TEHRAN UNIVERSITY OF MEDICAL SCIENCES

IJOH
INTERNATIONAL JOURNAL OF OCCUPATIONAL HYGIENE
eISSN: 2008-5435

IRANIAN OCCUPATIONAL HEALTH ASSOCIATION IOHA

ORIGINAL ARTICLE

Analysis of Occupational Injury and Forecasting the Number of Lost Days: A Machine Learning Approach

Elahe Jafari[1], Hamzeh Mohammadi[1], Majid Bayatian[1*], Maryam Behboudi[2], Davod Panahi[3]

[1]*Department of Occupational Health Engineering, Faculty of Health, Tehran Medical Sciences, Islamic Azad university, Tehran, Iran.*

[2]*Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran, Iran.*

[3]*Department of Occupational Health Engineering, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran.*

Received 2025-05-21; Revised 2025-07-24; Accepted 2025-08-17

ABSTRACT

Background: Work-related injuries data analysis helps the choice of measures to prevent accidents, therefore, the information and data of occupational injuries were analyzed to identify the overall trend of occupational injuries, and develop a forecasting model of lost workdays and provinces clustering.

Methods: To achieve the first goal, we calculated NFIR and FIR per 100000 workers and injury indices (including AFR, ASR, FSI, Safe T-Score, IR, MDR, and LTIR). To reach the second purpose, the FEE, FTE, FETE, and RE linear models and supervised machine learning alongside linear models (Random Forest, Extra trees, XG Boost, G Boost) were used. Finally, the AP clustering algorithm for provinces clustering, time series clustering (DTW method), and the KNN forecasting algorithm were applied. Data for 378826 occupational injuries, which occurred from 2001 to 2019, were extracted from the publications of the ISSO. Industries data were extracted from the Ministry of Industry publications.

Results: NFIR to FIR ratio ranged between 60 to 265.35 and injuries indices increase from 2001 to 2008 and then experienced a decline. In linear models, OLS and RE had the best performance in forecasting the loss days and the extra trees method had better performance as a blender than random forest method. The clustering results showed that Khuzestan, Markazi, Mazandaran, Qazvin, and Tehran provinces are in cluster 1 and other provinces are in cluster 2.

Conclusion: This study can be regarded to forecast occupational injuries and safety promotion planning.

KEYWORDS: *Analysis of Occupational Injury; Accident Data Capturing System; Lost Days; Machine Learning; Forecasting*

How to cite this article: Jafari E, Mohammadi H, Bayatian M, Behboudi M, Panahi D. Analysis of Occupational Injury and Forecasting the Number of Lost Days: A Machine Learning Approach. Int J Occup Hyg. 2025;17(3):138-153.

Corresponding author: *Majid Bayatian*
E-mail: *majid_bayatian@yahoo.com*

INTRODUCTION

Occupational injuries are incidents that occur in the course of work and result in a fatal or non-fatal injury [1]. These injuries are now one of major causes of death in the world and are considered as one of the most important health hazards in developed and developing societies [2, 3]. Occupational injuries have also become a major interest for countries and companies especially in the recent decade, due to the cost of such accidents and the subsequent loss of 4% of the Gross Domestic Product (GDP) worldwide [1, 4-6]. International Labor Organization (ILO) has reported that there are about 2.3 million deaths per year from work-related accidents and illnesses, equivalent to 6,000 deaths a day. Also, there are about 340 million work-related accidents and illnesses every year [4]. The WHO (World Health Organization) estimates that occupational injuries result in approximately 330,000 deaths per year per 2.7 billion workforces, and in the developing countries it tends to be higher [1, 7].

Occupational injuries are one of the most important consequences of globalization especially in developing countries [6-8]. Occupational health policies in many developed countries have led to the reduction of work-related injuries [9, 10]. Developing countries provide about 60% of global labors which 80% of them have hard and dangerous works [11].

Occupational injuries are leading to serious socioeconomic disadvantages such as disability, loss of working time, and increased health care [11]. Therefore, research on the causes of fatal and non-fatal occupational injuries is necessary to improve and promote preventive actions based on scientific evidence [10, 12]. A wide range of personal and occupational factors, such as age, gender, educational level, occupational status or lifestyles, lack of safety training and inadequate work experience, as well as personality, risky behaviors, smoking, workplace (proximal environmental conditions), the work environment (work organization and conditions) or even the social and political factors (employment or economic policies) are related to the risk of suffering a fatal or non-fatal occupational injury [7, 11, 12].

For occupational injury analysis, there exist several approaches, such as conventional techniques and modern data-driven learning approach. Epidemiological research of fatal and non-fatal occupational injuries becomes a priority, because it can help increase the knowledge of mechanisms and factors which cause the accidents and injuries, and to determine the effectiveness of the available control interventions and measures [12, 13]. Fatal and non-fatal occupational injury analysis is used to identify common factors contributing to occupational injuries and to improve accident prevention programs and decisions [3, 14]. Modern data-driven machine learning is divided into three sub groups, including supervised learning approaches (classification, regression), unsupervised learning approaches (clustering, dimension reduction, association rule mining, Bayesian network, neural network model, text mining and anomaly detection), and semi- supervised learning approaches (one-class SVM, like Primal Laplacian Support Vector Machine, S3VMs (Semi-Supervised Support Vector Machines) and TSVMs (Transductive SVMs)) [15]. Machine Learning (ML) methods are widely used in the prediction of occupational accident consequence[16].

The ML has been applied for different purposes in occupational injuries, with this playing a significant role in the choice of the utilized algorithms. For example, Tixier *et al.,* (2016) used the stochastic gradient tree boosting (STGB) to predict the type of energy involved in the occupational injuries, Poh *et al.,* (2018) used random forest (RF) method for the level of injuries severity; Tokdemir and Ayhan (2019) used artificial neural networks (ANNs) for occupational injuries consequence prediction [17], Duarte et al. (2019) reviewed occupational injuries in the mining industry; Samano-Rios et al. (2019) presented a review on occupational health interventions to protect workforces [15]; Choi et al., (2020) used a variety of algorithms for forecasting the likelihood of fatality[17].

This study was carried out in two phases. In the first phase, the study aim was to determine the pattern of occupational injuries among the workforce in Iran. For this purpose, data for 19 years (2001 - 2019) of fatal and non-fatal injury statistics were used to analyze the patterns and characteristics of occupational injuries that were published by the Iranian Social Security Organization (ISSO). The objective of this study in the second phase was to develop a forecasting model of lost workdays and regional clustering and scaling by machine learning approaches. The results of study can be intended to provide insight for prevention and control of the occupational-related fatal injuries.

MATERIALS AND METHODS

The study population included all insured workforce who were injured in 29 provinces of the industries in Iran between 2001 and 2019 years which are published annually by ISSO [18]. Also, the industry figures were taken from the Ministry of Industry Publications [19]. This study was carried out in two phases including descriptive analysis and statistical forecasting.

Descriptive Analysis

In this phase, Non-Fatal Injuries Rate (NFIR) and Fatal Injuries Rate (FIR) per 100000 workers and injury indices were calculated. These indices were included: Accident Frequency Rate (AFR) (Eq.1), Accident Severity Rate (ASR) (Eq.2), Frequency Severity Indicator (FSI), Eq. (3), Safe T-Score (Eq. 4) [10], Incident Rate (IR) (Eq. 5) [20], Mean Duration Rate (MDR) (Eq. 6) [21] and Lost-Time Injury Rate (LTIR) (Eq. 7) [22].

$$\text{AFR} = \frac{\text{Total number of injuries} \times 200000}{\text{Number of hours worked by all employees}} \quad (1)$$

Where 200000 (Ref value) is 100 (number of workforce) ´40 (number of working hours per week) ´50 (number of working weeks a year).

$$\text{ASR} = \frac{\text{Total loss times that are caused accident} \times 200000}{\text{Number of hours worked by all employees}} \quad (2)$$

$$\text{FSI} = \sqrt{\frac{\text{AFR} \times \text{ASR}}{1000}} \quad (3)$$

$$\text{Safe T. Score} = \frac{\text{AFR}_2 - \text{AFR}_1}{\sqrt{\frac{\text{AFR}_1}{\text{work hours of new year} / 200000}}} \quad (4)$$

