International Journal of Industrial Engineering & Production Research (2019) 30: 149-164
DOI: 10.22068/ijiepr.30.2.149

RESEARCH PAPER

Decision Support Approach to Occupational Safety Using Data Mining

Naghmeh Khosrowabadi[1] & Rouzbeh Ghousi[2]*

Received 03 July 2018; Revised 16 March 2019; Accepted 08 May 2019; Published Online 20 June 2019

ABSTRACT
For an industry to develop, occupational safety is a key factor in protecting the worker's health, achieving organizational goals, and increasing productivity. Therefore, research is required to investigate those factors affecting occupational safety. Based on the information gathered from the paint halts of an industrial unit in Tehran, this research uses data mining technique to identify significant factors. First, based on a literature review up to 2018, an insight into existing approaches and new ideas were obtained. Then, with significant 5600 units of data, the results of the charts, association rules, and K-means algorithm were used to extract the latent knowledge with the least error without human intervention by a six-step Crisp methodology. The results of charts, association rules, and K-means algorithm for clustering are in line and have been successful in determining effective factors such as important age groups and education and identifying important events and the halls; finally, the root causes of major events were the research questions. The results reveal the significant trace of very young and young age with often diploma education and low experience in major accidents involving bruising, injury, and torsion, often due to self-unsafe act and unsafe conditions such as slipping or collision with things. In addition, the important body members, hands, and feet in color retouching and surface color cabins are at risk. This paper uses a real case study and applies different approaches to improve safety strategy, which has received insignificant attention in this literature. Finally, suggestions for future research are presented.

KEYWORDS *Occupational safety, Data mining, CRISP, Association rules, K-means algorithm.*

1. Introduction

The significance of safety considerations in all areas, especially the safety of the workplace or occupational safety, is clear. Developing guidelines to prevent reducing probability or severity of the unexpected safety events can improve the safety of the workplace. Therefore, the purpose of this paper is to identify the factors affecting occupational safety in the paint halls of one of the industrial units in Tehran province. The hazardous activities of these halls cause significant damage to the workforce and, finally, result in productivity reduction in the industrial unit.

Recorded data on accidents of industrial units can be collected in a significant amount. In addition, the complexity of relationships between its effective variables requires a powerful approach. One of the most common approaches to dealing with such an issue is the data mining technique. An approach enjoying the least error and no human intervention can extract hidden knowledge from the data. Therefore, this technique has been used in this research.

Data mining is synonymous with one of the expressions of knowledge extraction, data retrieval, and data analysis. Therefore, the basis of data mining is an important process of identifying potential, useful, new, and ultimately understandable data models. Data mining can be used as an advanced tool for the introduction and analysis of decision-maker data. The term data mining is used by statisticians, data analysts, and the management information systems community.

According to one of the definitions, data mining, i.e., the discovery of knowledge in the databases, involves the extraction of potentially useful

* **Corresponding author:** ***Rouzbeh Ghousi***
Ghosui@iust.ac.ir

1. M.Sc., Department of Systems Optimization, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
2. Assistant Professor, Department of Systems Optimization, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.

information from data that were unknown. This issue includes some of clustering methods, data summarization, classifying rules, finding network relationships, analyzing changes, and discovering irregularities. Data mining is the discovery of interesting, unexpected and valuable structures from a huge amount of data and is an activity that basically matches the statistics and accurate data analysis [1].

There are a variety of software packages in this field, among which Clementine software is more popular because of its ease of use and no need for coding. To better illustrate the reason for choosing this tool, the two methods of statistical analysis and data mining are compared below.

In statistical analysis, statisticians always start with a hypothesis. They use numerical data. Statisticians should create relationships that are relevant to their hypotheses. They can identify inaccurate data during the analysis. In addition, they can interpret their work results and express them to managers. However, data mining does not require a hypothesis. Different data mining tools can use different types of data such as numerical data. The algorithms of this method automatically create relationships. Of course, data mining requires correct data. Data mining results are relatively complex and require interpretation by experts.

1-1. Explaining the problem

According to the purpose of this study, by extracting knowledge about important factors in paint halls' safety accidents, the question revolves whether data mining tools are used for the following purposes in the field of safety or not. This study follows the following issues:

1) Identification of factors affecting safety according to the related data such as the number of accidents.

2) Extracting factors that are difficult for humans such as the root causes of events.

3) Determining the success or failure rate of the halls in reducing the number of accidents and days lost.

4) Identifying the characteristics of the workforces who have caused major safety accidents.

A six-step CRISP-DM[2] approach has been used to analyze data using a data mining tool. CRISP is the most popular methodology in the field of data science, and it consists of six steps, as shown in Figure 1. The sequence of steps is not direct. Moving back and forth between different stages is always required. The output of each stage can determine the specific task of the next step or what stage should be performed. In Section 3, the steps are completely explained.

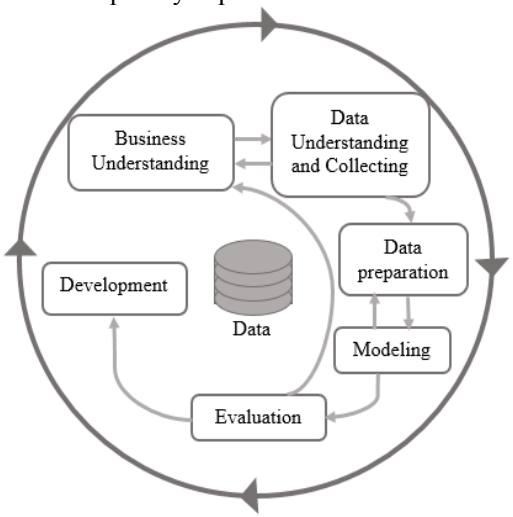


Fig. 1. The steps of the CRISP method

Following the above steps, in Clementine software, drawings of graphs in the graph tab are performed and, in the modeling tab, associative rules and K-means algorithm for clustering have been used. Association rules are one of the main techniques of data mining and, of course, the most important in discovering and extracting patterns in learning sets. This method retrieves all possible patterns in the database [3]. These rules represent the relationships and interdependencies between a large set of data. A common example is the shopping analysis. For example, customers who come to buy bread often buy milk [4]. Of course, there are different criteria for assessing the validity and generalizing the ability of these patterns. For this part, the GRI[3] algorithm has been used. The logic of another used algorithm, K-means, is that the elements inside the cluster are similar to each other and are different from those of other clusters. One of the things that should be considered in dealing with this algorithm is the optimal number of clusters, which is denoted by k in the software.

2. Literature Review

This section presents a brief description of studies with safety and data mining keywords; then, they have been categorized in the tabular form.

