MineríaSostenible

INTERNATIONAL CONFERENCE
SustainableMining

CONFERENCIA INTERNACIONAL
MinaríaSostible

09

LIBRO DE ACTAS

www.cims2009.com

Aplicación de técnicas de minería de datos para el análisis y predicción de accidentes en el sector minero

María Paz Paz Freire[1], Teresa Rivas[1], José Enrique Martín Suárez[2], Julio Francisco García Menéndez[2], José María Matías Fernández[3], Javier Taboada Castro[1]

Resumen

En este trabajo se modeliza la ocurrencia de accidentes/incidentes en dos empresas de los sectores de la minería y obra civil, con objeto de identificar los factores relevantes en dicha ocurrencia y obtener modelos predictivos al respecto. En el campo de investigación de riesgos laborales, el uso de metodologías estadísticas se limita a análisis descriptivos que no permiten un correcto análisis causa-efecto y la construcción de modelos predictivos que ayuden a anticipar la ocurrencia de este tipo de sucesos. El presente trabajo ofrece una aproximación basada en técnicas de minería de datos con el objeto de evitar estas deficiencias. Sobre una base de datos elaborada a partir de encuestas realizadas con posterioridad a incidentes y accidentes, se han aplicado: reglas de decisión, redes bayesianas, máquinas de soporte vectorial, árboles de clasificación y regresión logística. A través de los resultados obtenidos se ponen de manifiesto las ventajas de las reglas de decisión, árboles y redes bayesianas a la hora de predecir los accidentes/incidentes y de identificar los factores determinantes de los mismos.

Palabras clave: accidentes, árboles de clasificación, minería de datos, redes bayesianas, riesgos laborales.

Application of data mining techniques to analyse and predict work accidents in mining sector

Abstract

In the present work the occurrence of incident/accidents in two companies belonging to mining and construction sectors is modelled in order to identify the most important causes of this occurrence and to obtain predictive models on the matter. Current research on workplace risk is mainly performed using conventional descriptive statistics which limit the identification of the cause-effect relationships and also the construction of predictive models that help to anticipate the occurrence of this type of events. In order to avoid these methodological deficiencies, several data mining techniques (rule extraction algorithms, bayesian networks, support vector machines and classification trees), have been applied to accident/incident data obtained from no-delay interviews after incidents/accidents in mining and construction sectors. The results obtained are compared with classical statistical techniques like, for example, logistic regression.

1 Dpto. Ingeniería de los Recursos Naturales y Medio Ambiente. E.T.S.I. Minas. University of Vigo. Campus Lagoas 36310 Vigo-Spain. mpaz.minas@gmail.com, trivas@uvigo.es, jtaboada@uvigo.es

2 CIPP International. S.L. Parque Tecnológico de Asturias. Parcela 43. Of. 11. 33428 Llanera-Spain. jmartin@cippinternacional.com; jgarcia@cippinternacional.com

3 Dpto. Estadística e Investigación Operativa. E.T.S.I. Minas. University of Vigo. Campus Lagoas 36310 Vigo-Spain. jmmatias@uvigo.es

The result of this comparison reveals the advantages of rule extraction algorithms and bayesian networks in identifying the factors involved on the accidents. In terms on predictive capacity the best results are obtained by Bayesian networks and support vector machines.

Key words: accidents, classification trees, data mining, bayesian networks, labor risks.

Introducción

En España, en el año 2007 se produjeron un total de 1.022.067 accidentes laborales con baja en jornada de trabajo, siguiendo la trayectoria de aumento de accidentes que se viene repitiendo desde antes de la aprobación de la Ley 31/1995. No obstante, desde el año 2000 se ha experimentado una tendencia descendente de los índices de incidencia, tanto para el total de accidentes como para los accidentes mortales (Instituto Nacional de Seguridad e Higiene en el Trabajo, Ministerio de Trabajo e Inmigración). Dicho descenso puede ser explicado por la implicación cada vez mayor de las empresas en mejorar la seguridad de sus trabajadores mediante las acciones preventivas, ya sea por obligación legal o por iniciativa propia.

Para lograr estas mejoras, los empresarios deben obtener información rigurosa sobre las causas que están detrás de la ocurrencia de los accidentes y así poder llevar a cabo acciones correctivas eficaces.

En prevención de riesgos laborales, se vienen utilizando metodologías estadísticas convencionales para el manejo de los datos, fundamentalmente estadística descriptiva, de tal forma que los resultados obtenidos en estos estudios no van más allá de resúmenes históricos, porcentajes o índices (Groves et al., 2007; Ural y Demirkol, 2008) o métodos de regresión (Karra, 2005).

En esta situación, la interpretación global del fenómeno de la siniestralidad con el conjunto de todas las posibles causas se ve muy limitada y es difícil establecer una hipótesis de trabajo orientada a predecir la ocurrencia de los sucesos y por tanto a reducir la accidentalidad. Esto obliga a buscar otras herramientas que aprovechen eficazmente la información que ofrecen los datos para extraer relaciones de dependencia entre todas las variables consideradas y por tanto para obtener información rigurosa y global sobre la casuística de los accidentes.

La minería de datos se presenta como una importante disciplina que se viene utilizando en campos como la medicina, la ingeniería, el sector financiero, etc., obteniéndose buenos resultados en todos ellos (p. ej. Witten y Frank, 2005; Han y Kamber, 2006; Hastie y Tibshirani, 2001). En el campo de la gestión de la seguridad en el trabajo existen referencias sobre la utilidad de algunas de estas técnicas, tanto en lo que se refiere a la capacidad predictiva (Matías et al., 2008) como a la capacidad interpretativa del fenómeno (Matías et al., 2009). En estos estudios, y a partir de datos procedentes de partes de accidentes y de encuestas a los trabajadores y empresarios, quedan reflejadas las principales ventajas de la minería de datos frente a la estadística convencional, como son su función predictiva y la posibilidad de identificar interacciones entre las variables que intervienen en la ocurrencia de los accidentes.