Where AFR_2: accident frequency rate in this year; AFR_1: accident frequency rate in last year.

$$\text{IR} = \frac{\text{Number of Recordable Cases} \times 200000}{\text{Number of Employee labor hours worked}} \quad (5)$$

$$\text{MDR} = \frac{\text{No.of days lost as a result of x accidents}}{\text{x accidents}} \quad (6)$$

$$\text{LTIR} = \frac{\text{No.of Lost} - \text{time Injuries} \times 200000}{\text{total hoursed worked}} \quad (7)$$

Statistical Forecasting

Data Preprocessing

There was no missing value in all data. In supervised machine learning approaches, conventional methods have more accuracy using standardized data [23]. Therefore, the data were standardized in the form of Gaussian z-scores. This standardization causes the beta coefficients in regression models to be as standardized coefficients. Also, in supervised machine learning methods, the data were randomly divided with a ratio of 25 to 75 into training and test categories. Then, there were 303×2 and 101×2 arrays for training and test categories, respectively. For forecasting the number of lost days, the injuries data from 2006 to 2019 years for 29 provinces were used. Therefore, the database was included to observe 406 (29×14), and number of workers and injuries were independent variables and lost day was the dependent variable. Statistical analyses were carried out using R 4.1.3.

Linear Models (Panel data models)

In this section, the data were related to 14 years in 29 provinces, so the Pooled Ordinary Least Squares linear regression model (Pooled OLS) were not be able to identify year-specific and province-specific differences. Therefore, the Fixed Entity Effect (FEE), Fixed Time Effect (FTE), Fixed Time and Entity Effect (FETE) and Random Effect (RE) linear models were used to control the bias of the omitted variables.

In summary, the pooled OLS model was as follows:

$$y_{it} = \alpha + \beta_1 x_{1\,it} + \ldots + \beta_k x_{k\,it} + \epsilon\_\text{i} \quad (8)$$

where y_{it} is the dependent variable, x_{jit} the j^{th} explanatory variable (j=1, …, k), ϵ_{it} the error term, and $\hat{a}_1 \ldots\ldots \hat{a}_k$ are structural parameters coefficients [24, 25]. In FEE model, all coefficients can be different, therefore the model can be written as:

$$y_{it} = \beta_{0i} + \beta_{1i} x_{1\,it} + \ldots + \beta_{ki} x_{k\,it} + \epsilon\_\text{i} \quad (9)$$

This means that, i must be estimated as separate regressions. This model can be written as follows:

$$y_{it} = \beta_{0i} + \beta_1 x_{1\,it} + \ldots + \beta_k x_{k\,it} + \epsilon\_\text{i} \quad (10)$$

That is, slopes are considered the same. In this model, unite (entity) is considered as a dummy variable[24]. In the FTE model, the time is considered as dummy variable, which can be considered the features that are in all units but change over time. This model is expressed as follows:

$$y_{it} = \lambda_t + \beta_1 x_{1\,it} + \ldots + \beta_k x_{k\,it} + \epsilon\_\text{i} \quad (11)$$

where ϵ_{it} are independent and distributed with a zero mean and $\acute{o}_{\epsilon}^2$ variance [24]. In the RE model, the differences between individuals are random and follows a distribution with constant parameters. Therefore, the model may be written as:

$$y_{it} = \mu + \alpha_i + \beta_1 x_{1\,it} + \ldots + \beta_k x_{k\,it} + \epsilon\_\text{i} \quad (12)$$

where $\alpha_i \sim IID(0.\sigma_\epsilon^2)$, $\epsilon_{it} \sim IID(0.\sigma_\epsilon^2)$ and $\epsilon_{it} + \alpha_i$ are error term (α_i is constant over time). Also, these two components do not have correlation over time and are independent of x_{it} (for all i and t)[24].

Supervised machine learning

In this study, we used supervised machine learning alongside linear models. Supervised machine learning methods are more accurate alternatives in most cases compared to linear models although accuracy may not increase with supervised machine learning models. Using them alongside linear models can be as an exhaustive robustness check. The methods used in this section were decision trees (Random forest, Extra trees) and Gradient boosting (XG Boost, G Boost) models.

The Random Forest (RF) method is a popular classification algorithm, which can handle nonlinear data efficiently along with linear ones [26]. This method creates several random trees and merges them to have a more accurate and stable prediction [27]. In this method, instead of searching for the best feature in node splitting, it looks for the best feature inside a random subset. The maximum depth of each tree and the maximum number of features in a random tree are the two parameters that must be set.

The extra trees method is based on random attributing and selecting a cut-point while splitting a node. The randomization strength can be adjusted by selecting the appropriate parameters. These parameters include; the strength of the attribute selection process, the strength of averaging cut-point noise and strength of the variance reduction of the ensemble model aggregation [28].

Boosting is an ensemble meta-algorithm in machine learning method which is used for bias and variance reduction. This method is used in supervised learning as a way to convert weak learners to strong learners systems based on a combination of the results of different classifications [29, 30]. Boosting-based algorithms train weak learners repetitively and add to the previous set to correct previous predictions [31]. In this study, XG Boost and Stochastic Gradient Boost (SGB) methods were used. SGB method makes additive regression models by sequentially fitting a simple parametrized function to pseudo-residuals with least squares per iteration. By incorporating randomization in this method, both the approximation accuracy and the execution speed of the gradient amplification method can be significantly improved [32]. The XG Boost method is another algorithm that overcomes speed and scalability limitations[33, 34].

Integration of linear models and supervised machine learning methods

Statistical methods such as linear models are interpreted by the value coefficients, P-value and confidence intervals. While there are no such values in machine learning. Therefore, in this study, emulated feature importance method was used [35] which solve the interpretive gap between β coefficients and P-value in linear models and feature importance in machine learning methods with the following equation:

$$f_\upsilon = \frac{|\beta_\nu|(1-p_\nu)^\gamma}{\sum_{j=1}^{m}|\beta_j|(1-p_j)^\gamma} \quad (13)$$

where β is the model coefficients, p is the P-value and $\gamma \in \{1.2\}$ is an exponent. Using $\gamma = 2$ as an arbitrary choice, β coefficients with high P-value (have no statistical significance), obtain less feature. With this feature, supervised machine learning methods and linear models can be interpreted as complementary.

Stacking Generalization

The stacking generalization method was used as the final forecasting tool. This method uses the feature importance values of all the used methods to aggregate forecasting from all the used models into a set of meta-predictions [31]. This method stacks the predictions of all the used models as a level 0 in a new predictive model. The used models at level 0 are considered as an independent variable and days are a dependent variable. Then, a meta-learner (blender) considers the stacking model as level 1. This blender can be based on any linear model or supervised machine learning. In this study, the extra trees method is used as the blender. After level 1, the forecast model for days and feature importance for two independent variables (worker and accidents) were obtained.

Affinity Propagation method

This clustering algorithm is based on the concept of massage passing between data points. Unlike other methods, the Affinity Propagation (AP) method does not need to determine the number of clusters. This method considers the similarity between pairs of data points as input to the model[36].This algorithm proceeds with alternation between two steps of message transmission (massage-passing). This is done by updating two matrices, responsibility and availability, through which the exemplar in each cluster is determined. The responsibility matrix determines the well-suitedness of x_k as an exemplar for x_i compared to

other candidates, and the availability matrix shows how appropriate it is for x_i that x_k is chosen as its exemplar.

Time Series Clustering

Time series analysis is an important section of the study which can help identify the increasing or decreasing trends [37]. Also, clustering is an unsupervised learning approach whose main function is to group the data objects into a finite number of different clusters having strong similarities among the members within a group. Main function of clustering is not a prediction, but it can be used in building predictive modeling [15]. In this study, time series clustering was based on the Dynamic Time Wrapping (DTW) method, which measures the dissimilarities between the two time series. This algorithm compares the two time series and tries to find the optimum wrapping path between the two time series under certain constraints such as monotonicity [38]. Also, the window size constraint for the DTW method was the Sako-Chiba band [39].

KNN forecasting (K Nearest Neighbor)

This method is one of the prominent non-parametric techniques in classification and regression that has been used in time series forecasting. In this method, the k-Nearest Neighbors are used to forecast the next values in the time series. In this study, the FPTO-WNN algorithm was used [40].