Most models of occupational accidents in the construction industry are made up of several factors. However, statistical techniques can be

[2]Cross-Industry Standard Process for Data Mining

[3]Generalized Rule Induction

used to identify causal relationships between these factors. The multiplicity of factors and the complexity of communication between them make the identification of potential risk points and development of safety guidelines for managers in construction projects difficult. For this purpose, a study was conducted to investigate this problem by using association rules and a survey of 1347 construction incidents in Taiwan in 2000-2007. This study shows that events occur when a certain combination of factors occurs. Working at high altitudes without safety tools, loss of balance during movement, defective safety equipment, inadequate experience, and contacts with unstable structures can lead to serious accidents. These results can help managers formulate effective safety policies along with budget management and employee training [5].

Construction accident research involves the systematic ranking, categorization, and conceptual coding of databases related to injuries and casualties. A study investigates the causes and distribution of occupational accidents in the Taiwan construction industry by analyzing data with data mining and using the regression tree and classification. In this regard, 1542 incidents were conducted during 2000-2009 to find the causal relationship in this area. The results showed that the rules of falling probabilities and crashes in both types of private and public construction projects are used as key factors in predicting the probability of occupational injuries. The results of this study provided a framework for improving safety training and critical training programs to protect construction workers from destructive accidents [6].

The usefulness of using data mining in many areas is clear. However, in the analysis of occupational accidents, the application of this technique is still rare. A study of decision tree and association rules in Finnish occupational accidents was conducted on the basis of statistical damages related to these incidents, including slipping, stumbling, and falling associated with events of 2006 and 2007. The portion of these three types of incidents is 22% compared to all incidents. In addition, the significance of these incidents is remarkable, given the stop and slowness of work against other incidents. The results showed that the most important factor associated with these events is how physical activity develops during work. Their risk depends on the type of work and the age of the worker [7].

Tab. 1. Literature review on data mining in the field of safety

Ref.	Authors	models	Goal and Content	Field
[5]	(Chang et al., 2010)	Association rules	Discovered causal relationships in Taiwan Occupational Incidents	Construction
[8]	(Kokotos & Linardatos, 2011)	Classification tree	Provided a decision support system and safety improvements in limited water ships	Shipping
[9]	(Rivas et al., 2011)	Classification tree and Bayesian network	Identified the important factors in the occurrence of occupational accidents and provided an approach to their prediction	Mine, Construction
[6]	(Cheng et al., 2012)	Classification and Regression tree	Discovered the factors affecting job damage for Taiwan construction industry	Construction
[10]	(Silva & Jacinto, 2012)	Patterning	found patterns of occupational accidents in Portugal and support for strategic safety decisions	Extraction
[7]	(Nenonen, 2013)	Decision tree and Association rules	Analyzed the factors related to important occupational accidents including slipping, stumbling, and falling in Finland	Various industries
[11]	(Cheng et al., 2013)	Classification and Regression tree	Applied the data mining approach to the analysis of major occupational factors in Taiwan	Petrochemicals
[12]	(Maiti et al., 2014)	Hybrid classifications	Improved safety in the rail's event	Steel
[13]	(Verma et al., 2014)	Association rules	Found safety patterns for industrial events	Steel
[14]	(Argilés-Bosch et al.,	Experimental analysis	Data analysis of workplace incidents on financial performance, attention to return on investment	Various companies

	2014)			
[15]	(Hajakbari & Bidgoli, 2014)	Classification and other related methods	Presented a new system for assessing the risk of occupational accidents using data from Iran's Ministry of Labor	Various fields
[16]	(Sanmiquel et al., 2015)	Bayesian classification and Decision tree	Found a behavioral pattern against mining incidents and developed preventive policies	Mine
[17]	(Murè et al., 2017).	Self-organized Map, K-mines, Fuzzy approach, Neural Network	Provided job risk assessment guidelines, considering the factors affecting their dynamics	Manufacturing industries
[18]	(Sarkar et al., 2017)	Classification and Regression tree	Predicted occupational incidents using proactive and reactive data	Various industries
[19]	(Liu et al., 2018)	Hybrid	Integrated schematic framework supporting decision-making process management for water pipes	Water
[20]	(Das et al., 2018)	Association rules	Determined the relationship between effective factors in driving accidents	Transportation
[21]	(Gregoriades & Chrystodoulides, 2018)	Clustering and self-organized map	Analyzed historical traffic accident data to extract their safety management knowledge	Transportation
[22]	(Solanke & Gotmare, 2018)	Classification and Association rules	Found safety measures to reduce road traffic accidents and injuries	Transportation

In a study, the analysis of the data related to tools' exit from the line of steel plant was analyzed. The primary purpose of this research was to find a series of relationships between factors that affect tools' exit from the rails, which ultimately led to the development of conceptual rules for line protection. Then, 348 tools' exit incidents were collected over a period of 42 months, with 4 factors of shifts, location, reason, and department responsible for their analysis. The descriptive statistical tool showed that shifts did not change the likelihood of an accident. However, other factors were still noticeable. Raw materials, manual operations, and production represent 50%, 60%, and 28.88% of this type of incident, respectively. By using related analysis, it was found that the level of movement, human intervention, wagon management, and critical movements were latent root factors [12].

Another study was conducted to find patterns for the events of steel factories in India. Occupational events in these industries are essentially in the form of injury, disability, hurt, or as a combination of the various factors responsible for such incidents. An accident investigation scheme was presented. Association rules were used to discover the causal patterns by examining 843 events. Thirty-five Conceptual rules using this tool were presented, according to three indices including support, confidence, and lift. For example, the results showed that another unsafe act was repeated in injury cases (with 4.86% support, 78.8% confidence, and 3.2 lift). Keeping these results in mind, ideas for improving safety by preventing accidents were presented [13].

The mine is an economic sector with high accident rates. Mines are the workplaces at risk and their workers may suffer from a variety of injuries. With approximately 70000 reports of occupational accidents and casualties during the years 2003 to 2012 in Spanish mines, a research article aimed at analyzing the main causes of the incident was presented. Powerful statistical tools such as Bayesian classification and Decision tree were used [16].

Another study provided instructions for risk assessment of occupational accidents with their self-organized map, K-means algorithm, and fuzzy approach, which clearly explored the impact of occupational environments on the dynamics of occupational accidents. The guidelines included two assessment steps: 1) identifying the dynamics of occupational

accidents in the manufacturing sector through the use of the neural networks; 2) assessing the occupational accident risks for a specific plant with fuzzy logic and extracting knowledge related to the manufacturing industry events analysis. Information about an Italian manufacturing unit from 2006 to 2013 was used. The risk assessment was carried out, showing clearly how critical work environment affected the dynamics of accidents and their management improved safety [17]. A study presented a tool for analyzing historical traffic accident data using data mining techniques to extract the knowledge required for traffic safety management. Clustering analysis was carried out by a self-organizing map to identify the major points of events and their presentation on a map [21].

