]
BayesNet—HC—1 parent	7	85.48	[108,143]
BayesNet—HC—3 parents	7	79.03	[810,341]
BayesNet—TAN	10	83.87	[117,341]
Naive Bayes	7 & 10	83.87	[99,143]
Simple Naive Bayes	7 & 10	83.87	[99,143]
SVM—SMO	7	87.10	[117,143]
Logistic regression	7 & 17	72.58	[810,737]
Tree—ID3	7	77.42	[125,436]
Tree—J48	7, 10 & 17	87.10	[117,143]
Tree—LMT	7	82.26	[108,341]
Rule—PART	7	85.48	[117,242]
Rule—OneR	7, 10 & 17	75.81	[414,143]

It is noteworthy that variables that were associated with managing risk overall and that did not depend directly on the worker—namely risk assessment and risk management and supervision—were ranked 8 and 9.

Also of interest is the fact that job-related protective measures (whether individual or collective), which also correspond to the set of variables that define risk management, were ranked last.

3.2. Model predictive capacities

The predictive capacity of the models, as mentioned earlier, was evaluated using three tests: by selecting the 7 and 10 best covariables in the ranking given in Table 2 and by using all the covariables.

Predictive capacity was assessed on the basis of classification success (or error) in 10-fold cross-validation. The approach used was to exclude six observations from training and evaluate model training success (or error) for 56 observations. The procedure was performed for 10 randomly selected groups and the mean success rate was calculated. This approach enabled model predictive capacity to be assessed for new cases.

Table 4 shows the best results obtained for the models, indicating the number of covariables used, the mean cross-validation success rate and the confusion matrix. For models

producing similar success rates for different numbers of variables, confusion matrices are shown only for the lowest number of variables (seven variables).

The confusion matrices, shown on a single line for reasons of space, are composed of two rows with the comma indicating the end/beginning of a row. For a matrix [*a b*, *c d*], *a* is the number of correctly classified accidents, *b* is the numbers of accidents wrongly classified as incidents, *c* is the numbers of incidents wrongly classified as accidents and *d* is the number of correctly classified incidents.

In general, the different techniques functioned best with seven variables, although this does not mean that all the techniques used the same variables in the resulting models.

This would imply that training with many variables may ultimately be unproductive, confirming the importance of a variable pre-selection process. This can be observed particularly with the non-naive Bayesian networks, with the exceptions being the classification trees and the naive Bayesian networks. Classification trees have their own training algorithm and so successfully implement their own selection methods; the naive Bayesian networks do not select variables but directly estimate probabilities for all the covariables in accordance with the different values of the response variable.

As for success rates, the best (and very similar) results were obtained by the Bayesian networks with the K2 algorithm (88.71%), the J48 classification tree, the SVM (both 87.10%) and the PART rules (85.48%). PART produced the same structure as the J48 tree, in fact, with the small difference between success percentages resulting from cross-validation randomness.

The poorest performing predictive models were logistic regression (72.58%)—because of the excessive simplicity of its linear model—and the OneR algorithm for rule extraction (75.81%)—which produced rules that were based excessively on the particularities of the sample and so tended to overfit.

As for the confusion matrices, it can be observed that, with seven variables, most errors in the models occurred in classifying the accidents. This can be explained by the fact that accidents are less well-represented in the sample than incidents, making it more difficult for the models to characterize the former. When all the variables were used, the model classification errors were more evenly distributed between accidents and incidents, although the success rates fell, indicating that using all the variables led to overfitting.

3.3. Model interpretation

Of all the models evaluated, those with the best explanatory capacity (whether because they establish rules or decision lists or graphically model the relationship between variables) were the decision rules, the classification trees and the Bayesian networks. Below we analyse the results obtained by these models from the perspective of interpreting accident occurrence in terms of the number of variables used in the models.

The simplest result explains accident occurrence in terms of two variables: task duration in hours and company contractual status (ranked 1 and 2, respectively, in Table 2). This result was obtained for seven covariables by the three best models: the PART rule, the Bayesian networks and the classification trees.

