NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2023; 12(3), 778-798

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Predicting the severity of occupational accidents in the construction industry using standard and regularized logistic regression models

İnşaat sektöründe standart ve düzenlileştirilmiş lojistik regresyon modelleri kullanılarak iş kazalarının şiddetinin tahmini

Şura Toptancı[1,*], Nihal Erginel[2], Ilgın Acar[3]

[1,2] *Eskişehir Technical University, Industrial Engineering Department, 26555, Eskişehir, Türkiye*
[3] *Western Michigan University, Industrial and Entrepreneurial Engineering and Engineering Management Department, Michigan, USA*

Abstract

Occupational accidents in the construction industry occur more frequently when compared with other industries. Construction occupational accidents still have not been prevented at the desired level. Several studies in the literature have been conducted to predict the occurrence frequency of these accidents using classical statistical and machine-learning techniques. However, some challenges regarding imbalanced and multicollinearity problems present in the dataset are not considered while analyzing data with a large size and a large number of categorical variables. This study aims to predict the severity of non-fatal construction accidents considering mentioned challenges to obtain more accurate results. In this study, standard binary logistic regression, Firth, Ridge, Lasso, and Elastic Net Regularized logistic regression models were used for the prediction of lost workdays in the construction industry and results were compared. The data used were classified into five groups: victim, workplace, accident time, accident and sequence of events, and post-accident state-related variables. The results showed that Firth's logistic model is the best-performing model and age, education, vocational education, workplace size, project type, working environment, accident month and year, general and specific activities, material agent, type of injury, and part of body injured are the most significant variables. This study, by providing interpretable machine learning tools, is the first attempt to use proposed models in the area of construction safety in the literature.

Keywords: Occupational accidents, Construction industry, Logistic regression, Machine learning, Accident severity.

Özet

İnşaat sektöründe iş kazaları diğer sektörlere kıyasla daha sık meydana gelmektedir. İnşaat iş kazaları hâlâ istenilen düzeyde önlenememiştir. Literatürde klasik istatistiksel ve makine öğrenmesi teknikleri kullanılarak bu kazaların meydana gelme sıklığını tahmin etmek için birçok çalışma yapılmaktadır. Ancak, büyük boyutlu ve çok sayıda kategorik değişken içeren veriler analiz edilirken, veri setinde bulunan dengesizlik ve çoklu bağlantı sorunlarına ilişkin bazı problemler dikkate alınmamaktadır. Bu çalışma daha doğru sonuçlar elde edebilmek için bahsedilen problemleri dikkate alarak, ölümcül olmayan inşaat kazalarının şiddetini tahmin etmeyi amaçlamaktadır. Bu çalışmada, inşaat sektöründe iş günü kaybının tahmini için standart ikili lojistik regresyon, Firth, Ridge, Lasso ve Elastik Net düzenlileştirilmiş lojistik regresyon modelleri kullanılmış ve sonuçlar karşılaştırılmıştır. Kullanılan veriler kazazede, iş yeri, kaza zamanı, kaza ve olaylar zinciri ve kaza sonrası durumla ilgili değişkenler olmak üzere beş gruba ayrılmıştır. Sonuçlar, Firth'in lojistik modelinin en iyi performans gösteren model olduğunu ve yaş, eğitim, mesleki eğitim, işyeri büyüklüğü, proje türü, çalışılan ortam, kaza ayı ve yılı, genel ve özel faaliyetler, kullanılan materyal, yaranın türü ve yaranın vücuttaki yerinin en önemli değişkenler olduğunu göstermiştir. Yorumlanabilir makine öğrenimi araçları sağlayan bu çalışma, literatürde inşaat güvenliği alanında önerilen modelleri kullanmaya yönelik ilk girişimdir.

Anahtar kelimeler: İş kazaları, İnşaat sektörü, Lojistik regresyon, Makine öğrenmesi, Kaza şiddeti.

1 Introduction

Occupational accidents are complex and serious health and safety problems of the working life. Millions of people suffer from injuries and fatalities resulting from occupational accidents in the workplaces every year. Not only do these incidents impact the workers' health, but they also create significant burdens for their families, employers, and society as a whole [1], [2]. In addition, occupational accidents may result in absenteeism from work, loss of income, loss of job, and time loss due to disabling injuries and medical check-ups after the injured worker returns to work [3], [4].

According to the estimation model of the International Labour Organization (ILO), there are approximately 3.5 billion workers in the world [5]. Unfortunately, these workers are at risk of experiencing occupational accidents and diseases, resulting in 2.78 million deaths and 374 million

* Sorumlu yazar / Corresponding author, e-posta / e-mail: sani@eskisehir.edu.tr (Ş. Toptancı)
Geliş / Recieved: 30.11.2022 Kabul / Accepted: 22.05.2023 Yayımlanma / Published: 15.07.2023
doi: 10.28948/ngmuh.1212385

non-fatal accidents that cause more than four days of work missed each year [6]. Occupational accidents also lead to significant economic costs including long-term workday loss, safety corrections, medical treatment, survivor benefits, death-related costs, fines, and numerous indirect costs to the employers for occupational accidents are much and assorted. The economic cost of poor occupational safety and health (OSH) practices is estimated to be 3.94% of the global Gross Domestic Product (GDP) annually [6], [7].