3. Research Methodology

In this research, the Clementine Software 12, which is a
SPSS[4] software series, is used for data analysis and modeling. As mentioned, the CRISP-DM method was used to explore the knowledge of the data.

In order to analyze the safety data through the Clementine software, it is necessary to enter the relevant data after preparation and preprocessing into the software, which has many features. The greater number of data and variables may enhance the tool's ability to discover meaningful relationships. In this research, about 5,600 units of data on the events of the paint halls of one of the industrial units located in Tehran were investigated.

3-1. Business understanding

In this basic stage, the considered organization has been investigated. In order to realize this aim, issues such as data recognition, the purpose of the study, and existing problems have been investigated. Accuracy in this stage is meant to imply an increase in the accuracy of the results and their proximity to reality. This study was performed based on the data collected from the safety information of an industrial unit, which aims to identify the factors that affect the safety of paint halls. The unit was established in 1345 and, now, the production volume of Salon 1 includes about 440 vehicles per day. Paint hall is the second hall of this industrial unit. In this section, by applying advanced technicians and equipment, the painting process is performed accurately by using robots. In this hall, the metal body of the car requires a protective layer after cleaning; therefore, the car room is placed on a movable base and immersed in the color pond. Then, the car body enters a furnace to cook the protective layer in the room. Finally, after finishing the insulating, sealing, and primer paints, the final painting of the car begins. Color spraying with 16 robots is done, and advanced paint lines can paint up to 30 cars' room per hour. The main objective of the industrial unit is to increase production efficiency and customer satisfaction. Of course, like all other active units of the country, it faces challenges and external factors, such as currency fluctuations, and requires scientific methods. The aim of this study is to improve the safety and working conditions of the employees in order to improve the efficiency and reduce days lost.

3-2. Data understanding and collecting

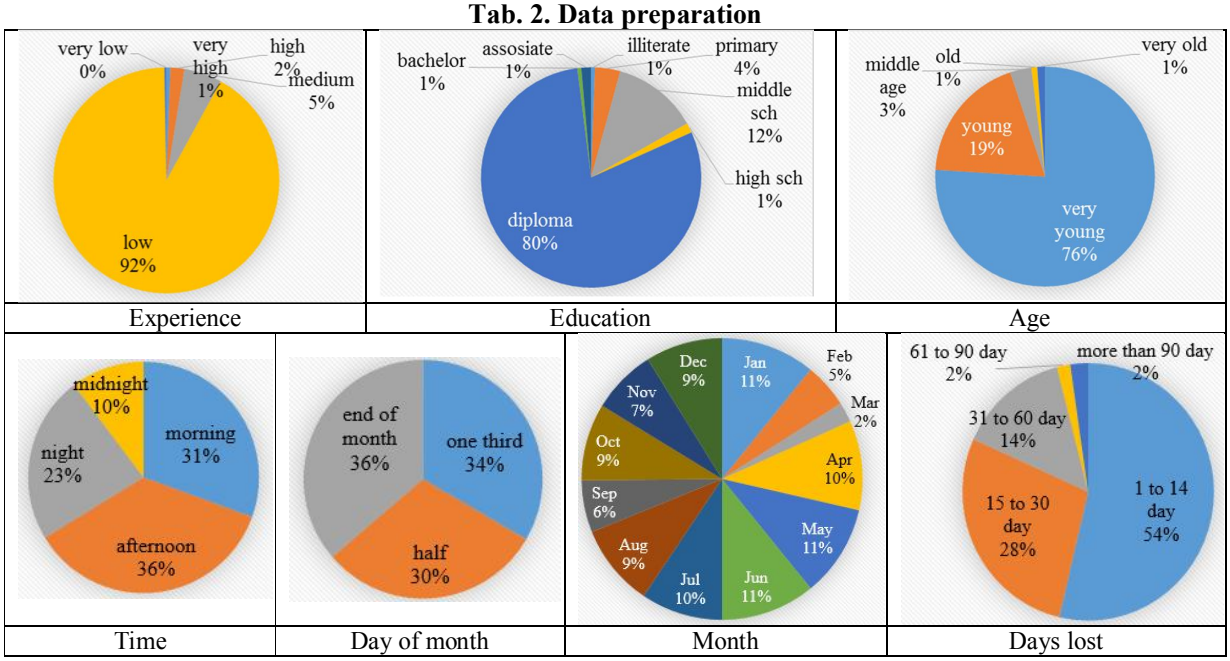
Data have been collected from the safety information of paint halls in 14 years from 2002 to 2017. The Excel file contains 350 rows and 16 main columns, which, according to the models and graphs in the software, are separated by some columns. The first column is the row number that identifies the injured workforce. This is a unit column that does not depend on any item. The other columns include experience, education, age, time, day of the month (one third of the month, half or end of the month), month, year, day of the week, injured body member, number of days lost, management (i.e., responsibility of the event hall), location of the accident, cause, the way the accident occurred, and the result (i.e., type of accident).

3-3. Data preparation

Data preparation is required to better utilize graphs and modeling algorithms. Here is a description of the data preparation. Each group is shown in Table 2.

[4]Statistical Package for Social Science

Tab. 2. Data preparation



The column of injured body members, causes, how, and results do not need to be grouped, and they are denoted by their own name. Now, the data have been prepared and ready to use descriptive statistical tools such as graph and modeling algorithms.

3-4. Modeling

In Clementine software, the down tabs provide all the necessary features according to the CRISP approach from the data entry stage to the presentation of the results. Draw graph types in the graph section and in the modeling section, association rules, and K-means algorithms can be used. One of the methods for determining the number of optimal cluster K is amounting it at an interval of 0 to 10. The cluster number with the least sum of squares error could be chosen. This is in fact similar to the elements within the cluster with a significant difference from other clusters. In this research, four indicators to determine the optimal cluster are used. The cluster number that is optimized for more indicators is eventually selected. The following is a brief explanation of these indicators.

Dunn Index: This clustering indicator is used to identify a dense cluster set, which is a small difference between the members of each cluster and a large difference with other clusters. As a result, the average clusters are different from their cluster members [23]. The method for calculating this index is shown in Equation (1). Greater amount of this indicator is desirable.

$$DI_m = \frac{min_{1\le i<j\le m}\ \delta(C_i, C_j)}{max_{1\le k\le m}\Delta_k} \quad (1)$$

In this formula, $\delta(C_i, C_j)$ is the distance between the center of cluster $C_i$and the center of clusterC_j, while i ≠ j. Δ_i is the maximum distance between all the pairs in cluster i.