RESULTS AND DISCUSSION

Descriptive Analysis

The trend of occupational injury and workforce number variation is shown in Figure 1. The trend of occupational injury number variation increased from 14114 (2001) to 24075 (2007) which are injuries lowest and highest rates, respectively. Then, the number of occupational injuries decreased slowly from 2007 to 2018 and increased from 2018 to 2019. Most of the researchers report that the number of occupational injuries increase with the increasing workforce number every year. For instance, Salguero et al reported that the number of occupational injuries experienced a decline by decreasing the workforce number from 2009 to 2012 in Spanish [41]. In this study, the percentage of occupational injuries increased by 215.23%, while the percentage of the workforce number increased by 129.39% from 2001 to 2019. This result shows that ascending trend of occupational injury numbers was less than the trend of the workforce, therefore this can be indicative of safety improvement in Iran (except for 2018-2019).

The average of the workforce total and social insured workforces were 24 (24043230) and 10 million (10564555), respectively [42]. Therefore, the 338826 occupational injuries for the mentioned insured workforce include half of the total workforce because many workplaces are not covered by the ISSO. This is the most important reason for the unreliability of the occupational injury statistics that causes these statistics not to be used for estimates of regional and global occupational injuries. Also, it cannot easily be accepted that the number of occupational injuries is doubled and injury does not change much. Since the survey of safety improvement cannot be correct based on the occupational injury number, therefore, we calculated Non-Fatal Injuries Rate (NFIR) and Fatal Injuries Rate (FIR) per 100000 workforces from 2001 to 2019 (Table 1). The NFIR and FIR highest were 336 in 2006 and 5 in 2019, respectively. Also, NFIR and FIR averages were 220.052 and 1.74 in 19 years, respectively, which are lower than other countries in the world. For instance, NFIR were 3600, 3400 and 3200 in Portugal, France,

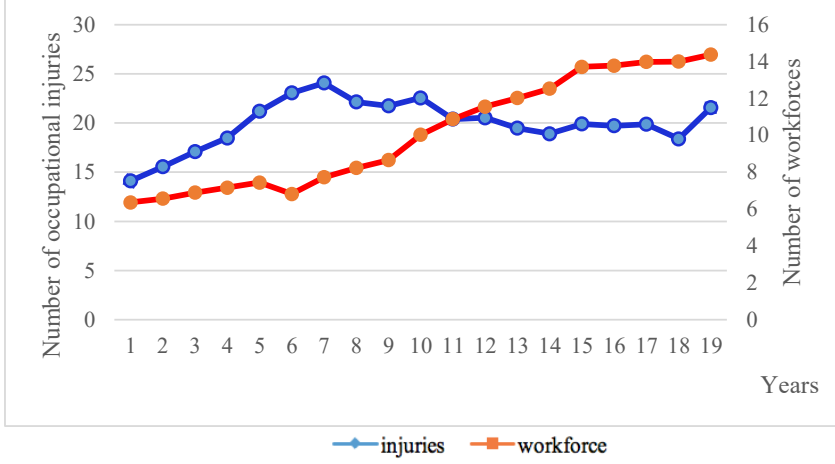

Figure 1. Trend of occupational injury (multiply by 1000) and workforce (multiply by 1000000) from 2001 to 2019

and Spain in 2014 years, respectively [43]. Bureau of Labor Statistics (BLS) reported FIR was 3.6 and 3.5 in 2016 and 2017 years, respectively [44], and HAS reported this rate was 2.5 in 2015 [45] that are high rates in comparison to the studies of all years. In the USA and Europe (2014) FIR were 11.2 and 13.4 in the industrial sector, respectively [46], that are higher than FIR in Iran. NFIR and FIR variation trends increased from 2001 to 2006, then decreased from 2006 to 2017 and increased from 2017 to 2019. These decreases have been reported in other studies. For instance, an FIR decrease has been reported from 2.1(1981) to 0.4 (2017) in England [47] or a 49% decrease from 2007 to 2016 in Australia [48] (in comparison with a 44% decrease in Iran).

Generally, there is a constant ratio between NFIR and FIR, if the occupational injury statistics record are correct. This ratio has not been constant in Iran and has ranged from 60 (2019) to 265.35 (2008) with a 175.86 average (Table 1). NFIR to FIR ratio was 850 in European Union countries in 2014 [43] and 138.8 in HAS report, which is very different from the results of this study. This ratio has a decreasing trend in various studies due to safety improvement, but it does not have a clear trend in Iran.

The FR, SR, FSI, IR, and LTIR indices increased from 2001 to 2006 and then decreased to 2017 (Table 2). FR and SR indices maximum were 0.355 and 16.19 in 2006, respectively. SR increase can be for two reasons: the number of injuries and magnitude of the injury's consequences underwent a spike. FR index is the highest in 2006; therefore, it can be said that the increase in occupational injuries rate is the reason for the high SR index. FR and SR indices study cannot be solely a good index for the safety survey, therefore FSI index was calculated for an accurate survey of occupational injuries. FSI minimum and maximum were 0.03 (2016 and 2017) and 0.0757 (2006). FR and SR indices were minimum and maximum values. IR maximum was 3.38 in 2006 which represented the highest number of occupational injuries per 1000 hours of work during the studied years. This index decreased from 2007 to 2017 because occupational injuries decreased and safety improved in Iran. The safe T-Score index shows safety

Table 1. NFIR and FIR per 100000 workforces from 2001 to 2019

Years	NFIR	FIR	NFIR to FIR Ratio
2001	220	1.6	137.5
2002	235	1.87	125.67
2003	247	1.41	175.18
2004	257	0.99	259.6
2005	284	1.34	211.94
2006	336	2.04	164.7
2007	310	1.51	205.3
2008	268	1.01	265.35
2009	250	1.27	196.85
2010	224	1.09	205.5
2011	187	0.78	239.74
2012	177	0.98	180.61
2013	161	0.97	165.98
2014	150	0.96	156.25
2015	144	0.79	182.28
2016	142	0.85	167.06
2017	141	0.84	168.21
2018	148	2.01	73.63
2019	300	5	60

Table 2. Occupational injury indices from 2001 to 2019

Years	FR	SR	FSI	Safe T- Score	IR	MDR	LTIR
2001	0.23	9.974	0.047	-5.15	2.2	42.68	79.79
2002	0.25	10.367	0.05	10.68	2.36	41.63	83.008
2003	0.26	11.164	0.054	5.25	2.48	42.74	89.31
2004	0.27	11.81	0.056	5.22	2.85	43.47	94.49
2005	0.3	12.53	0.061	15.74	2.85	41.8	100.29
2006	0.355	16.19	0.0757	23.84	3.38	45.52	129.56
2007	0.328	15.82	0.072	-9.39	3.11	48.22	126.57
2008	0.283	13.95	0.062	-24.96	2.69	49.27	111.6
2009	0.264	13.34	0.06	-11.12	2.51	50.46	106.76
2010	0.237	11.74	0.052	-12.42	2.25	49.58	93.95
2011	0.2	9.97	0.044	-26.92	1.87	50.48	79.75
2012	0.19	9.94	0.043	-7.4	1.77	53.15	79.55
2013	0.17	9.035	0.038	-15.9	1.62	52.98	72.28
2014	0.16	8.36	0.036	-8.58	1.51	52.58	66.86
2015	0.14	7.2	0.031	-18.51	1.45	49.65	57.67
2016	0.14	6.54	0.03	-1.94	1.43	45.76	52.38
2017	0.14	6.52	0.03	-0.98	1.42	45.88	52.17
2018	0.15	7.44	0.033	10.13	1.5	49.58	59.51
2019	0.305	15.12	0.068	14.79	3.05	49.58	121.01

level decreased from 2002 to 2006 years (Safe T-Score more than 3) and increased from 2007 to 2015 (Safe T-Score less than -3). Finally, safety level decreased from 2016 to 2019. MDR index increased from 42.68 (2001) to 53.15 (2012) and then decreased to 45.88 in 2017 and finally, increased to 49.58 in 2019. MDR increase (2001 to 2006) is reasonable because lost days' number increases by the increase of the number of occupational injuries. The reason for the increase in MDR could be due to the magnitude of the occupational injuries, resulting in an increase of lost days from 2006 to 2012. LTIR index increased from 79.79 (2001) to 129.56 (2006), then decreased to 52.17 in 2017. This increase (until 2006) and decrease (until 2017) in the trend of LTIR was due to the increase (until 2006) and decrease (until 2017) of occupational injuries.