As an illustrative example, Fig. 1 shows the structure of the J48 tree, analogous to that produced by the PART rule and interpreted as follows:

- If a task lasts less than 4 h, the outcome is an accident, with a confidence level of 85.71% and coverage of six cases.
- If a task lasts 4–8 h, the outcome is an incident, with a confidence level of 100% and coverage of six cases.
- If a task lasts more than 8 h, the outcome depends on the company's contractual status: an accident if the company is subcontracted (confidence, 100%; coverage, 5) or an incident if the company is the main contractor (confidence, 84.09%; coverage, 37).

The Bayesian networks offered a much more interesting perspective on interpreting accidents than the other models. They agreed with the other models in granting greatest weighting in terms of accident risk to task duration in hours (when less than 4 h) and company contractual status (when subcontracted).

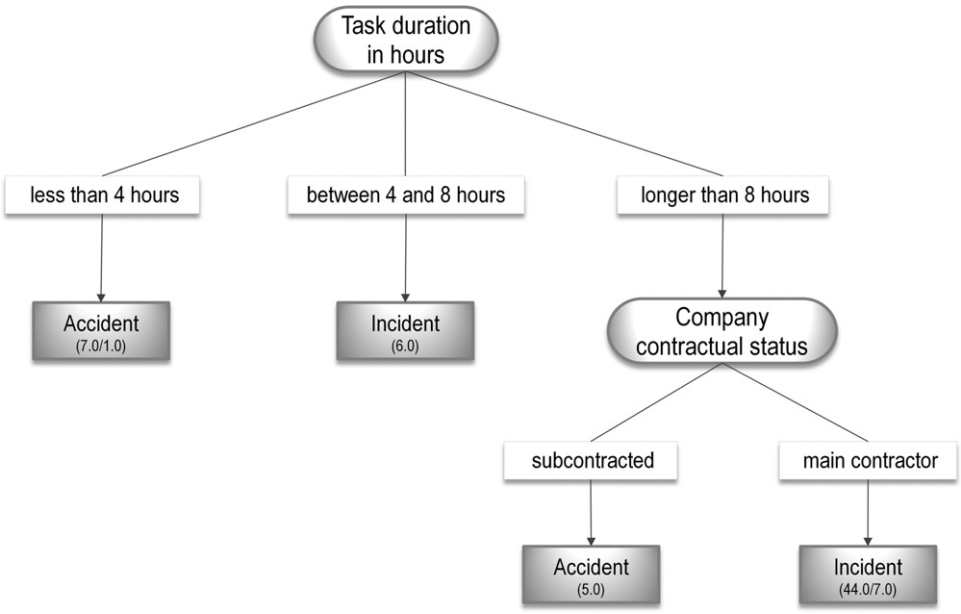

Fig. 1. J48 tree structure. The numbers in brackets indicate the number of observations in each class (accident/incident) and the letter indicates the majority class (A for accident and I for incident). Thus, for example, the first node on the extreme left indicates that if task duration in hours is less than 4, then more accidents (7) are likely compared to incidents (1).

However, as graphic models that are automatically updated when new evidence is included (i.e., they permit an immediate what-if query), the Bayesian networks have an advantage over the other models in terms of the overall interpretation of workplace accidents.

The simplest Bayesian model—and also the one with one of the best prediction success rates—was the greedy K2 Bayesian network based on the first seven causal variables, depicted in Fig. 2(a). In the structure learning phase a restriction of three parents was established as the maximum for each node. Nonetheless, from the data, the algorithm estimated the network shown in the figure, with just a maximum of two parents per node, indicating that the data do not require greater complexity. This figure highlights one of the distinguishing features of this technique: its capacity for identifying possible relationships not only between the covariables and the response but also between the covariables themselves. The network obtained revealed direct relationships between time on the job and the type of job, task duration and company contractual status and the worker's training.