Davies–Bouldin Index: This index was introduced in 1979 for measuring clustering with respect to the dispersion within each cluster and their separation [24]. The method used for calculating this index is shown in Equation (2). Scattering measure within each cluster is $S_i = (\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j - C_i|^p)^{1/p}$. The number of elements in each cluster is T_i and element j of dataset is X_j. The center of cluster j is C_i. Scattering measure among clusters is $M_{ij} = (\sum_{k=1}^{n}|C_{ki} - C_{kj}|^p)^{1/p}$, including C_{ki} as the K^{th} dimension for the center of cluster j. If it is assumed that $R_{ij} = \frac{S_i+S_j}{M_{ij}}$ and $D_i = max_{j\neq i}R_{ij}$, and the number of clusters is N, then

$$DBI = \frac{1}{N}\sum_{i=1}^{N} D_i \quad (2)$$

Sum of Squared Error Index: The sum of squares is the distance of each element from its center of cluster. The lower amount of this index is desirable. The calculation method is given in

Equation (3) [25]. To determine the optimal cluster, the ratio of this index is used as $\min(\frac{SSE_i}{SSE_{i-1}})$.

$$SSE_i = \sum_{k=1}^{i} \sum_{j=1}^{T_k} (X_j - C_k)^2 \quad (3)$$

Silhouette Index: This is an index for the measurement of cluster adaptation, which indicates the similarity and dependency of each element to/on its cluster and the difference or degree of separation from other clusters, ranging from -1 to 1. The greater amount of this index, which is shown in Equation (4), shows a greater coordination of the element with its own cluster and its incompatibility with other clusters. If, for most clusters, negative or small values are obtained, there is a sign of a very small or very large number of clusters [26].

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \quad (4)$$

where a_i is the average distance between *i* and all other elements in its cluster. a_i can be assumed as a measure of how well *i* is assigned to its cluster. b_i is the lowest average distance of *i* to all elements in any other cluster, and *i* is not a member.

3-5. Evaluation

Evaluation and validation of the results ensure the user that they can use the results in practice. There are different approaches to validating the created models. Otherwise, it is not possible to find attractive rules. Suppose that there is a rule called R, defined as A giving to B, where A and B are subsets of the objects. In order to validate the two criteria, support and confidence must be introduced.

From the ratio of the transactions, where both A and B are there, to the total number of records, the support number has been calculated, which has a value between zero and one in the form of percentage. The greater amount of this measure shows that the two objects in the law are more related. By specifying the threshold for this index, the user can only obtain rules with a higher number of supports than the minimum specified. In doing so, the reduction of the search space can, in turn, reduce the time required to find association rules. The use of the support measurement is not enough alone. Hence, confidence measurement is introduced. The value of this index is also a number between zero and one in the percentage form. Confidence is derived from the ratio of supports A and B to the number of supports A. Equation (5) is used to express this. The greater value of this index increases the quality of the rule. The concept of this measure is used to determine those cases in which the occurrence of element A leads to the occurrence of element B. In other words, if the confidence level is 100%, whenever A occurs, it will definitely result in B's consequence.

$$Confidence = \frac{Support\ (A,B)}{Support\ (A)} \quad (5)$$

The application of this measure along with the support percentage is a good complement to the evaluation of association rules. However, it is still possible for a law with high confidence and low quality to exist. Among the other measures for evaluating association rules, Lift index is considerable. This is a numerical index between zero and infinity. In fact, the coincidence of characteristics is considered with regard to Lift. Contrary to the confidence measure, it considers the occurrence of consequent section B alone. The values close to 1 represent the independence of the two parts of the rule, and the values less than 1 represent the inverse relationship between the antecedent and consequent parts. The value of this index higher than 1 means that A provides more information about B. In other words, this criterion indicates the degree of independence or dependence of factors A and B. Thus, the attractiveness of rule (A gives B) will increase. Therefore, sorting the rules according to this index can also be helpful.

3-6. Development

After performing all the steps and evaluating models, the validated results need to be implemented. At first, it is necessary to take some basic actions: the recording of more accurate data, more information about the environment, the interaction of line officials and workers, and the willingness of managers to collaborate and implement recommendations from data mining results. In this way, it is possible to implement the ideas, stream them, and gain feedback on how they function in the industrial unit.

4. Research Results

The data of this research includes all the necessary specifications for the salon safety

performance. An example of data is presented in Table 3.

Tab. 3. An example of data

Variables	Prepared data			
Number	1	2	3	4
Experience	medium	medium	medium	Low
Education	diploma	diploma	diploma	Diploma
Age	very young	young	very young	very young
Time	afternoon	midnight	midnight	morning
Day of month	one third	half	end of month	one third
Month	Apr	Apr	May	Jun
Year	2009	2009	2009	2009
Day of week	Tue.	Tue.	Wed.	Wed.
Body member	leg	head	leg	other
Days lost	about 2 months	2 weeks or less	about 2 months	2 weeks or less
Management	color 2	color 2	color 2	color 2
Location of the accident	Salon surface 2	Pussy 2	PT color 2	Surface color 2
Cause	self-unsafe act	unsafe conditions	self-unsafe act	self-unsafe act
How	slip	falling objects	fall down	falling objects
Result	Torsion	Bruising	Fracture	Bruising

The first item to consider is whether the unit in question needs to be improved and what the current situation is. By charting the number of accidents and days lost collected over 14 years, obtained results are in a downtrend. However, the ratio of days lost to the number of accidents in Figure 2 indicates that the unit needs to be improved.

After identifying the need for an improvement approach, the pie chart for age, shown in Table 2, was used to identify the important age group, indicating that most of the workforce and age group are very young and young.

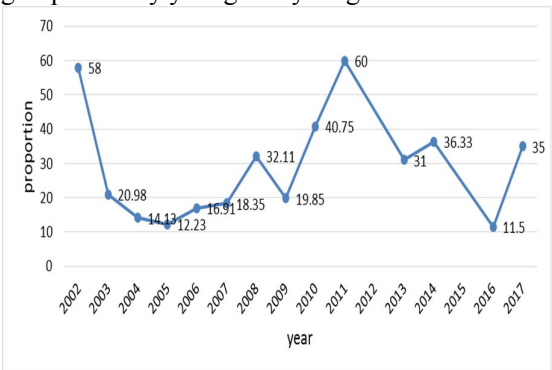


Fig. 2. The ratio of the number of the days lost to the number of accidents in different years

Currently, in order to survey the work experience of this important age group, the age versus experience chart is drawn in Figure 3, which shows that most workforce has insignificant experience.