Since occupational injury depends on number of workforce [41], the average of occupational injury rate was calculated in each province per 100000 workforce. In this case, Markazi (689.26), Qazvin (584.69), Zanjan (503.16) and Semnan (483.59) provinces had the highest occupational injury and Hormozgan (112.4) and Sistan and Baluchestan (74.58) provinces had the lowest occupational injury, respectively (Figure 2). Though in some research reports Tehran has the occupational injury highest[11, 13], in this study it was found that Tehran (164.11) province had the 22nd ranking.

Statistical Forecasting

For statistical forecasting, data (number of workers, occupational injury, loss days) related to 14 years (2006 to 2019) were selected in 29 provinces of Iran. The purpose was to forecast the loss days in the years 2020 and 2021 and to classify the situation of Iranian provinces in these two years. First, all 14-year data were standardized for the number of workers, occupational injury, and loss days. 75% of the data were considered as training data and 25% of the data as test data.

Loss days was considered as dependent variable and the number of workers and occupational injury were considered as independent variables, then, Pooled OLS, FTE, FEE, FTEE and RE linear models were implemented and their beta coefficients and P-value were calculated. All beta coefficients obtained were significant (P-value<0.05). They were used as emulated feature importance in equation 13. Also, supervised machine learning models were implemented and the feature significance results obtained for them were recorded. Table 3 shows linear models and feature importance of supervised machine learning models results; feature importance are probability vectors, these values indicate which variable each model assigns

more weight to. In Pooled OLS, RE and FTE methods, the occupational injury variable was more influential, while in other models, the number of worker's variable was more influential. The FETE and FEE methods assigned a heavy weight to the number of workers and a small weight to the occupational injury. Davoudi et al tested the performance of supervised machine learning algorithms for classification (support vector machines with linear, quadratic, and RBF kernels, Boosted Trees, and Naïve Bayes) in modeling and predicting occupational incidents severity in agribusiness industries. The results showed the significance of statistical analysis of occupational injury data in the workplace [49].

Table 4 shows accuracy statistics for each model in the training and test data (including: Mean Bias Error (MBE), Root Mean Square Error (RMSE) and r^2 method. In machine learning methods, the accuracy of the methods is based on test data, so models are compared according to the results of them in test data. The model is fitted ideally, if MBE and RMSE tend to zero and r^2 tend to 1. Also, the negative value of MBE means that the forecasted values are smaller than the observed values on average.

In linear models, OLS and RE had the best performance in loss days forecasting. The FTE model eliminates the provinces effect and had a similar performance to the OLS method in terms of predictive accuracy. FEE and FTEE models had similar results due to the entity effect and their accuracy was lower than other linear models. In supervised machine learning methods, extra trees and random forest methods had the highest accuracy and the G Boost method had the lowest accuracy. Figure 3 shows the performance of the methods used in loss days' forecasting in training and test data.

Then, Gini coefficients for inequality and Simpson's index (equation14) for concentration were calculated. Simpson's index is computed as follows:

$$\lambda = \sum_{i=1}^{N} fe_i^2 \qquad (14)$$

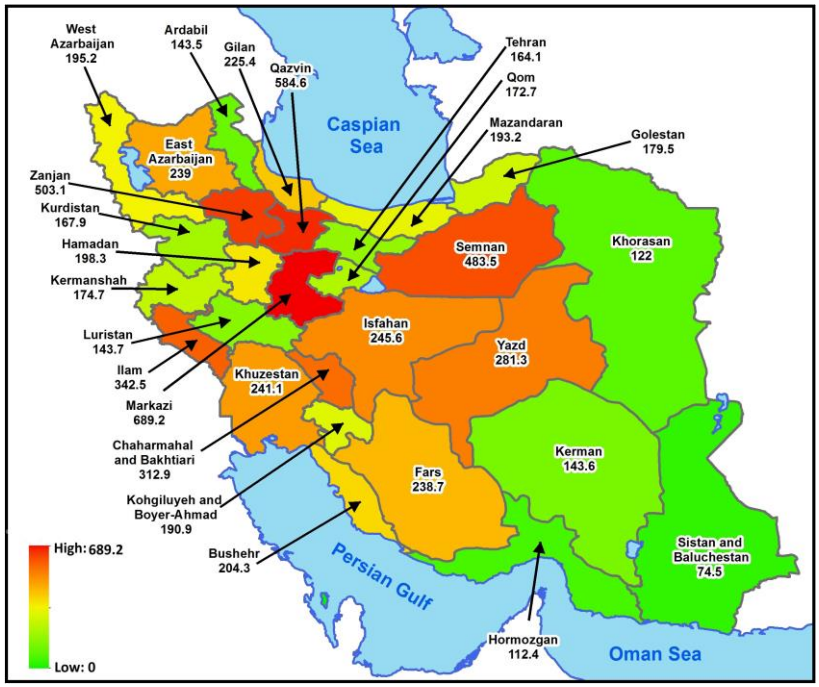

Figure 2. Occupational injury rate average in each province per 100000 workforces

Table 3. Emulated feature importance's for linear models and feature importance's of supervised machine learning models

Models	Pooled OLS	FEE	FTE	FETE	RE	Random Forest	Extra Trees	XG Boost	G Boost
Occupational injury	0.620	0.0001	0.646	0.004	0.526	0.476	0.456	0.471	0.336
Number of workers	0.380	0.9999	0.353	0.996	0.444	0.524	0.544	0.529	0.664

Table 4. MBE, RMSE, and r^2 are accuracy statistics for all used methods

Models	MBE	RMSE	r^2
Poold OLS (train)	-0.391	0.6429	0.3191
Poold OLS (test)	0.1488	0.6543	0.71476
FEE (train)	-3.3066×10^{-17}	0.057415	0.45699
FEE (test)	0.1869	0.725384	0.552112
FTE (train)	-7.578×10^{-17}	0.62125	0.36424
FTE (test)	0.13944	0.681096	0.605497
FTEE (train)	-1.5589×10^{-16}	0.54244	0.515311
FTEE (test)	0.16256	0.72526	0.511077
RE (train)	0.002606	0.06439	0.31731
RE (test)	0.151623	0.655179	0.72737
Extra trees (train)	$2.12968*10^{-17}$	0.29244	0.874827
Extra trees (test)	0.04235	0.561299	0.717047
Random Forest (train)	-0.00596	0.297186	0.874438
Random Forest (test)	0.075608	0.593824	0.692227
XG Boost (train)	0.00010146	0.05963877	0.9950311
XG Boost (test)	0.068167	0.652078	0.5693979
G Boost (train)	-0.005061	0.604691	0.398998
G Boost (test)	0.178425	0785802	0.4736499

where, fe_i is feature importance.

A value of 1 for these two indexes means that the model assigns all feature importance to one model. $\frac{1}{\lambda}$ shows variables number used by model. All methods except FEE and FETE methods require both workers number and occupational injury variables to forecast loss days and these two methods carried out forecasting only through the workers number

Because random forest and extra trees methods have more accuracy than the 9 methods performed[31, 35], these methods were selected separately as blenders. Zhu et al reported random forest method is an integrated algorithm and its accuracy is often higher than a single algorithm. Also, this method has high performance on the training set and test set[16]. In Kang et al study, the random forest method was applied to predict the occupational injury types of construction sites. They suggested a model to derive feature importances and verified the prediction correctness[50]. Figure 4 shows the results obtained from loss days forecasting or using stacking with these two methods.

As can be seen, the extra trees method had better performance as a blender than the random forest method. Table 6 shows fitting value error results in stacking with random forest and extra trees methods.