En este trabajo se pretende identificar qué metodologías de minería de datos, de entre un gran grupo preseleccionado, extraen conocimiento útil sobre el fenómeno de la

siniestralidad a partir de una base de datos extraída de encuestas realizadas con posterioridad a la ocurrencia de incidentes/accidentes en empresas del sector minero y movimiento de tierras para obra civil. En España, ambos sectores, minería y construcción de obra civil, han sido tradicionalmente muy castigados por los accidentes y de hecho son los que encabezan el mayor porcentaje de accidentes con baja laboral (año 2006) según el Ministerio de Trabajo y Asuntos Sociales.

Hemos evaluado como herramientas de análisis las redes bayesianas, reglas de clasificación, árboles de clasificación, regresión logística y máquinas de soporte vectorial. Los resultados obtenidos permitirán concentrar los trabajos de prevención en los puntos más conflictivos con la finalidad de reducir la siniestralidad laboral de una manera eficaz.

Material y métodos

Fuentes de información y descripción de datos

La información utilizada en el presente estudio se ha obtenido a partir de una encuesta llevada a cabo entre los trabajadores de dos empresas pertenecientes al mismo grupo empresarial, una dedicada a la minería y otra a obra civil, que desarrollan actividades en España, en las comunidades de Aragón, Asturias y Valencia. Durante el período de estudio, septiembre de 2007 a marzo de 2008, se obtuvieron un total de 62 encuestas.

En la empresa dedicada a la minería fueron encuestados trabajadores procedentes de cuatro explotaciones activas: una explotación a cielo abierto de carbón para producción de energía (que emplea a 48 trabajadores), dos explotaciones a cielo abierto de arcillas destinadas al sector cerámico (que emplean a 72 trabajadores en total) y una cantera de cuarcitas, para uso de relleno en obra pública, y obtención de árido y escollera, que emplea 13 trabajadores. La empresa de obra pública realiza trabajos de movimiento de tierras para la construcción de carreteras, autovías, ferrocarril y AVE. Cuando este tipo de obras se encuentran en producción, la plantilla está constituida por unos 50 trabajadores.

Los diferentes campos cubiertos en la encuesta aportan información sobre las causas básicas e inmediatas (Bird y Germain, 1990) que llevan a la ocurrencia del suceso. Cada uno de los campos de la encuesta constituye una variable y en total se ha obtenido información sobre 17 variables que pueden separarse en tres tipos: relativas al suceso, al trabajador y a la obra y a la empresa. A continuación se describe cada una de las variables.

Relativas al suceso:

- Hora del día (H), con las siguientes posibilidades: a primera hora del día (PH), después de comer (DC), dos últimas horas de la jornada laboral (DU), horas extras (HE) y otras horas del día (Ot).

- Día de la semana (D), expresado mediante seis posibilidades: lunes (L), martes (M), miércoles (Mx), jueves (J), viernes (V) y sábado (S).

- Mes (M): Los valores que toma en la base de datos son: septiembre (S), octubre (O), noviembre (N), diciembre (D), enero (E), febrero (F) y marzo (Mz).

Relativos al trabajador:

- Edad del trabajador (Ed_d): Para trabajar con esta variable se ha discretizado creando cuatro rangos de edad: menos de 27 años de edad (s1_below_27), entre

27 y 35.5 años de edad (s2_27_35c5), entre 35.5 y 40.5 años de edad (s3_35c5_40c5) y más de 40.5 años de edad (s4_40c5_up).

- Nacionalidad (Na): Se han agrupado las nacionalidades en 4 grupos: trabajadores de nacionalidad española (Nac), trabajadores de Europa del este (Armenia, Bulgaria, Polonia y Rumanía) (EE), trabajadores de países africanos (Libia y Marruecos) (Afr), y trabajadores de Latinoamérica (Chile, Colombia y Ecuador) (LA).

- Puesto de trabajo (PT): Se han dividido los puestos de trabajo en 5 grupos: puestos que implica el manejo de vehículos (PV); puestos en los que se maneja máquinas de gran tamaño (PGM); puestos de oficina y sin uso de máquinas (PSM); otros puestos de trabajo en el que se incluirían electricistas, mecánicos, ayudantes,... (OP) y un último grupo para aquellos casos en los que no se reflejó en la encuesta el puesto de trabajo del operario (SD).

- Antigüedad en la empresa (An_d): Esta variable se ha discretizado en 5 grupos: menos de medio año en la empresa (s1_below_0c5), más de medio año y menos de un año en la empresa (s2_0c5_1), más de un año y menos de año y medio (s3_1_1c5), entre año y medio y dos años en la empresa (s4_1c5_2) y más de dos años de experiencia en la empresa (s5_2_up).

- Tiempo del trabajador en la obra (TO): Con esta variable se intenta reflejar cuanto tiempo el empleado llevaba trabajando en la operación en la que se ha producido el suceso. Los posibles estados son: menos de una semana (s1_below_1S), entre una semana y un mes (s2_S_M) o más de un mes (s3_M_up).

- Formación (F): Se han definido tres estados posibles: formación genérica y específica de su puesto de trabajo (de carácter aplicado y práctico en el puesto de trabajo, según la ley 31/95) (G_E), formación genérica (formación teórica sobre riesgos generales) (G) y carencia de formación sobre riesgos laborales (N).