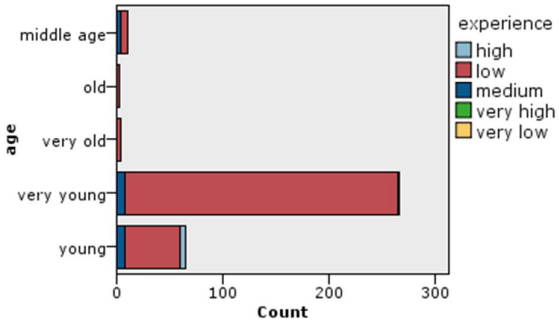


Fig. 3. Age vs experience chart

Further, Figure 4 shows the aim of investigating education status in the important age groups, clearly demonstrating that most of them hold a diploma degree. Meanwhile, at a young age, few workforce with a middle-school degree is also seen.

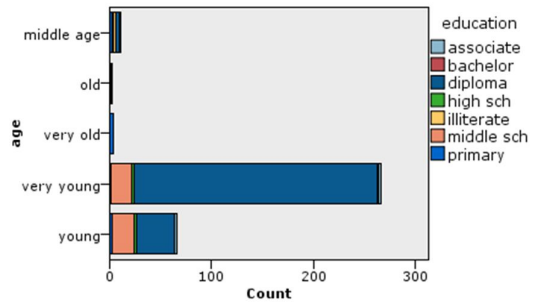


Fig. 4. Age vs education

One of the useful graphs is the multiple charts of the body member versus age, results, and days

lost. Figure 5 shows that young and very young age are more diverse in results and days lost.

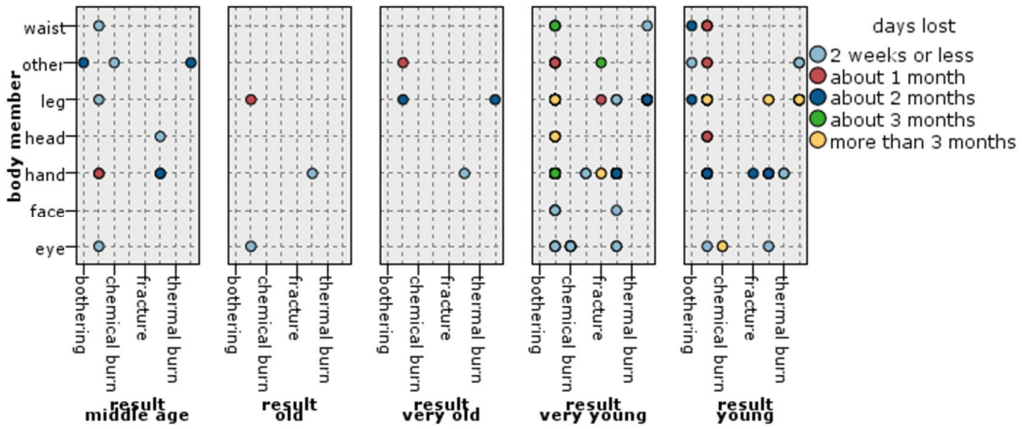


Fig. 5. Body member vs age, accident and days lost

Leg body member has the most variety of results, and this member suffers from more severe accidents due to having more than three months of days lost. In addition to the leg, hand, head, and eye are important according to the days lost. In important age groups, leg and hand members are important. In terms of accidents' significance, bruising has more variation in body member and days lost. After that, fractures and injuries are also important accidents, which affect various body members and their effects are also severe.
The results are presented in Table 4. Note that, this table considers only 3 months and more than 3 months of days lost.

Tab. 4. Variety of accidents and body member

Age	Body member	accident
Young	Waist and hand	Bruising
	other	Fracture
Very young	Hand and leg	Bruising
	Hand	Fracture
	Eye	Chemical burn
	leg	Injuries
	leg	Torsion

Figure 6 illustrates the distribution of body members at different ages to re-examine the most seriously injured members. As has been mentioned, at the young and very young age, the body members are of significant importance.

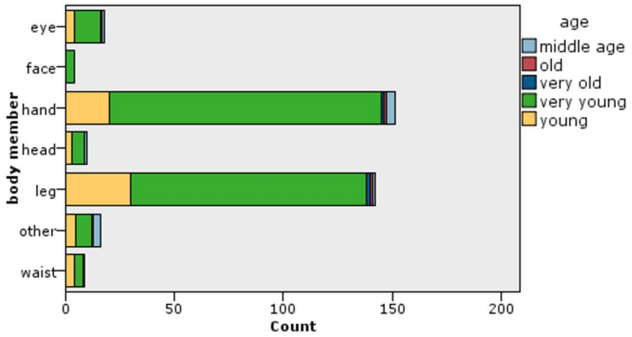


Fig. 6. The distribution of body members at different ages

Given the importance of the hand and leg as body members, the questions are as follows: what are the most consequences related to these body members? Are the accidents that have been previously identified as important are re-approved or some accidents have been ignored?

As shown in Fig. 5, the variation of the elements is controversial, and there is no information about the number of each element. According to Fig. 7, most of the accidents on the leg area involve bruising and injuries and, also, bruising and torsion related to the hand area. Both cases share bruising. Therefore, this result is important and should be focused on its number and severity.

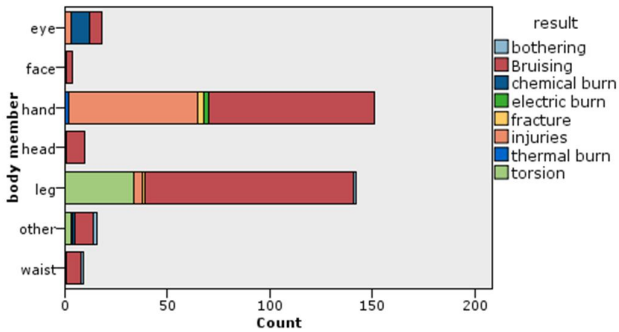


Fig. 7. Body members versus results

Based on the charts containing the results on important body members, for the legs, slip and fall from height and collision with things are more noticeable; for the hand, dealing with sharp levels, collision with things, falling objects, and slipping are more prevalent. For example, Figure 8 shows that the two cases of slip and collision with things are the same in both leg and hand cases.

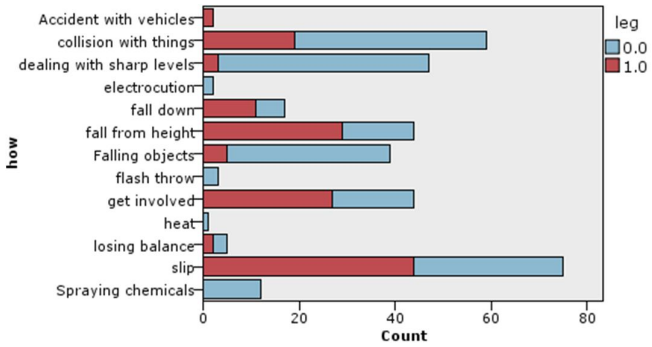


Fig. 8. How vs injured leg body member

After determining how the accidents occur, the root causes of these important events in the critical age group must be found. For this purpose, the chart showing how they occur based on the reason for the occurrence of Figure 9 is used.