Stacking with random forest and extra trees methods had much better results than any of the 9 methods used in step 0. Also, the extra trees method had better performance than the random forest method because r^2 is higher (Figure 5). Then, the 2 × 9 F matrix was obtained by emerging all the feature importance's in step 0 and the 9 × 1 w matrix of the feature importance related to step 1 of the stacking method in the 9 models. The vector v was obtained with the dot product of these two matrices, ($v = F.W$). It is a 1 × 2 vector and shows the feature importance for the number of workers and occupational injury variables in the final model. Table 7 shows the Gini coefficient, the Simpson index and $\frac{1}{\lambda}$ in the stacking method in random forest and extra trees methods. The concentration is almost the same in the two methods, and both models use 4 variables to forecast, in other words, they are a 4-variable model. Kerim et al, surveyed four tree-based ensemble machine learning models: Random Forest, XGBoost, AdaBoost, and Extra Trees, as well as a state-of-the-art optimization method for hyper parameter tuning, Genetic Algorithm (GA) in construction projects. They reported GA-XG Boost had the highest prediction rate with 0.8292 in terms of accuracy[51].

because stacking with extra trees as a blender had a higher accuracy, extra trees were used as the final blender. Figure 6 shows the feature importance obtained from stacking by extra trees for the number of workers and occupational injury variables. The values in Fig 6 represent the v matrix. These values are the final aggregation for feature importance's. According to the results, the aggregated feature importance for the

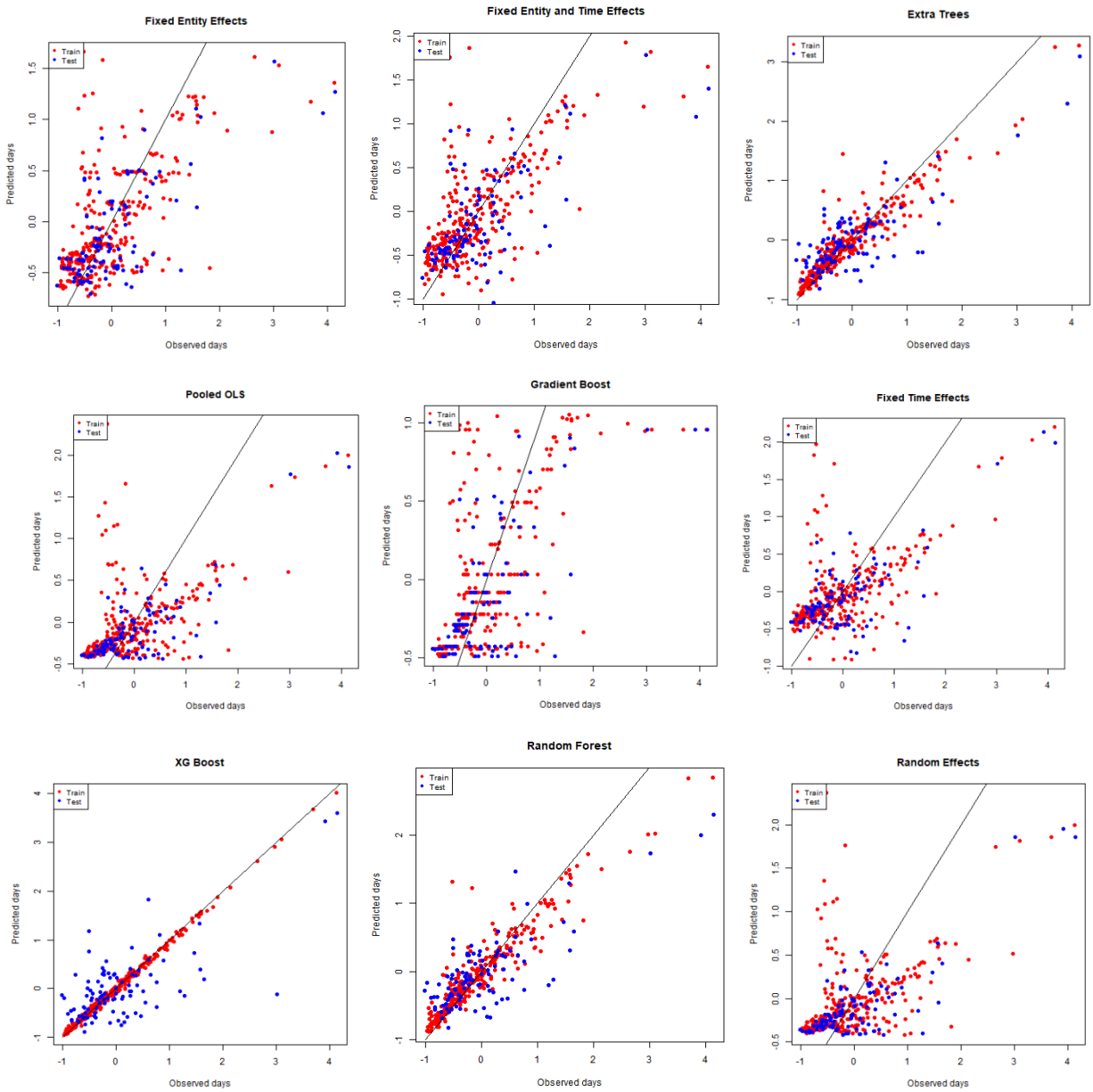

Figure 3. Fitted values vs actual data (training and test data). Red points: training data, Blue points: test data

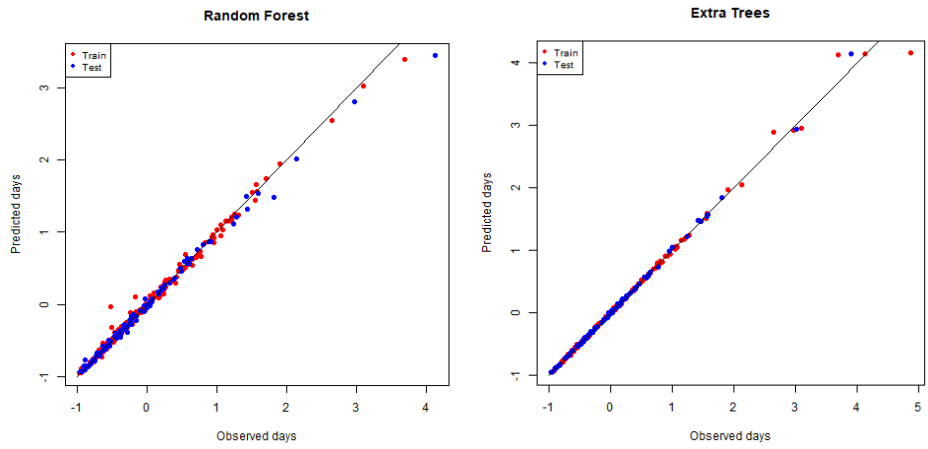

Figure 4. Fitted values vs actual data (training and test) in stacking using random forest and extra trees methods. Red points: training data, Blue points: test data.

number of workers and occupational injury variables were 0.59 and 0.41, respectively.

Accuracy-weighted data was calculated by re-scaling the data using the feature importance aggregation. This method provides more reliable prediction and classification than preliminary data[35]. Then, the provinces were clustered into two groups based on the accuracy-weighted number of worker data and accuracy-weighted occupational injury data, using the DTW method. This number of clusters was optimal. Figure 7 shows the time-series clustering of provinces using the DTW method. Each cluster included the number of workers and occupational injury variables time series that are separated by a vertical dotted line in the chart. In each cluster, the time series on the left related to the number of workers variable and on the right to the occupational injury variable. The results showed that Khuzestan, Markazi, Mazandaran, Qazvin, and Tehran