- Tipo de contrato del trabajador (TC): Se recoge con esta variable el tipo de contrato del operario: obra y servicio (OS), temporal (T) e indefinido (In).

- Reconocimiento del peligro (RP): Se trata de una variable subjetiva, que define si el operario reconoció que se encontraba ante una situación de peligro antes del accidente/incidente, siendo los posibles estados: el trabajador se percató de la situación de peligro (Si), y el trabajador no se dio cuenta del peligro (No).

- Factores personales (FP): Ésta es una variable subjetiva, pero se incluye en el estudio porque es indudable la influencia que tiene en la ocurrencia de los sucesos factores como el exceso de celo, el estrés, la monotonía, etc. Se definen las dos situaciones siguientes: sí hubo factores personales que pudieron contribuir al acontecimiento del suceso (Si) y no hubo ningún factor personal que pudiera haber influido (No).

Relativos a la obra y a la gestión en seguridad de la empresa:

- Duración de la operación en horas (HO_d): Esta variable va a influir directamente sobre el uso de las medidas de protección. Se han definido tres estados: operaciones de menos de 4 horas de duración (s1_below_4), operaciones de duración entre 4 y 8 horas (s2_4_8) y ope-

raciones de más de 8 horas de duración (s3_8_up).

- Régimen de contrato (R): Con este atributo se define si la empresa estaba trabajando en régimen de subcontrata (Sb) o bien, como contrata principal (CP).

- Dirección y/o supervisión (DS): Se trata de una variable que hace referencia a si existen recursos humanos preventivos en la empresa. Este parámetro se define a través de tres posibles escenarios:

  – Existía una supervisión y/o control de las condiciones de trabajo en el lugar del suceso (S/C).

  – Existían recursos preventivos en la empresa aunque habitualmente no se llevaban a cabo acciones de control y/o supervisión de los trabajos que dieron lugar al suceso (R).

  – No existían recursos preventivos en la empresa que controlaran y/o supervisaran las condiciones de trabajo que dieron lugar al suceso (SR).

- Evaluación del riesgo (ER): La variable ER refleja si se ha realizado (Si) o no (No) la evaluación de riesgos según se describe en el Real Decreto 39/97.

- Condiciones de trabajo (CT): Hace referencia a las condiciones de trabajo antes del suceso relativas a la dotación de recursos (protección individual y colectiva). Las posibles situaciones son: existían las protecciones (Si) o no (No).

Además de las variables causales anteriores, también se incluye en los análisis de los modelos la variable respuesta denominada Suceso (S) que puede tomar los valores Accidente (A) o Incidente (I).

Metodología de trabajo

Para la utilización de las diferentes técnicas de minería de datos se utilizaron las siguientes herramientas. Para la selección de variables y determinación de su relevancia, así como para la estimación de la mayor parte de las técnicas de minería de datos se utilizó el software de libre distribución WEKA (Waikato Environmental for Knowledge Analysis) desarrollado por la Universidad de Waikato (Witten y Frank, 2005). Además, se utilizó el software Genie (Graphical Network Interface) desarrollado por la Universidad de Pittsburgh para la estimación de las redes bayesianas.

El trabajo se ha planteado en dos fases. Una primera fase se ha centrado en la aplicación de distintos métodos para seleccionar las variables más relevantes. La segunda fase ha consistido en la aplicación de diferentes técnicas de minería de datos para, con un número variable de atributos relevantes, conocer su capacidad predictiva del fenómeno accidente/incidente y discutir su capacidad interpretativa.

Selección de variables

Se ha llevado a cabo una primera fase dirigida a identificar, entre las diecisiete variables consideradas, las más relevantes a la hora de explicar la variable respuesta.

Para la selección de las variables relevantes se han realizado 3 pruebas mediante validación cruzada con 10, 5 y 3 grupos respectivamente. La herramienta WEKA cuenta con un gran número de métodos de evaluación para el estudio de la relevancia de las variables de la base de datos en relación con una variable respuesta.

A través de la función selección de atributos (select attributes) de WEKA se han utilizado 31 modelos distintos agrupados según dos grandes métodos basados en:

- La frecuencia con que las variables eran seleccionadas en cada ronda de entrenamiento de la validación cruzada. En este grupo se utilizaron dos técnicas, *CfsSubSetEval y WrapperSubSetEval*. La primera selecciona las variables mediante su grado de correlación con la variable respuesta, y la segunda a partir de la mejora que la incorporación de cada variable produce en la capacidad predictiva de diferentes modelos de minería de datos. Se utilizaron 19 modelos con la primera técnica y 6 con la segunda técnica.

- Algún indicador de mérito de cada una de las variables, produciendo un ranking de las mismas según su relevancia. Se utilizaron 6 modelos de selección que ordenan por relevancia.

Las opciones de evaluación así como los clasificadores y métodos de búsqueda empleados en algunos de los modelos se detallan en la Tabla 1.

Método de evaluación	Clasificador / Opciones	Método de búsqueda
Métodos seleccionadores de atributos		
WrapperSubSetEval	BayesNet, algoritmo K2 – 1 P	Greedy
	BayesNet, algoritmo K2 – 3 P	BestFirst
	BayesNet, algoritmo HillClimber – 3 P	BestFirst
	NaiveBayes, valores predeterminados	GeneticSearch
	Logistic, valores predeterminados	BestFirst
	Tree J48, valores predeterminados	BestFirst
CfsSubSetEval	Valores predeterminados	ExhaustiveSearch
	Valores predeterminados	RankSearch
Métodos que ordenan por relevancia		
ChiSquareAttributeEval	Valores predeterminados	Ranker
OneRAttributeEval	Valores predeterminados	Ranker
GainRatioAtrributeEval	Valores predeterminados	Ranker

Tabla 1: Algunos de los métodos de selección de variables de WEKA utilizados, junto con el tipo de clasificador y método de búsqueda de cada uno de ellos.