Table 5 depicts the conditional probabilities for the node referring to task duration in hours, showing the conditional probabilities for each state (accident or incident) estimated from the data using the maximum likelihood method. It can be observed that

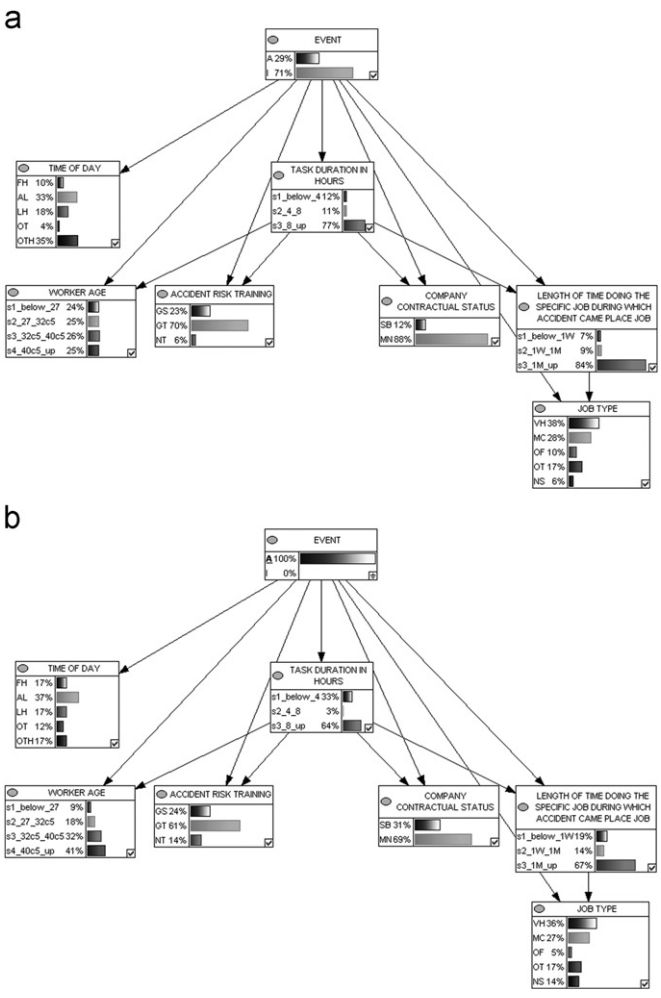

Fig. 2. Greedy K2 Bayesian network built with seven variables and a maximum of three parents. The bars indicate the estimated probabilities for each variable: (a) structure inferred from the data and (b) what-if analysis regarding the possibility of an accident ($p(A)=100\%$).

Table 5
Conditional probabilities for the task duration node. Shown are the probabilities for each state (accident or incident) estimated from the data.

Task duration (h)	Accident	Incident
$d \leq 4$	0.3333	0.0330
$4 < d \leq 8$	0.0256	0.1429
$d > 8$	0.6410	0.8249
	1.0000	1.0000

when an accident occurred, the probabilities that a task lasted more than 8 h or less than 4 h were 0.6410 and 0.3333, respectively; since both are higher than the probability for a task lasting between 4 and 8 h, the conclusion is that an accident is less likely in this intermediate task duration interval.

However, this analysis can be more easily performed by exploiting the other great feature of this technique: the possibility of performing what-if analyses based on the structure inferred from the data. In Fig. 2(b), which shows the results of this analysis regarding the occurrence of accidents, it can be observed that the probability of an accident rose significantly—compare the conditional probabilities in Fig. 2(b) with the unconditional probabilities in Fig. 2(a)—for workers aged over 40 years, workers on the job for less than a week, subcontracted workers and workers performing tasks lasting less than 4 h. These accidents, moreover, were associated with jobs categorized as other or unspecified, and also featured a higher proportion of workers with insufficient risk training. In terms of the time frame, the accidents were associated with the first hours of the day, the hours after lunch and overtime hours.