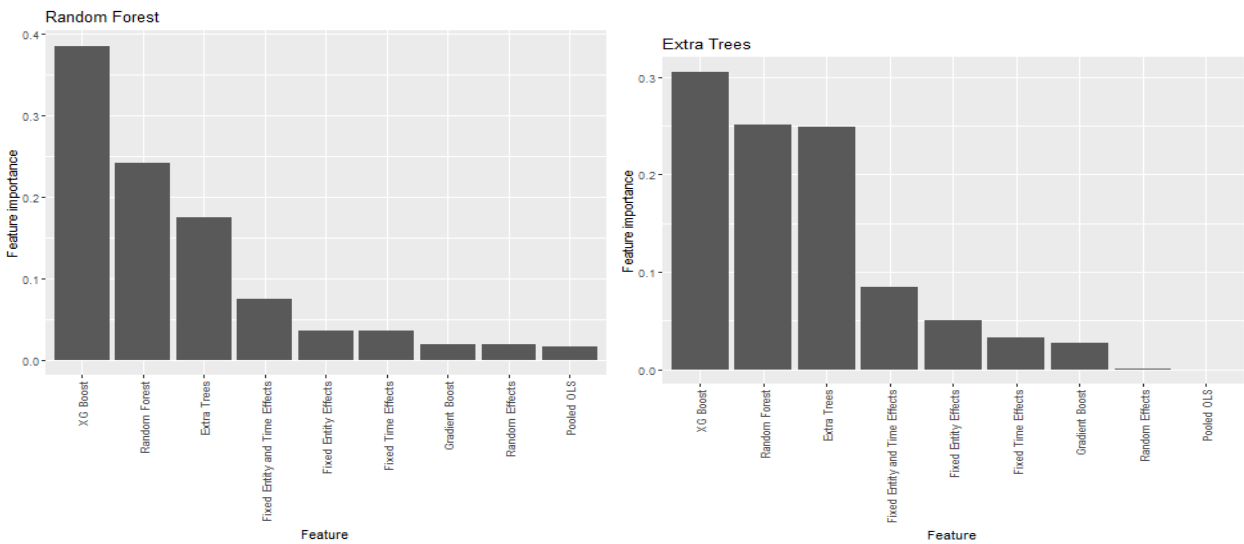

Figure 5. Feature importance's obtained from stacking by random forest and extra trees methods

Table 5. Gini coefficient, Simpson's index and variables number in used models

Diversity or concentration index	Random forest	Extra trees	XG Boost	G Boost	Pooled OLS	FEE	FTE	FETE	RE
Gini Coef.	0.941	0.9431	0.963	0.679	0.629	0.686	0.639	0.725	0.624
Simpson's index (λ)	0.501	0.504	0.502	0.554	0.529	0.999	0.543	0.991	0.501
$\frac{1}{\lambda}$	1.99	1.98	1.99	1.81	1.89	1	1.84	1.008	1.99

Table 6. Accuracy statistics in the stacking with extra trees and random forest methods

Model	MBE	RMSE	r^2
Stacking by random forest (train)	0.000569	0.0719966	0.99272
Stacking by random forest (test)	0.012212	0.088135	0.99338
Stacking by extra trees (train)	-7.547×10^{-17}	0.0027994	0.99872
Stacking by extra trees (test)	0.00155222	0.0405862	0.997966

Table 7. Gini coefficient, the Simpson index and 1/λ in the stacking method in random forest and extra trees methods

Diversity or concentration index	Stacking by Random forest	Stacking by Extra trees
Gini Coef.	0.9413	0.9431
Simpson's index (λ)	0.245	0.2295
$\frac{1}{\lambda}$	4.077	4.335

provinces are in cluster 1 and other provinces are in cluster 2. Kumar et al, proposed a framework to analyze road accident time series data in India. This framework segments the time series data into different clusters. They reported that the trend of road accidents was going to increase in certain clusters and those districts should be the prime concern to take control measures to overcome the road accidents[37].

Then, the number of worker and occupational injury variables were forecasted in each province by the KNN method for 2020 and 2021 years. The parameters used for the KNN method were considered for each province and each of the two variables separately using the FPTO-WNNS algorithm. For instance, Figure 6 shows the forecast in Alborz province. Sarkar et al, used KNN method for predicting and analyzing injury severity in the steel manufacturing plants. They reported that the classification algorithms indicated that occupational injury does not just occur at random, and there exists an underlying pattern, which can be explored by the machine learning approaches[52].

Next, the stacking by extra trees method was carried

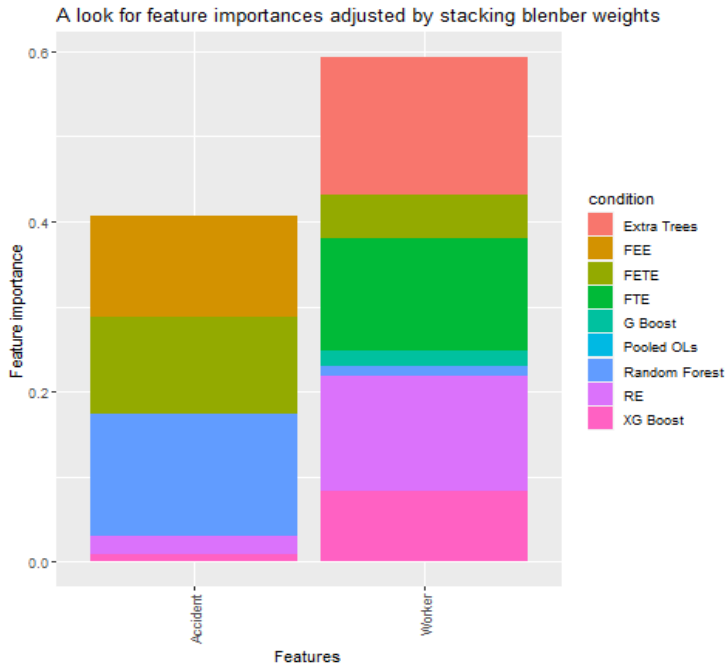

Figure 6. Feature importance's obtained from stacking by extra trees for the number of workers and occupational injury variables.

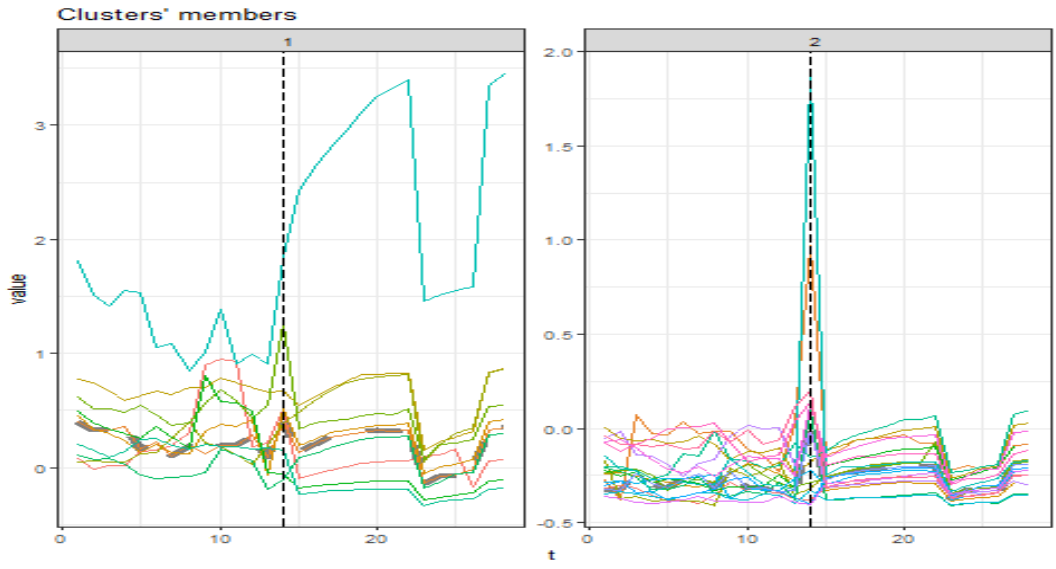

Figure 7. Time-series clustering of provinces according to number of worker and occupational injury during 14 years using DTW method

out for independent variables (number of workers and occupational injury) and dependent variable (lost days). In each of the two clusters of provinces with accuracy-weighted data of 14 years. Table 8 shows the results obtained for aggregated feature importance in these two clusters.

Because the aggregated feature importance obtained in the two clusters were significantly different, the data was rescaled with these weights once again. Then, using the forecasted number of workers and occupational injury results in the KNN method, the loss days in each province for 2020 and 2021 were forecasted (Figure 8). Since the number of workers and occupational injury, and loss days values were first standardized and then re-scaled several steps, the value of IR and SR indexes are in $(-\infty,\infty)$ interval. In these cases, smaller values of IR and SR indicate worse safety and higher values indicate a better safety condition.