Aplicación de los modelos y análisis de su capacidad predictiva e interpretativa.

Una vez ordenadas las variables según su peso-relevancia en la variable predicción Suceso, se procede a la aplicación de diecisiete distintos modelos o técnicas de minería de datos obteniéndose de cada uno de ellos el porcentaje de casos correctamente clasificados y su matriz de confusión.

Con el objetivo de comprobar la influencia del número de variables introducidas en la capacidad predictiva y explicativa del fenómeno por parte de los modelos, se han estructurado éstos en tres pruebas distintas: introduciendo las siete variables causales más relevantes, obtenidas en la primera fase del estudio, y la variable predicción, las primeras diez variables causales y la variable de predicción y finalmente introduciendo todas las variables.

Los modelos aplicados son reglas de clasificación, árboles de clasificación, redes bayesianas, máquinas de soporte vectorial y regresión logística. A continuación se describen brevemente las características de las diferentes técnicas utilizadas.

Técnicas utilizadas

Las técnicas de minería de datos utilizadas para la modelización de la influencia de las variables explicativas en la variable respuesta (Suceso) fueron las siguientes (además de las referencias específicas, una descripción más detallada de todas puede verse por ejemplo en Witten y Frank, 2005; Han y Kamber, 2006, y de la segunda y las dos últimas en Hastie et al., 2001):

- **Reglas de clasificación**. Son técnicas orientadas a producir reglas del tipo "*si A entonces B*", es decir, A=>B, donde, el antecedente es una combinación de valores de las variables explicativas y el consecuente, es un valor de la variable respuesta, en este caso accidente o incidente. La bondad de las reglas obtenidas a partir de los datos se evalúa según los siguientes índices:

  – Cobertura o soporte: número de instancias que la regla predice correctamente, es decir, la cobertura de A=>B es el número de instancias de los datos que verifican A y B a la vez. Puede escribirse *Cobertura(A=>B) = P(A∩B)*, donde la probabilidad se entiende referida a la muestra de datos.

  – Confianza o precisión: porcentaje de instancias que la regla predice correctamente cuando se puede aplicar, es decir, la confianza de A=>B es el número de instancias de los datos que satisfaciendo B, satisfacen también A. Este número coincide con la cobertura dividida por el número de instancias que satisfacen A. Puede escribirse *Confianza(A=>B) = P(A|B) = P(A∩B)/ P(A)*.

En este trabajo se han utilizado los algoritmos RuleOneR y RulePART de WEKA.

- **Árboles de clasificación**. Los árboles de clasificación son técnicas estadísticas de clasificación que pueden ser representados gráficamente mediante diagramas de árbol.

  A pesar de sus diferentes variantes, en general, el entrenamiento de un árbol consiste en dividir progresivamente los datos en grupos utilizando alguna condición sobre

una de las variables explicativas de tal manera que cada grupo es lo más homogéneo posible en términos de la variable respuesta. A su vez, cada grupo obtenido en la etapa anterior vuelve a dividirse utilizando una nueva condición basada en una variable explicativa buscando incrementar la homogeneidad de los grupos resultantes. Y así sucesivamente, hasta que se satisface algún criterio de parada. Los árboles de clasificación aplicados en este trabajo son: TreeID3, TreeJ48 y LMT implementados en WEKA.

- **Redes bayesianas** (Jensen, 2001). Las redes bayesianas son gráficos acíclicos dirigidos con uso descriptivo y predictivo. Mediante su estructura en red de nodos y arcos, dan información sobre las relaciones de independencia/dependencia (arcos) entre las variables (nodos). En este trabajo se ha utilizado como algoritmo de aprendizaje de la red el GreedyK2, el NetHillClimber y el NetTAN implementados en WEKA, con diferentes restricciones al número de padres. También se utilizó el Naive Bayes, caso particular de estas redes, cuya estructura posee sólo dos niveles y un único padre, la variable respuesta, que apunta a todas las covariables.

  El entrenamiento de estas redes se realiza mediante búsqueda *greedy* del espacio de las estructuras posibles, seleccionándose la mejor en base a algún criterio de bondad específico del algoritmo elegido.

- **Máquinas de soporte vectorial** (SVM) (Scholkopf y Smola, 2002). Las máquinas de soporte vectorial implementan una regla de clasificación lineal que maximiza la distancia entre las clases (maximizar el margen de la solución) en un espacio de dimensión mayor que es el resultado de transformar adecuadamente las variables del espacio de entrada. Como consecuencia, en el espacio original se obtiene una frontera entre las clases de carácter no lineal.

  El entrenamiento de las SVM se realiza resolviendo un programa cuadrático con restricciones lineales, que posee solución única. Para ello, el algoritmo utilizado en este trabajo es el denominado SMO.

- **Regresión logística** (Seber, 1984): La regresión logística es una técnica de regresión lineal generalizada que en lugar de modelizar directamente la variable respuesta en términos de las covariables, modeliza el logaritmo del ratio (*odds ratio)* entre la probabilidad de la clase de interés (accidente) y la probabilidad de la otra clase (incidente). La estimación de la regresión logística se realiza mediante el método de la máxima verosimilitud.