Like the other models, the networks gave greatest weight to task duration in hours and company contractual status, and it can be observed that this association arose among workers employed by subcontractors. According to the network, this group of workers is composed of employees aged over 40 years who perform activities for 4 h or less, on the job for less than a week and carrying out unspecified jobs. In comparison, workers employed by a main contractor tend to have a slightly higher level of job-specific accident-risk training.

In regard to the same network built with more variables, a number of other interesting observations can be made:

- Certain variables had little or no influence on the accident/incident distribution of the variables, for example, risk assessment, age, type of employment contract and length of time in the company.
- The risk management and supervision variable was associated with the risk assessment and event variables; thus, when risk management and supervision was appropriate, risk was evaluated better, meaning that incidents were more likely than accidents (80% of cases). In contrast, if risk management and supervision was defective, the accident rate increased (51% of cases).
- A close link could be observed between job-related protection and risk assessment. Thus, if risk was assessed, then protection was appropriate in 83% of the cases; otherwise, this rate fell to 67%. Likewise, good job-related protection was associated with a greater proportion of tasks lasting longer than 8 h, workers who had been on the job for more than 1 month and specified jobs (vehicle operators, heavy machinery operators and office workers)—all associated with a higher incident rate. In contrast, poor job-related protection was associated with brief tasks and unspecified jobs—all associated with a higher accident rate according to all the models.

4. Discussion

The results of this research represent an important advance in terms of managing information on workplace accidents, obtained in this case, from surveys conducted soon after the occurrence of incidents/accidents in the construction and mining sectors.

First of all, we established a statistical analysis protocol, based on data-mining techniques, that enables the following: (1) selection of the most relevant information from all the data available, which ultimately enhances model prediction success rates and explanatory capacities; and (2) identification of the most useful data-mining tools in terms of predictive and explanatory capacities.

Secondly, the conclusions extracted from the models capable of explaining the causes of workplace incidents/accidents in the surveyed companies revealed the circumstances with a bearing on accidents, thereby enabling causes to be clearly defined. Nonetheless, such conclusions were not the main aim of the research, which was conceived from the outset as a pilot project aimed at profiling a methodology for analysing workplace accidents that can be safely applied to a subsequent large-scale project. These conclusions, therefore, are only valid for the specific companies and period studied and should not be extrapolated further.

In the variable pre-selection phase, all the models applied in our research coincided in pointing to the same few variables (of a longer list of possible variables) with most weight in terms of predicting an accident/incident. This approach screens out non-relevant information and so ensures better training and prediction by the models. Furthermore, it also means that further questionnaires can be prepared for the same employees that will be easier to complete and therefore more effective, whether in terms of recording information or including other exploratory factors that better delimit the causes of accidents/incidents.

In terms of the information provided about the accidents/incidents that occurred in the surveyed companies, the most useful of the models studied were the rule-based (classification trees and decision rules) and Bayesian models, as they were the most successful in terms of prediction. This conclusion generally coincides with that obtained in previous research [10], which demonstrated that these techniques offered the best predictive capacities in regard to workplace falls studied using a database created from accident reports. It can only be concluded that the structure of these techniques is very adaptable to the data generally available in this area of study.

Decision rules and classification trees provided identical information on workplace accidents when relatively few variables (the most relevant ones) are included in the models. They are usefully applied when the aim is to identify the circumstances typically associated with a greater probability of an event; an example from this particular research is the association between accidents and both task duration and company contractual status. The Bayesian networks—which offer the same quality of information as the rules and trees in regard to the most relevant variables with a bearing on accidents/incidents—have an additional advantage, which is that their what-if analytical capacity allows data to be explored in greater depth, enabling different scenarios for workplace accidents/incidents to be depicted.

From the risk management perspective, our results point to the typical scenario in which workplace accidents occurred in the companies that were surveyed for our research: tasks of a short duration, jobs that had not been specified, workers employed by subcontractors and workers who had not been long on the job. Once an associated scenario like the one described has been characterized, graphic models like the Bayesian networks enable an investigation into the possible causes of an accident; thus, for example, referring to the particular scenario described above, we were able to rule out an association between a higher rate of accidents and poorer risk training, as the Bayesian networks indicated that risk training was similar and sometimes even better for workers employed by subcontractors.