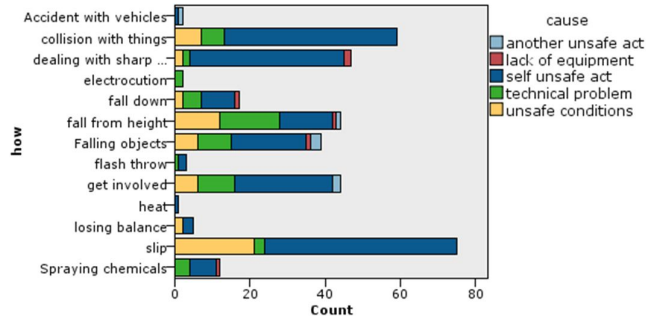


Fig. 9. How vs causes

This figure shows that self-unsafe action and unsafe conditions have been conducive to slipping, collision, and dealing with things, which constituted important accidents. Moreover, the technical problem is another root cause of the fall from height in the case of leg injury. Using the

distribution diagrams of the causes based on important results, it can be reassessed that the root causes, self-unsafe actions, and unsafe conditions are associated with major accidents, as previously identified. The distribution of causes based on an accident is shown in Figure 10. This figure illustrates the major contributing factors in the bruising results. Therefore, the root causes that have been found are right.

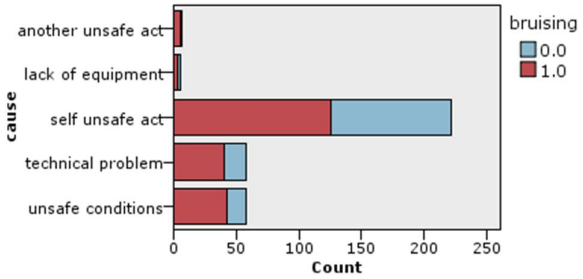


Fig. 10. Distribution causes based on bruising

According to the chart of the location of the accidents based on causes, important halls can be identified. At this stage, locations chart expresses the same interpretation either based on results or how, because they are related concepts. For this reason, in Figure 11, the column related to causes is used to clarify those halls where the worker's unsafe actions and conditions had made the greater role. According to this figure, Retouch Hall 2 and Color Halls 1 and 2 are associated with the greater number of causes than others. Therefore, these halls are noticeable for improvement.

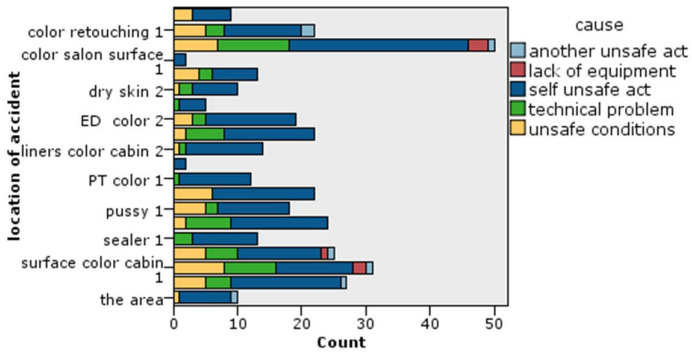


Fig. 11. Location of accidents distribution based on the causes

In addition, it should be determined what important halls are under the supervision and management of which parts. Figure 12 illustrates the share of management of Colors 1 and 2, especially 2. Color 2 consists of two parts of monitoring, which are related to the events of Paint Hall 2.

All the results have been obtained from graphs and visual aspect so far. Based on the Association Rules modeling (GRI) tools and the K-Mean algorithm, the accuracy of the results is tested. Parts of the rules derived from the GRI algorithm, which are sorted according to the lift index, are given in Table 5. Given the support, confidence, and lift index, the validity of the rules is approved.

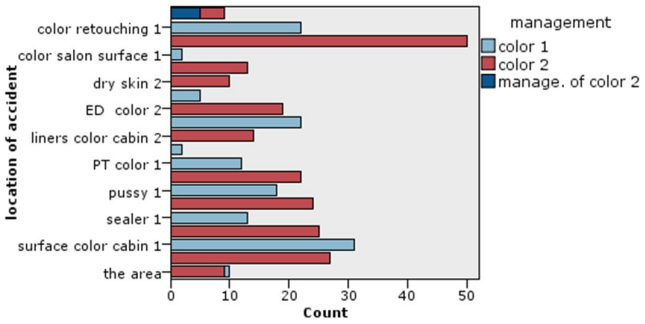


Fig. 12. Location of accidents distribution based on management

In this table, variables such as age, time, day of the month, month, year, day of the week, and important body members, foot and hand are selected as the input variables. Of course, Table 6 uses all the variables as inputs with body members and results (accidents) as the output variables, in which, for a more detailed analysis, separation sources in the software are used based on the elements of some columns such as the body members.

Tab. 5. Result of GRI with 7 variables

Rules	Support	Confidence	Lift
Working in the afternoon and leg slipping torsion.	5.43	73.68	6.79
Contact with sharp levels for hands injuries.	12.29	100	4.86
Contact with sharp levels with self-unsafe act cause for hands injuries.	10.86	100	4.86
Collision with sharp levels in very young age injuries.	9.71	100	4.86
Collision with sharp levels for workers under management by Color 2 injuries.	8	100	4.86

With the implementation of the algorithm, it is concluded that sensitive age groups are very young and young, often with little work experience and a diploma. Among the injured members, the members of the foot and hand in this important group have been prone to the most severe accidents. These important accidents have been associated with bruising, injury, and torsion due to self-unsafe act, unsafe conditions, and, in some cases, technical problems, with slipping and collision with things. The days lost result from the damage made to important body members, three months and more, highlighting the importance of focusing on this group. Locations or, in other words, important event halls include Retouching Hall 2 and Paint Halls 1 and 2 are under the management of Colors 1 and 2.

Tab. 6. Result of GRI with 20 variables

Rules	Support	Confidence	Lift
Collision with sharp levels for diploma workers with self-unsafe act injuries.	9.14	100	4.86
Very young worker with the cause of self-unsafe act in contact with sharp levels suffering from injuries.	8.57	100	4.86
In the half of the month with self-unsafe act, collision with sharp levels resulting in injuries.	5.43	100	4.86
Working in the morning, contact with sharp levels resulting in injuries.			
Low experience worker with very young age in contact with sharp levels injured.	9.43	96.97	4.71
Low experience worker with diploma, contact with sharp levels injured.	9.43	96.97	4.71
Worker with self-unsafe act in contact with sharp levels, not bruised.	11.71	100	2.63
In very young age because of self-unsafe act, contact with sharp levels resulting in hand injury.	8.57	100	2.32

One of the cluster analysis methods is the K-means algorithm. The explanation of and method for calculating its optimal cluster are presented in the research methodology section. To determine the optimal cluster, it is necessary to calculate the four introduced indexes. First, the total squared error ratio is calculated. Then, the minimum value of each index column is subtracted by all of its elements, and the resulting elements are divided by the maximum number of their column.