Resultados

Selección de variables

Los distintos métodos de evaluación de variables produjeron resultados bastante similares, sin embargo, con objeto de construir una única ordenación de las variables por orden de importancia, los resultados de los métodos basados en frecuencia de selección en los procesos de validación cruzada, se tradujeron a orden de importancia considerando más importantes las variables más frecuentemente utilizadas. Finalmente, con objeto de obtener un único ranking se calculó la media de las ordenaciones de cada uno de los métodos, obteniéndose el orden de precedencia mostrado en la Tabla 2.

Número de orden	Variable	
1	HO_d	Duración de la operación en horas
2	R	Régimen de contrato
3	TO	Tiempo del empleado en la obra
4	PT	Puesto de trabajo
5	Ed_d	Edad del trabajador
6	H	Hora del día
7	F	Formación
8	ER	Evaluación de riesgos
9	DS	Dirección y supervisión
10	FP	Factores personales
11	D	Día de la semana
12	An_d	Antigüedad en la empresa
13	RP	Reconocimiento del peligro
14	TC	Tipo de contrato
15	M	Mes
16	Na	Nacionalidad
17	CT	Condiciones de trabajo

Tabla 2: Orden de las variables obtenido como media de los resultados de las distintas técnicas de evaluación de la relevancia en la variable respuesta. Para los métodos de ordenación se utilizaron los órdenes obtenidos por cada una. Para los métodos de selección se consideró el número de veces que cada variable era seleccionada en cada iteración de la validación cruzada.

Al respecto de dicha tabla, se destaca lo siguiente:

- Se observa que las variables de mayor peso tienen relación con el tipo de trabajo que realiza el trabajador o con aspectos directamente relacionados con el empleado. Así, entre las ocho variables de mayor peso se encuentra la edad del trabajador y la formación.

- Las variables que no dependen directamente del trabajador y que están relacionadas con la gestión de la seguridad, como es la evaluación de riesgos y la dirección y supervisión quedan en un octavo y noveno lugar.

- Resulta también llamativo que la variable condiciones de trabajo (relativa a la existencia de dispositivos de seguridad individual o colectiva), agrupada dentro del conjunto de variables que definen la gestión de seguridad, haya quedado en último lugar.

Capacidad predictiva de los modelos

La evaluación de la capacidad predictiva de los modelos se ha realizado en tres contextos diferentes: seleccionando las siete y diez mejores covariables según el ranking de la Tabla 1, y utilizando todas las covariables sin realizar la selección previa.

La evaluación de la capacidad predictiva de los diferentes modelos obtenidos se realizó mediante la tasa de acierto (o de error) de clasificación obtenida en el proceso de validación cruzada con 10 grupos. Este método consiste dejar fuera del entrenamiento 6 observaciones y evaluar la tasa de acierto (o de error) del modelo entrenado con las restantes 56. Este proceso se realiza para 10 grupos seleccionados aleatoriamente y se obtiene la media de dichas tasas de acierto. Ello permite estimar la capacidad predictiva de los modelos sobre casos que nunca han sido vistos por ellos.

La Tabla 3 muestra los mejores resultados obtenidos con los modelos utilizados indicando el número de covariables utilizado en ellos, la tasa de acierto media del proceso de validación cruzada y la matriz de confusión. Cuando un modelo produce análogas tasas de acierto para distinto número de variables se muestra sólo la tabla de confusión obtenida con el menor número de variables (7 variables).

Las matrices de confusión se muestran en una sola línea en la tabla por razones de espacio, pero están constituidas por dos filas (separadas por una coma). Si la matriz es [*a b*, *c d*] entonces *a* es el número de accidentes clasificados correctamente como tales, *b* es el número de accidentes clasificados erróneamente como incidentes, *c* es el número de incidentes clasificados erróneamente como accidentes y *d* es el número de incidentes clasificados correctamente como tales.

Sobre dicha tabla realizamos las siguientes consideraciones:

- En general, las distintas técnicas alcanzan sus mejores tasas de acierto con 7 variables, lo que no significa que todos utilicen las mismas variables en los modelos resultantes.

- Ello significa que, en general, el entrenamiento con muchas variables puede entorpecer la eficacia del mismo, lo que ratifica la importancia de realizar una selección previa de variables. Esto se observa sobre todo en las redes bayesianas no Naive. Excepciones a esta norma son los árboles de clasificación y las redes bayesianas tipo Naive. Las primeras, por su propio algoritmo de entrenamiento, implementan con éxito un método propio de selección. Las segundas porque no realizan selección de variables sino que proceden directamente a la estimación de probabilidades de todas las covariables condicionadas a los distintos valores de la variable respuesta.

- En cuanto a las tasas de acierto, los mejores resultados obtenidos son similares para un grupo de técnicas: destacan las redes bayesianas con algoritmo K2 con un 88.71%, los árboles J48 y las SVM, ambas con un 87.10% y las reglas PART con un 85.48%. Esta última técnica produce la misma estructura que los árboles J48 por lo que la ligera diferencia entre tasas de acierto se debe a la aleatoriedad de la validación cruzada.

- En sentido contrario, los peores modelos en predicción son: la regresión logística

(72.58%), que acaba pagando el coste de la excesiva simplicidad de su modelo lineal, y el algoritmo OneR de extracción de reglas (75.81%) que, por su carácter produce reglas excesivamente basadas en las particularidades de la muestra y tiende al sobreajuste.

- Teniendo en cuenta las matrices de confusión, se observa que con siete variables, la mayor parte de los errores de los modelos se producen en la clasificación de los accidentes. Ello es debido a que la muestra posee menor representación de este tipo de suceso lo que dificulta a los modelos su caracterización. Cuando se utilizan todas las variables los errores de clasificación de todos los modelos se reparten más entre accidentes e incidentes, pero la tasa de acierto disminuye lo que indica que la utilización de todas las variables favorece el sobreajuste.