Another interesting fact is that, in the variable pre-selection phase for our sample, the variables associated with company risk management were ranked in the intermediate (two variables) and lowest (one variable) positions. The Bayesian networks thus revealed no direct association, for our sample, between accidents and variables such as risk assessment, job-related protective measures and risk management and supervision. Note that we were not researching accidents as such; we were analysing possible causes that might affect the level of accidents in a specific group of workers. Leaving aside the particular circumstances of subcontractors in our sample, what remains clear is the need to investigate the circumstances associated with risk management and supervision that would enable a better causality model for a company or sector to be built.

These results were obtained from a database of 62 cases. This database size, which was determined by the kind of unforeseeable factors that inevitably affect a pilot project, may be considered insufficient to enable conclusions to be drawn. Nonetheless, the ideal sample size cannot be pre-established, as it depends on the variability in the data (variance) and in the structure of relationships between variables. If variability was zero, a single record for each type of event (incident or accident) would be sufficient to infer cause. In contrast, if there were no causal relationship between the variables and the relationship between variables were unpredictable, even an unlimited quantity of data would be useless.

For this reason, in large-scale projects, this variability should be estimated through pilot sampling, as done in our research, which was conceived as a preliminary project aimed at designing a methodology that can be applied to different scenarios in order to estimate the minimum sample size (among other organizational, statistical and computational aims) that would depict the underlying causal structure.

Table 2 shows that the most successful techniques had success rates above 85%. It should be borne in mind that these rates were obtained by means of cross-validation; therefore, their predictive capacity was evaluated for data not processed by the different models. These results speak for themselves regarding the coverage provided by the size of the sample in this pilot project. If the sample size had been insufficient, predictions regarding new data (that is, data not used in building the models) would only be slightly better than random predictions. It is difficult to conceive that success rates of over 90% could be achieved for problems of this nature, with so many circumstantial variables conditioning accident type.

On the other hand, the lower percentages for a linear model like logistic regression suggest non-linear relationships between the variables, justifying the use of machine learning techniques for this kind of problem, due to their capacity for reproducing non-linear structures that are not known a priori (the non-parametric philosophy).

5. Conclusions

The results of this research have enabled us, using different data-mining techniques, to define an efficient protocol for handling and analysing workplace accident data that (1) identifies the most immediately relevant variables and consequently improves prediction success rates and explanatory capacities, and (2) permits conclusions to be drawn regarding the causes of accidents/incidents.

The protocol is implemented in two phases. Firstly, the most important variables are pre-selected using different approaches. In our case, all the approaches used produced similar results in terms of ranking the most relevant variables. Secondly, relevant information regarding accident/incident causes is extracted using different prediction models. Decision rules, classification trees and Bayesian networks constructed with the K2 algorithm best explained accidents/incidents. Furthermore, the Bayesian/K2 networks had the added advantage of allowing what-if analyses, thereby enabling data to be explored in greater depth and different workplace accident/incident scenarios to be depicted.

The results of this research represent an important advance in terms of managing information on workplace accidents. The satisfactory results of the decision rules, classification trees and Bayesian networks constructed with the K2 algorithm indicate these to be reliable tools for studies of workplace accidents and their causes. The quality of the results is such that we will be able to plan and design a more ambitious, larger-scale study aimed at gaining a deeper understanding of the causes of accidents in a range of industrial sectors, but, in particular, in the mining and construction sectors. The methodologies, moreover, are sufficiently consistent to be used in periodic studies aimed at comparing developments over time.

Acknowledgement

J. M. Matías's part of this research was funded by the Spanish Ministry of Science and Innovation under Project MTM2008-03010. Ailish M.J. Maher assisted with the English in a version of this manuscript.