Thus, the values of the columns become dimensionless for decision-making. In addition, the values of the sum of squares of the error must be reversed. In other words, the elements of this column are deducted from number 1. The reason is that, unlike the rest, the lower value is attractive for this index. By reversal, the type of all indexes will be equal and, after calculating the sum of the values of the indexes of each cluster, it is possible to calculate the optimum cluster according to the maximum number obtained. The calculated pure indexes are shown in Table 7.

Tab. 7. Pure calculated indexes

Cluster num.	Dunn	Davis-Bouldin	SSE	Silhouette
One cluster	0	0	3.05E+06	0
Two clusters	0.1906	1.7361	1.14E+06	0.3542
Three clusters	0.2597	1.4356	7.26E+05	0.3892
Four clusters	0.2843	2.3027	4.11E+05	0.4414
Five clusters	0.3056	0.8957	2.54E+05	0.4475
Six clusters	0.4152	0.8944	1.02E+05	0.4776
Seven clusters	0.2886	0.9166	5.99E+04	0.4249
Eight clusters	0.2897	0.9211	3.96E+04	0.4344
Nine clusters	0.3108	0.9816	2.92E+04	0.4397
Ten clusters	0.3059	1.0802	2.03E+04	0.4428

The final stage of decision-making calculations for the optimal number of clusters is presented in Table 8.

Tab. 8. The last process on indexes for final decision-making

	Dunn	Davis-Bouldin	SSE	Silhouette	Sum of score
One cluster	0	0	0	0	0
Two clusters	0	0.402329	1	0	1.402329
Three clusters	0.307658	0.615707	0.283042	0.28363	1.490037
Four clusters	0.417186	0	0.472485	0.706645	1.596316
Five clusters	0.512021	0.999077	0.334957	0.756078	2.602133
Six clusters	1	1	0.929359	1	3.929359
Seven clusters	0.436331	0.984236	0.40949	0.572934	2.402991
Eight clusters'	0.441229	0.981041	0.214105	0.649919	2.286293
Nine clusters	0.535174	0.938081	0	0.692869	2.166124
Ten clusters	0.513357	0.868068	0.123628	0.71799	2.223043

Figure 13 shows the implementation result of this algorithm with respect to the 6 optimum clusters on the research data. This algorithm also confirms all the previous results by determining the importance all variables except the day of the week and time, which have not been highlighted in association rules and graphs. All the results are interrelated and justifiable. This makes the content more relevant once again. In addition, the results responded to research questions well.

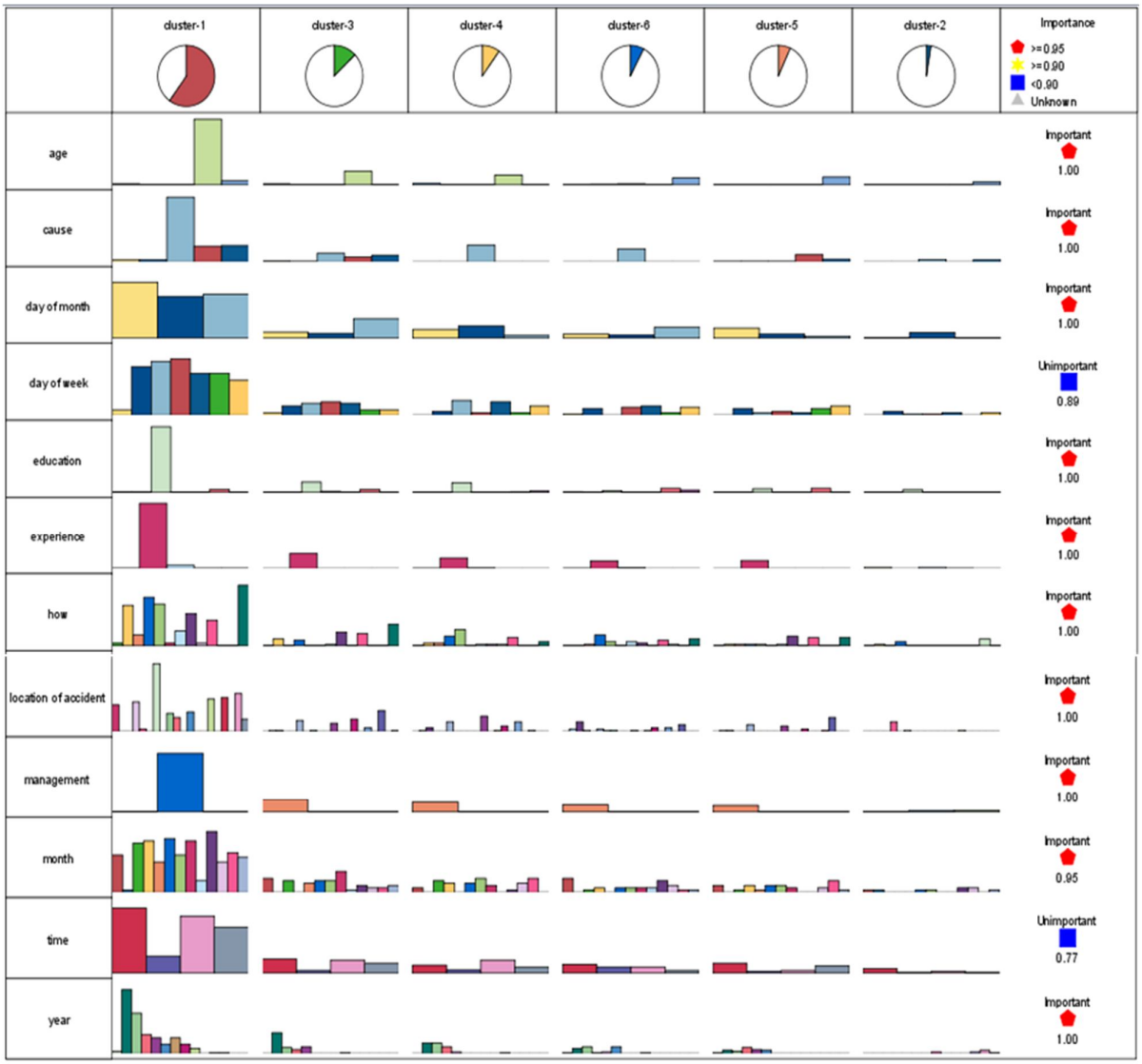


Fig. 13. K-means algorithm result

5. Discussion and Conclusion

Occupational safety is an important issue that has received less attention. This study was performed to identify effective factors in improving the safety of workers in the paint halls of one of the industrial units in Tehran province. The six-step methodology of CRISP in the Clementine software with this unit data collected during 14 years was implemented. The reasons for using data mining technique include the extraction of hidden knowledge from a significant amount of data with the least error, more accuracy, and the lowest level of human intervention.