Modelo	nº variables	% acierto	Matriz de confusión
BayesNet – K2 – 1 padre	7 y 10	85,48%	[10 8, 1 43]
BayesNet – K2 – 3 padres	7	88,71%	[12 6, 1 43]
BayesNet – K2 – 8 padres	7	88,71%	[12 6, 1 43]
BayesNet – HillClimber – 1 padre	7	85,48%	[10 8, 1 43]
BayesNet – HillClimber – 3 padres	7	79,03%	[8 10, 3 41]
BayesNet – TAN	10	83,87%	[11 7, 3 41]
Naive Bayes	7 y 10	83,87%	[9 9, 1 43]
Naive Bayes Simple	7 y 10	83,87%	[9 9, 1 43]
SVM – SMO	7	87,10%	[11 7, 1 43]
Logistic Regresion	7 y 17	72,58%	[8 10, 7 37]
Tree – ID3	7	77,42%	[12 5, 4 36]
Tree – J48	7, 10 y 17	87,10%	[11 7, 1 43]
Tree – LMT	7	82,26%	[10 8, 3 41]
Rule – PART	7	85,48%	[11 7, 2 42]
Rule – OneR	7, 10 y 17	75,81%	[4 14, 1 43]

Tabla 3: porcentajes de acierto y matrices de confusión de los modelos aplicados para el número de variables que proporcionó los mejores resultados.

Interpretación de los modelos

De todos los modelos evaluados, los que poseen mayor capacidad interpretativa (ya sea porque establecen reglas o lista de decisiones o modelizan gráficamente la relación entre variables) son las reglas de clasificación, los árboles de clasificación y las redes bayesianas. A continuación se analizan los resultados obtenidos por estos modelos desde el punto de vista de la interpretación del fenómeno de la accidentabilidad, analizando esta capacidad para un número de variables introducidas diferente.

El resultado más sencillo que se extrae de la interpretación de los modelos es aquel que explica el fenómeno de la accidentabilidad por medio de dos variables, Horas de operación (HO_d) y el Régimen de la empresa a la que pertenece el trabajador (R). Este resultado lo obtienen los tres modelos, la regla PART, las redes bayesianas y los árboles

de clasificación cuando se seleccionan siete covariables.

Como ejemplo de lo anterior, la Figura 1 muestra la estructura del árbol J48 que es análoga a la producida por el modelo PART y se interpreta de la siguiente forma:

- Si los trabajos que se están realizando son de corta duración (menos de 4 horas) el resultado es accidente (con una confianza del 85,71% y cobertura de 6 casos).
- Si la duración es de 4 a 8 horas, el resultado es incidente (cobertura 6, confianza del 100%).
- Si la duración es de más de 8 horas, el resultado depende del régimen de contratación de la empresa, siendo accidente si es subcontratada (5 casos correctos y confianza del 100%) o incidente si es contrata principal (37 casos correctos y 84,09% de confianza).

Las redes bayesianas ofrecen una perspectiva mucho más interesante que los modelos anteriores en cuanto a la interpretación del fenómeno. Coinciden con los modelos anteriores en darle mucho peso a las variables Hora de operación y Régimen, encontrándose en todos los modelos que la probabilidad de accidente es mayor en la actividades que duran menos de cuatro horas y que se desarrollan bajo un régimen de subcontrata. Pero además, al tratarse de modelos gráficos que se actualizan automáticamente tras la incorporación de nuevas evidencias (es decir, permiten un análisis *what if* inmediato), las redes muestran una ventaja con respecto a los otros modelos en la interpretación global del fenómeno de la accidentalidad.

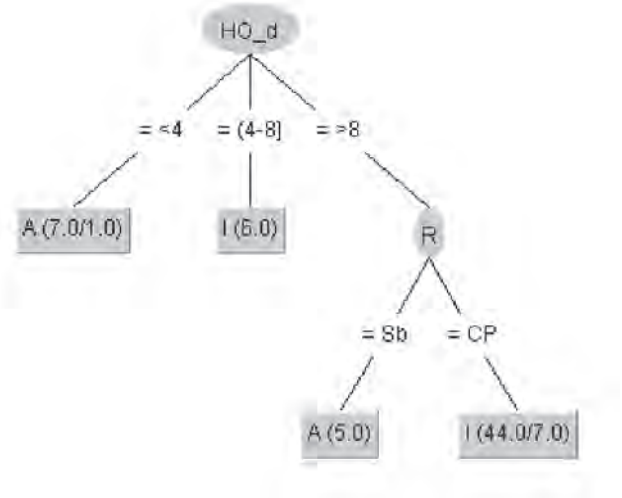

Figura 1: Estructura producida por el árbol J48. Los números entre paréntesis indican el número de observaciones de cada clase (accidente o incidente) y la letra en mayúscula indica la clase mayoritaria. Así por ejemplo, el primer nodo de la izquierda indica que si la variable HO_d es menor o igual que 4, se obtiene una mayoría de accidentes (7 casos) por un solo caso de incidentes.