References

[1] Spanish National Institute for Occupational Health and Safety and the Spanish Ministry of Labour and Immigration. Informes estadísticos sobre siniestralidad laboral 2007–2008–2009 [Statistics on workplace accidents 2007–2008–2009]. Available from: 〈http://www.oect.es/portal/site/Observatorio/〉 [in Spanish].
[2] Ley 32/2006, de 18 de octubre, reguladora de la subcontratación en el Sector de la Construcción [Law 32/2006 regulating subcontracting in the construction sector]. Official State Bulletin (BOE), 19 October 2006; 250. p. 36317–23 [in Spanish].
[3] Real Decreto 1109/2007, de 24 de agosto, por el que se desarrolla la Ley 32/2006, de 18 de octubre, reguladora de la subcontratación en el Sector de la Construcción [Royal Decree 1109/2007 deploying Law 32/2006 regulating subcontracting in the construction sector]. Official State Bulletin (BOE), 25 August 2007; 204. p. 35747–35764 [in Spanish].
[4] Groves WA, Kecojevic VJ, Komlejenovic D. Analysis of fatalities and injuries involving mining equipment. Journal of Safety Research 2007;38:461–70.
[5] Ural S, Demirkol S. Evaluation of occupational safety and health in surface mines. Safety Science 2008;46:1016–24.
[6] Karra VK. Analysis of non-fatal and fatal injury rates for mine operator and contractor employees and the influence of work location. Journal of Safety Research 2005;36:413–21.
[7] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann; 2005.
[8] Han J, Kamber M. Data mining: concepts and techniques. Morgan Kaufmann; 2006.
[9] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. data mining, inference and prediction. Springer; 2001.
[10] Matías JM, Rivas T, Martín JE, Taboada J. A machine learning methodology for the analysis of workplace accidents. International Journal of Computer Mathematics 2008;85:559–78.
[11] Martín JE, Rivas T, Matías JM, Taboada J, Argüelles AA. Bayesian network analysis of workplace accidents caused by falls from a height. Safety Science 2009;47:206–14.
[12] Liao CW, Perng YH. Data mining for occupational injuries in the Taiwan construction industry. Safety Science 2008;46:1091–102.
[13] Liao CW, Perng YH, Chiang TL. Discovery of unapparent association rules based on extracted probability. Decision Support Systems 2009;47:354–63.
[14] Bird FE, Germain GL. Practical loss control leadership, revised ed. International Loss Control Institute; 1990. p. 446.
[15] Ley 31/1995, de 8 de noviembre, de prevención de riesgos laborales [Law 31/1995 on the prevention of workplace risk]. Official State Bulletin 269, 10 November 1995. Transposition of Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work. Official Journal L183, 29/06/1989. p. 0001–0008.
[16] Decreto Real. 39/1997, de 17 de enero, por el que se aprueba el reglamento de los servicios de prevención [Royal Decree 39/1997 approving regulation of risk prevention services]. Official State Bulletin (BOE) 1997;31(27):3031–45. [in Spanish].
[17] Seber GAF. Multivariate Observations. New York: John Wiley and Sons; 1984.
[18] Holte RC. Very simple classification rules perform well on most commonly used datasets. Maching Learning 1993;11:63–91.
[19] Frank E, Written IH. Generating accurate rule sets without global optimization. In: Fifteenth international conference on machine learning 1998. p. 144–51.
[20] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Wadsworth; 1984.
[21] Quinlan JR. Induction of decision trees. Machine Learning 1986;1(1):81–106.
[22] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers; 1993.
[23] Landwehr N, Hall M, Frank E. Logistic model trees. Machine Learning 2005;95(1–2):161–205.
[24] Jensen FV. Bayesian networks and decision graphs. Springer; 2001.
[25] Cooper GF, Herskovits EA. Bayesian method for the induction of probabilistic networks from data. Machine Learning 1992;9(4):309–47.
[26] Nocedal J, Wright SJ. Numerical optimization. Springer; 1999.
[27] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning 1997;29(2–3):131–63.
[28] Schölkopf B, Smola AJ. Learning with kernels. MIT Press; 2002.