Table 8. Aggregate feature importance's for number of workers and occupational injury in two provinces' clusters

Clusters	Occupational injury	Number of workers
1	0.43	0.57
2	0.17	0.83

Table 9. Clusters formed using affinity propagation algorithm

	2020		2021	
Clusters	provinces	exemplar	provinces	exemplar
1	Alborz	Alborz	Gilan, Kermanshah	Gilan
2	Ardabil, Chaharmahal and Bakhtiari, Fars, Golestan, Ilam, Isfahan, Kerman, Kermanshah, Khorasan, Kohgiluyeh and Boyer-Ahmad, Kurdistan, Luristan, Qom, Semnan, Sistan and Baluchestan, Tehran, Zanjan	Chaharmahal and Bakhtiari	Kerman	Kerman
3	East Azarbaijan, Hamadan, Hormozgan, Markazi	Hormozgan	Kurdistan, Bushehr, Qazvin, Zanjan, Semnan, Sistan and Baluchestan, Tehran, Ilam, Kurdistan, Khorasan, Kohgiluyeh and Boyer-Ahmad, Ardabil, Luristan, Golestan, East Azarbaijan, Fars, Hormozgan, Markazi, Alborz, Mazandaran, West Azarbaijan	Kurdistan
4	Khuzestan, Mazandaran	Khuzestan	Qom, , Chaharmahal and Bakhtiari,	Qom
5	Gilan, Qazvin, Yazd, Bushehr	Qazvin	Hamadan	Hamadan
6	West Azarbaijan	West Azarbaijan	Isfahan, Yazd	Isfahan

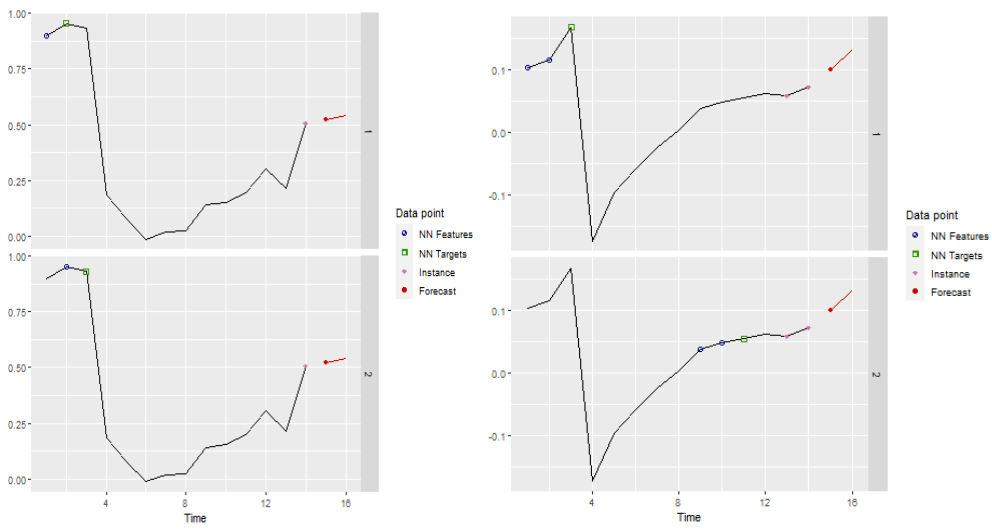

Figure 8. Forecast the values of number of workers and occupational injury variables in Alborz province in 2020 and 2021.

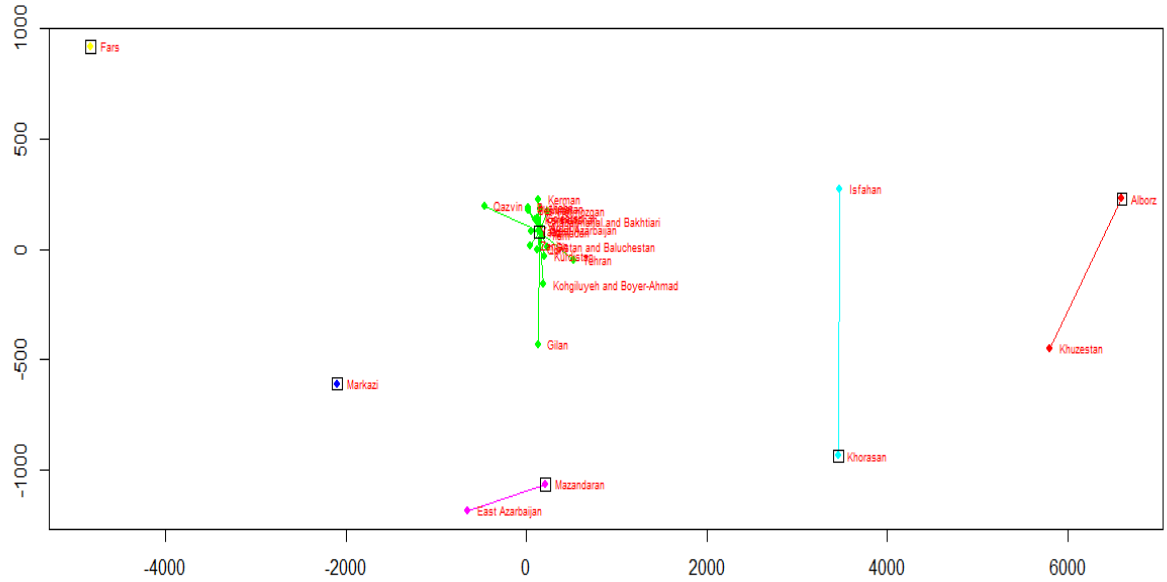

Figure 9. Six clusters designated and their exemplar in each cluster for the 2021

Finally, using the AP method, the provinces were clustered according to IR_{re} and SR_{re} indexes in 2020 and 2021. The cluster number was automatically determined by the algorithm. Table 9 shows the clustering and exemplar in each cluster in 2020 and 2021. Also, Figure 9 shows the clusters determined by this algorithm in 2021. In this case, Fars province has the worst safety condition among the other provinces and the Alborz and Khuzestan provinces have the best safety condition.

CONCLUSION

This study focused on occupational injury analysis of national data reported by the ISSO and was carried out in two phases 1. descriptive analysis and 2. statistical forecasting in in Iran. In the descriptive analysis, we calculated NFIR and FIR per 100000 workers and injury indices (including AFR, ASR, FSI, Safe T-Score, IR, MDR and LTIR). In statistical forecasting, first, the FEE, FTE, FETE and RE linear models were used to control the bias of the omitted variables. Also, we used supervised machine learning alongside linear models (Random forest, Extra trees, XG Boost, G Boost). Then, we applied the AP clustering algorithm for provinces clustering, time series clustering based on the DTW method and KNN forecasting for time series forecasting with the FPTO-WNN algorithm.

The first phase revealed that the ISSO injury data, covering only a portion of the workforce, contained a major deficiency in Iran's official statistics, a limitation that alone renders national injury statistics unreliable; these statistics did not include occupational injuries statistics for the white color workers, such as oil ministry workers, defense workers etc. This omission alone renders national injury statistics unreliable and undermines effective planning for reducing occupational injuries and their consequences, contributing to a continued increasing trend in injury indices. The second phase showed, we can cluster the provinces based on injury and workforce number and lost days and forecast them for the future years. Then, cluster the provinces based on the forecast results. In this regard, blender's method had better performance because it reduced errors and increased accuracy in forecasting.

Despite limitations in data quality, completeness, and accessibility, conventional data analysis can still play an important role in extracting and disseminating valuable information to relevant systems and stakeholders. Such analyses help identify priorities and inform preventive strategies; however, greater attention to data accessibility and integration is required. Accordingly, we recommend that occupational injury recording and publication involve multiple authorities, including Ministry of labor and social affairs, Ministry of Industry, Mine and Trade, Ministry of Oil, Ministry of Health and Medical Education, Ministry of Energy, Ministry of Roads & Urban Development, Iranian Traffic Police and Scientific Associations (such as Iranian Health Association, Iranian Occupational Health and Safety Association). Also, we suggest that in future works, statistical analysis and forecasting be done with Python software because there are more analytical parameters in this software.

Conflict of Interest
The authors declare no conflict of interest.

Funding
This study was not supported by any funding.