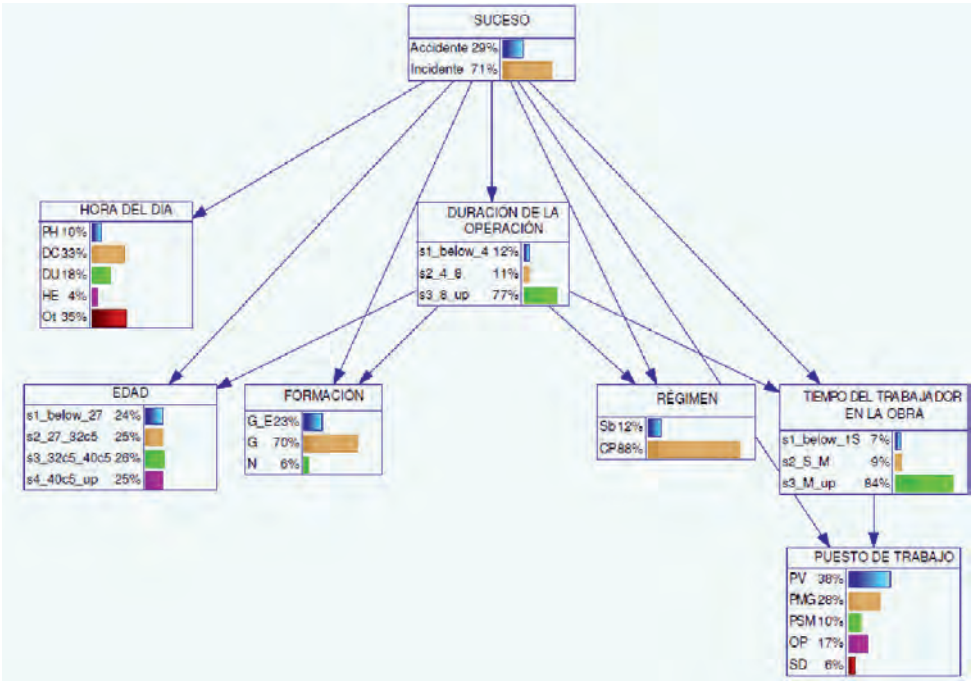

Figura 2: Red bayesiana greedy K2 estructurada con siete variables y 3 padres. Las barras indican la probabilidad estimada de cada uno de los valores de las variables.

El modelo bayesiano más sencillo que ofrece este resultado, que es el que además muestra uno de los mejores porcentajes de acierto en la predicción, es la red bayesiana greedy K2 con las siete primeras variables causales y 3 u 8 padres, cuya estructura se muestra en la Figura 2. Este modelo indica que un mayor porcentaje de accidentes está relacionado con empleados procedentes de subcontratas, que realizan actividades de duración menor de cuatro horas, de edades mayores de 40 años y con un tiempo de permanencia en la obra menor de una semana. Los accidentes están asociados con puestos del tipo "Sin determinar" u "Otros puestos" y con un mayor porcentaje de casos de formación insuficiente.

Coincidiendo con los otros modelos, las redes parecen darle el mayor peso a las variables Horas de operación y Régimen, permitiendo comprobar que esta asociación reside en las particularidades del colectivo de empleados procedentes de subcontratas: la red señala que este colectivo se corresponde con empleados de edades mayores de 40 años, con un tiempo de permanencia en la obra menor de una semana, trabajos de muy corta duración y puestos no determinados. Sin embargo, comparativamente al grupo de empleados procedentes de la contrata principal, poseen un nivel de formación específica ligeramente más elevada.

En la misma red pero estructurada con más variables, se extraen otros hechos interesantes:

- Escasa o nula influencia en la distribución de accidentes/incidentes de las variables Evaluación de riesgos, Edad, Tipo de contrato y Antigüedad.

- La variable Dirección y supervisión se relaciona con las variables Evaluación de riesgos y Suceso de manera que cuando la dirección y supervisión es correcta son mayores los casos en los que se evalúan los riesgos y en esta situación son más probables los incidentes (80% de los casos). Por el contrario, si la dirección y supervisión es deficiente se incrementa el número de casos de accidentes (51%).

- Se observa una estrecha relación entre la variable Condiciones de trabajo y Evaluación de riesgos, obteniéndose que si se evalúa el riesgo, las condiciones de trabajo son buenas en el 83% de los casos; si no se evalúa el riesgo, el número de situaciones en las que las condiciones son buenas se reduce al 67%. Igualmente, las condiciones de trabajo buenas están asociadas a un mayor porcentaje de actividades que duran más de ocho horas, a actividades de trabajadores que llevan más de un mes en la obra y a puestos de trabajo específicos (las tres primeras clases de esta variable); todas estas situaciones están asociadas a un mayor número de incidentes. Por el contrario, las condiciones de trabajo no aceptables se asocian con un mayor número de actividades de corta duración y con puestos no determinados, condiciones que todos los modelos asocian con un mayor número de accidentes.

Discusión

Los resultados obtenidos en este trabajo suponen un importante avance en el manejo de la información relacionada con la siniestralidad en el trabajo, en este caso procedente de encuestas posteriores a la consecución de incidentes/accidentes en el sector de la construcción y minería. Por una parte, permiten definir un protocolo de análisis estadístico basado en técnicas de minería de datos que permite: 1) seleccionar la información más relevante de toda la disponible, lo que incrementa los porcentajes de acierto de los modelos predictivos y la capacidad interpretativa, y 2) identificar las herramientas de minería de datos más interesantes, como modelos predictivos e interpretativos, en el campo de estudio de la siniestralidad.

Por otra parte, las conclusiones que se extraen de los modelos con capacidad interpretativa acerca de las causas de los accidentes/incidentes del grupo de trabajadores encuestados son muy interesantes ya que identifican las circunstancias que rodean la ocurrencia de accidentes permitiendo definir claramente su casuística.