In this regard, after explaining and implementing the steps of the CRISP approach, the results of various graphs, association rules (Table 5:6), and K-means algorithm (Figure 13) were presented. The results of all the three methods were in line with each other and helped determine effective factors such as age and educational groups, identify important events, both in terms of number and severity, and identify the most important location of the accidents; in addition, the root causes of major accidents were addressed in research questions. The importance of work safety is clear; however, despite its importance, the number of research in this area is still quite limited in scope. According to the dangerous accidents of industrial paint halls, the study aims to clarify the vision of improving safety using data mining tools. The results of this research can be generalized for all of the similar fields and methods.

For further research, it is suggested performing more investigation on the case studies and combination of the data mining approach and other methods, including methods for optimizing and implementing this technique with different

methods and comparing the results of those methods. The overall result of this research illustrates the importance of a very young and young age group, often with a diploma and low experience for whom important accidents occur such as slipping, injury, and torsion, which are due to self-unsafe act or unsafe conditions in the form of slipping or collision with things in most cases. Critically injured body members were determined in retouching and painting halls, which have major roles. These results help improve safety strategies.

Reference

[1] Han, J., Kamber, M., & Pei, J. Data mining: concepts and techniques, (the Morgan Kaufmann series in data management systems). (2006), pp. 230-240.

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. CRISP-DM 1.0 Step-by-step data mining guide (2000).

[3] Kantardzic, M. Data Mining: Concepts, Models, Methods, and Algorithms. Technometrics, Vol. 45, No. 3, (2003), p. 277.

[4] Agrawal, R., Imieliński, T., & Swami, A. Mining association rules between sets of items in large databases. In ACM sigmod record Vol. 22, No. 2, (1993), pp. 207-216. ACM.

[5] Cheng, C. W., Lin, C. C., & Leu, S. S. Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. Safety science, Vol. 48, No. 4, (2010), pp. 436-444.

[6] Cheng, C. W., Leu, S. S., Cheng, Y. M., Wu, T. C., & Lin, C. C. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accident Analysis & Prevention, Vol. 48, (2012), pp. 214-222.

[7] Nenonen, N. Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database. Applied Ergonomics, Vol. 44, No. 2, (2013), pp. 215-224.

[8] Kokotos, D. X., & Linardatos, D. S. An application of data mining tools for the study of shipping safety in restricted waters. Safety Science, Vol. 49, No. 2, (2011), pp. 192-197.

[9] Rivas, T., Paz, M., Martín, J. E., Matías, J. M., García, J. F., & Taboada, J. Explaining and predicting workplace accidents using data-mining techniques. Reliability Engineering & System Safety, Vol. 96, No. 7, (2011), pp. 739-747.

[10] Silva, J. F., & Jacinto, C. Finding occupational accident patterns in the extractive industry using a systematic data mining approach. Reliability Engineering & System Safety, Vol. 108, (2012), pp. 108-122.

[11] Cheng, C. W., Yao, H. Q., & Wu, T. C. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. Journal of Loss Prevention in the Process Industries, Vol. 26, No. 6, (2013), pp. 1269-1278.

[12] Maiti, J., Singh, A. K., Mandal, S., & Verma, A. Mining safety rules for derailments in a steel plant using correspondence analysis. Safety Science, Vol. 68, (2014), pp. 24-33.

[13] Verma, A., Khan, S. D., Maiti, J., & Krishna, O. B. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Safety Science, Vol. 70, (2014), pp. 89-98.

[14] Argilés-Bosch, J. M., Martí, J., Monllau, T., Garcia-Blandón, J., & Urgell, T. Empirical analysis of the incidence of accidents in the workplace on firms' financial performance. Safety science, Vol. 70, (2014), pp. 123-132.

[15] Hajakbari, M. S., & Minaei-Bidgoli, B. A new scoring system for assessing the risk of occupational accidents: A case study using data mining techniques with Iran's Ministry of Labor data. Journal of Loss Prevention in the Process Industries, Vol. 32, (2014), pp. 443-453.

[16] Sanmiquel, L., Rossell, J. M., & Vintró, C. Study of Spanish mining accidents using data mining techniques. Safety Science, Vol. 75, (2015), pp. 49-55.

[17] Murè, S., Comberti, L., & Demichela, M. How harsh work environments affect the occupational accident phenomenology? Risk assessment and decision making optimisation. Safety Science, Vol. 95, (2017), pp. 159-170.

[18] Sarkar, S., Verma, A., & Maiti, J. Prediction of Occupational Incidents Using Proactive and Reactive Data: A Data Mining Approach. In Industrial Safety Management (2018), pp. 65-79. Springer, Singapore.

[19] Liu, D., Chen, J., Li, S., & Cui, W. An integrated visualization framework to support whole-process management of water pipeline safety. Automation in Construction, Vol. 89, (2018), pp. 24-37.

[20] Das, S., Dutta, A., Jalayer, M., Bibeka, A., & Wu, L. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using 'Eclat'association rules to promote safety. International Journal of Transportation Science and Technology, Vol. 7, No. 2, (2018), pp. 114-123.

[21] Gregoriades, A., & Chrystodoulides, A. Extracting Traffic Safety Knowledge from Historical Accident Data. In Adjunct Proceedings of the 14th International Conference on Location Based Services (2018), pp. 109-114. ETH Zurich.

[22] Solanke, N. A., & Gotmare, A. D. Analysis of Roadway Traffic using Data Mining Techniques for providing safety Measures to Avoid Fatal Accidents. International Journal of Safety Research Science. Engg. Tech. Vol. 4, No. 4, (2018), pp. 348-355.

[23] Dunn, J. C. Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics, Vol. 4, No. 1, (1974), pp. 95-104.

[24] Davies, D. L., & Bouldin, D. W. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, No. 2, (1979), pp. 224-227.

[25] Ketchen, D. J., & Shook, C. L. The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal, Vol. 17, No. 6, (1996), pp. 441-458.

[26] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Vol. 20, (1987), pp. 53-65.

Follow This Article at The Following Site:

Khosrowabadi N, Ghousi R, Decision Support Approach on Occupational Safety Using Data Mining. IJIEPR. 2019; 30 (2) :149-164
URL: http://ijiepr.iust.ac.ir/article-1-836-en.html