Author's Contributions
Study Conception or Design: M Bayatian, M Behboudi
Data Acquisition: M Bayatian, D Panahi
Data Analysis or Interpretation: E Jafari, M Behboudi
Manuscript Drafting: M Bayatian, H Mohammadi
Critical Manuscript Revision: M Bayatian, M Behboudi, D Panahi, E Jafari, H Mohammadi
All authors have approved the final manuscript and are responsible for all aspects of the work.

AI Statement
The authors confirm that no AI tools or services were used during the preparation of this work.

REFERENCES

1. Al-Abdallat EM, Oqailan AMA, Al Ali R, Hudaib AA, Salameh GAM. Occupational fatalities in Jordan. J Forensic Leg Med. 2015;29:25-9.
2. Dorman P. The economics of safety, health, and well-being at work: an overview. Geneva: ILO; 2000.
3. Alizadeh SS, Mortazavi SB, Sepehri MM. Analysis of occupational accident fatalities and injuries among male group in Iran between 2008 and 2012. Iran Red Crescent Med J. 2015;17(10):e18975.
4. International Labour Organization. Safety and health at work. Geneva: ILO; 2018.
5. Rohani JM, Johari MF, Hamid WHW, Atan H, Adeyemi AJ, Udin A. Occupational Accident Direct Cost Model Validation Using Confirmatory Factor Analysis. Procedia Manuf. 2015;2:286-90.
6. Asady H, Yaseri M, Hosseini M, Zarif-Yeganeh M, Yousefifard M, Haghshenas M, et al. Risk factors of fatal occupational accidents in Iran. Ann Occup Environ Med. 2018;30(1):29.
7. Rahmani A, Khadem M, Madreseh E, Aghaei HA, Raei M, Karchani M. Descriptive Study of Occupational Accidents and their Causes among Electricity Distribution Company Workers at an Eight-year Period in Iran. Saf Health Work. 2013;4(3):160-5.
8. Mehrdad R, Seifmanesh S, Chavoshi F, Aminian O, Izadi N. Epidemiology of occupational accidents in Iran based on social security organization database. Iran Red Crescent Med J. 2014;16(1):e10359.
9. Kim Y, Park J, Park M. Creating a Culture of Prevention in Occupational Safety and Health Practice. Saf Health Work. 2016;7(2):89-96.
10. Hamidi N, Omidvari M, Meftahi M. The effect of integrated management system on safety and productivity indices: Case study; Iranian cement industries. Saf Sci. 2012;50(5):1180-9.
11. Bakhtiyari M, Delpisheh A, Riahi SM, Latifi A, Zayeri F, Salehi M, et al. Epidemiology of occupational accidents among Iranian insured workers. Saf Sci. 2012;50(7):1480-4.
12. Villanueva V, Garcia AM. Individual and occupational factors related to fatal occupational injuries: A case-control study. Accid Anal Prev. 2011;43(1):123-7.
13. Vahabi N, Kazemnejad A, Datta S. Empirical Bayesian Geographical Mapping of Occupational Accidents among Iranian Workers. Arch Iran Med. 2017;20(5):298-304.
14. Lin YH, Chen CY, Luo JL. Gender and age distribution of occupational fatalities in Taiwan. Accid Anal Prev. 2008;40(4):1604-10.
15. Sarkar S, Maiti J. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. Saf Sci. 2020;131:104900.
16. Zhu R, Hu X, Hou J, Li X. Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Saf Environ Prot. 2021;145:293-302.
17. Shayboun M, Kifokeris D, Koch C. Machine Learning for Analysis of Occupational Accidents Registration Data. In: Proceedings of the 36th Annual ARCOM Conference; 2020 Sep 7-8; Leeds, UK. p. 618-27.
18. Iranian Social Security Organization. Statistics and information [Internet]. Tehran: ISSO; [cited 2026 Jun 4]. Available from: https://tamin.ir
19. Ministry of Industry, Mine and Trade. Statistics and information [Internet]. Tehran: MIMT; [cited 2026 Jun 4]. Available from: https://mimt.gov.ir
20. Occupational Safety and Health Administration. OSHA statistics. Washington (DC): OSHA; 2014.
21. Channing J. Safety at work. 7th ed. London: Routledge; 2013.
22. International Association of Drilling Contractors. Incident statistics program. Houston (TX): IADC; 2018.
23. Müller AC, Guido S. Introduction to machine learning with Python: a guide for data scientists. Sebastopol (CA): O'Reilly Media; 2016.
24. Verbeek M. A guide to modern econometrics. 5th ed. Hoboken (NJ): John Wiley & Sons; 2017.
25. Wooldridge JM. Introductory econometrics: A modern approach. 6th ed. Boston (MA): Cengage Learning; 2015.
26. Sarkar S, Chain M, Nayak S, Maiti J. Decision support system for prediction of occupational accident: a case study from a steel plant. In: Verma N, Ghosh AK, editors. Emerging technologies in data mining and information security. Singapore: Springer; 2019. p. 787-96.
27. Majumder M, Bhattacharyya S. An alternate arrangement of geofoam blocks and air pocket to mitigate confined blast induced vibration. Int J Geotech Eng. 2021;15(1):52-65.
28. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3-42.
29. Schaal S, Atkeson C. From isolation to cooperation: An alternative view of a system of experts. In: Touretzky D, Mozer M, Hasselmo M, editors. Advances in Neural Information Processing Systems 8 (NIPS 1995). Cambridge (MA): MIT Press; 1995. p. 605-11.
30. Breiman L. Bias, variance, and arcing classifiers. Berkeley (CA): Statistics Department, University of California; 1996. Technical Report No. 460.
31. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2nd ed. Sebastopol (CA): O'Reilly Media; 2019.
32. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367-78.
33. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785-94.
34. Zhou Y, Li T, Shi J, Qian Z. A CEEMDAN and XGBOOST-based approach to forecast crude oil prices. Complexity. 2019;2019:4391675.
35. Chen JM, Zovko M, Šimurina N, Zovko V. Fear in a Handful of Dust: The Epidemiological, Environmental and Economic Drivers of Death by PM2.5 Pollution. Int J Environ Res Public Health. 2021;18(16):8688.
36. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315(5814):972-6.
37. Kumar S, Toshniwal D. A novel framework to analyze road accident time series data. J Big Data. 2016;3(1):1-11.
38. Sardá-Espinosa A. Comparing time-series clustering algorithms in R using the dtwclust package. R Package Vignette. 2017;12:41.
39. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process. 1978;26(1):43-9.
40. Tajmouati S, Wahbi BE, Bedoui A, Abarda A, Dakkoun M. Applying k-nearest neighbors to time series forecasting: two new approaches. arXiv [Preprint]. 2021:arXiv:2103.14200.
41. Salguero-Caparros F, Suarez-Cebador M, Rubio-Romero JC.

Analysis of investigation reports on occupational accidents. Saf Sci. 2015;72:329-36.
42. Statistical Center of Iran. Iranian labor force survey report 2005-2017. Tehran: SCI; 2018.
43. Eurostat. Accidents at work statistics [Internet]. Luxembourg: European Commission; 2015 [cited 2026 Jun 4]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Accidents_at_work_statistics
44. Bureau of Labor Statistics. National census of fatal occupational injuries in 2017. Washington (DC): BLS; 2017.
45. Health and Safety Authority. Summary of Workplace Injury, Illness and Fatality Statistics 2014-2015. Dublin: HSA; 2016.
46. Hämäläinen P. Global estimates of occupational accidents and work-related illnesses 2017. Espoo: Finnish Institute of Occupational Health; 2017.
47. Health and Safety Executive. Fatal injuries arising from accidents at work in Great Britain 2017. Bootle: HSE; 2017.
48. Safe Work Australia. Key Work Health and Safety Statistics Australia 2017. Canberra: SWA; 2017.
49. Kakhki FD, Freeman SA, Mosher GA. Evaluating machine learning performance in predicting injury severity in agribusiness industries. Saf Sci. 2019;117:257-62.
50. Kang K, Ryu H. Predicting types of occupational accidents at construction sites in Korea using random forest model. Saf Sci. 2019;120:226-36.
51. Koc K, Ekmekcioğlu Ö, Gurgun AP. Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. Autom Constr. 2021;131:103896.
52. Sarkar S, Pramanik A, Maiti J, Reniers G. Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. Saf Sci. 2020;125:104616.