Así, en la fase de selección de variables, todos los modelos aplicados coinciden en señalar las mismas variables, entre todas las consideradas, que más peso tienen en la variable de predicción accidente/incidente. Esto constituye un método riguroso de filtrar información no relevante que no sólo va a permitir obtener mejores resultados de aprendizaje y predicción de los modelos sino también elaborar encuestas mucho más sencillas de cumplimentar y por tanto más eficaces en el registro de la información.

De entre todos los modelos estudiados, los más interesantes en cuanto a la información que ofrecen del fenómeno de accidentes/incidentes en el colectivo analizado resultan ser los modelos que ofrecen reglas (árboles y reglas de clasificación) y los modelos bayesianos, los cuales son, todos ellos, los que muestran mejores porcentajes de acierto en cuanto a predicción. Este resultado coincide en general con el obtenido en un trabajo anterior (Matías et al., 2008) pues las mismas

técnicas mostraron mayor capacidad de predicción de tipos de caídas a partir de una base de datos extraída de partes de accidentes. Ello indica que la estructura de este tipo de técnicas se adecúa muy bien a la de los datos que habitualmente se dispone en este campo de estudio.

Las reglas y árboles ofrecen una información idéntica del fenómeno estudiado cuando se incorporan pocas variables (las más relevantes) en los modelos. Pueden resultar por tanto interesantes de aplicar cuando se pretende identificar la situación tipo asociada a una mayor probabilidad de un suceso, como ha sido en este caso la asociación entre accidente y dos situaciones determinadas relacionadas con la duración de las operaciones y el régimen de la empresa. Las redes bayesianas suponen una ventaja añadida: no sólo ofrecen la misma información que los árboles y reglas en identificar cómo influyen las variables más relevantes en el fenómeno sino también que gracias a su capacidad de análisis tipo what-if permiten una exploración más profunda de los datos y la obtención de los distintos escenarios que se relacionan con los accidentes y con los incidentes en el colectivo analizado.

Desde un punto de vista de la gestión de la seguridad, los resultados obtenidos señalan claramente cuál es el escenario asociado a la accidentabilidad en el colectivo analizado: se trata de actividades que duran poco tiempo, que no son específicas y que son realizadas por empleados procedentes de empresas subcontratadas que llevan poco tiempo trabajando en la obra. Una vez identificado el escenario asociado a la accidentalidad, los modelos gráficos permiten investigar las posibles causas asociadas a esta siniestralidad: así, se descartaría, por ejemplo, que la mayor accidentalidad en este grupo de empleados concretos (es decir, los que definen el escenario anterior) pueda deberse a un menor nivel de formación, ya que las redes bayesianas indican que estos empleados poseen formación de nivel similar o incluso superior a empleados de contrata principal.

Otro hecho relevante es que en la fase de selección de variables, las variables relacionadas con la gestión de la seguridad hayan quedado en puestos intermedios o en los últimos puestos de relevancia. Así, las redes bayesianas no muestran una relación directa, en el colectivo estudiado, entre la accidentalidad y variables como la Evaluación de riesgos, Condiciones de trabajo o Dirección y Supervisión, pero lo que sí encuentran, por ejemplo, es que los accidentes no se asocian con situaciones en las que existe una Dirección y Supervisión, Evaluación del Riesgo y Condiciones de trabajo adecuadas. Desconociendo las circunstancias particulares que pueden rodear la actividad desarrollada por las subcontratas en este colectivo, es indudable que estos hechos indican la necesidad de investigar las causas de los accidentes que estén relacionadas con las actividades de gestión y control.

Agradecimientos

La participación de J. M. Matías ha sido financiada por el Ministerio de Ciencia e Innovación del Gobierno de España mediante el proyecto MTM2008-03010

Bibliografía

Bird, F.E.; Germain, G.L. (1990). Practical loss control leadership. International Loss Control Institute (Publ.) Revised Edition. 446 pp.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Classification and Regression Trees. Wadsworth.

Groves W.A., Kecojevic V.J., Komljenovic D. (2007). Analysis of fatalities and injuries involving mining equipment. Journal of Safety Research 38 461-470.

Han J., Kamber, M. (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hastie T., Tibshirani, R., Friedman J. (2001) The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer.

Jensen, F. V. (2001) Bayesian Networks and Decision Graphs. Springer.

Karra V. K. (2005). Analysis of non-fatal and fatal injury rates for mine operator and contractor employees and the influence of work location. Journal of Safety research 36 413-421.

Ley 31/1995, de 8 de noviembre (Ley de Prevención de Riesgos Laborales). BOE (Boletín Oficial del Estado) nº 269, de 10.11.1995. Transposición de la 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.

Martín, J. E., Rivas, T., Matías, J. M., Taboada, J., Argüelles, A. (2009) A Bayesian network analysis of workplace accidents caused by falls from a height. Safety Science, 47, 206-214.

Matías, J. M., Rivas, T., Martín, J. E., Taboada, J. (2008) A machine learning methodology for the analysis of workplace accidents. International Journal of Computer Mathematics, 85, 559-578.

Paul, P.S.; Maiti, J. (2007). The role of behaviour factors on safety management in underground mines. Safety Science 45, 449-471.

REAL DECRETO 39/1997, de 17 de enero, por el que se aprueba el Reglamento de los Servicios de Prevención. BOE núm. 27 de 31 enero.

Scholkopf, B. and Smola, A.J. (2002) Learning with Kernels. MIT Press.

Seber G. A. F. (1984) Multivariate Observations. John Wiley.

Ural S., Demirkol S. (2008). Evaluation of occupational safety and health in surface mines. Safety Science 46 1016-1024.

Witten, I. H., Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.