Despite the fact that annual surveys of occupational accidents in developing countries state that there is a decrease taking place in the incidence rate for occupational accidents, some problems may occur (i.e., identifying, recording, and reporting) in surveys and observation programs. Thus, this annual survey cannot show the rates precisely for occupational injuries and illnesses to evaluate the load of occupational accidents of the Nations correctly. There can be also other discussions for this decrement such as the results of surveys may have statistical artifacts or the survey may be completed in the higher unemployment times, and or shutdown period due to the epidemic. Besides, this decrement mentioned is not valid for every year and in general, the trend in the rates of injury and illness in workplaces and their costs is still upward. For these reasons, it is necessary to establish an OSH system in workplaces and carry out OSH studies and accident prevention activities in a systematic way in order to eliminate or reduce occupational accidents and their effects.

In order to direct OSH-related activities in workplaces, occupational accident data should be analyzed. It is generally handled in terms of frequency and severity of the occupational accidents which indicates the number of incidents occurring in a given working period and expresses the effect on people when the accident occurs, respectively. In literature, accident frequency estimation is widely studied to accident severity. Besides, occupational accidents may result from various causes or variables. For this reason, much research has also been conducted to understand the etiologic mechanisms of occupational accidents, with most analyzing accident data to identify variables that may trigger them or to investigate the relationship between dependent and a limited number of independent accident variables. Therefore, it is important to analyze occupational accident data for learning overlapping characteristics of accidents, predicting future events, and reducing the frequency and severity of injuries.

One industry that experiences a high frequency and severity of occupational accidents is the construction industry. In comparison to other industries globally, construction activities pose unique risks and have high rates of fatal and non-fatal injuries resulting from occupational accidents.

According to the Bureau of Labor Statistics (BLS), in 2021, 951 of 5.190 fatal occupational accidents in the USA and 386 of 1.382 fatal occupational accidents in Türkiye occurred in the construction industry [8], [9]. Despite the current legislation (Occupational Health and Safety Law No. 6331 made in 2012 and then many regulations and communique gradually published under this law) that applies to OSH, numerous safety-related actions carried out, fines applied, and precautions taken to improve safety and reduce workplace accident rates in Türkiye, the number, and severity of accidents in the construction industry have not been reduced to the desired level. In fact, the number of these accidents in the construction industry in Türkiye has exhibited an increasing trend every year. According to the report of the Social Security Institution (SSI), 10.08% of the working group under the framework of article 4/1.a of Law No. 5510 were employed in the construction industry in 2021. Data from the same year indicate that occupational accidents that occurred in the construction industry constituted 11.37% of all occupational accidents, 27.93% of all fatal accidents, and 23.53% of all accidents that cause permanent incapacity in Türkiye. These findings indicate that the construction industry is where accidents occur most frequently.

In literature, there are two main sources of data used to analyze occupational accidents which are government statistics and empirical data from organizations. Besides, government statistics that have high-dimensional database is less used in accident studies. Moreover, researchers commonly use traditional statistical approaches such as frequency analysis and standard regression models to analyze occupational accident data with small sample sizes. Cakan [10] used logistic regression models to study the effects of a few accident variables on fatal and non-fatal construction fall accidents by examining case reports obtained from the Occupational Safety and Health Administration (OSHA) based on the degree of injury. Onder [11] analyzed non-fatal occupational accidents in an open-pit mining operation using logistic regression method to estimate the workday loss. Akboga [12] assessed the risk factors affecting injury severity scores in construction accident reports obtained from the General Directorates of Social Security Institution (SSI) for three metropolitan cities using logistic regression. Bilim [13] used cross-tabulation and logistic regression analyses for highway and railway construction accidents that occurred from 2013 to 2016.

Nevertheless, a comprehensive database is necessary to investigate occupational accidents and learn the links among variables addressed in the study. However, traditional statistical modelling strategies may not accurately analyze and interpret high-dimensional accident databases. Recently, machine learning and data mining techniques have been used with large data sets. These approaches are used to predict accident outcomes, learn from the data, and take actions related to the variables that make that prediction to avoid recurrent accidents. Table 1 outlines some of the prominent studies that have analyzed construction accident data using machine learning and data mining techniques.

Moreover, in general, the accident data include mostly rare events, many independent variables, and mostly categorical variables with many levels. Furthermore, highly imbalanced distribution, and multicollinearity which is a situation in which some independent variables are too similar to one another and highly correlated also exist among accident variables. All these issues may arise a quasi-

Table 1. Some prominent studies on analyzing construction accident data

Author (year)	Application Field (data period)	Technique(s)	Prediction purpose
Tixier et al. [14]	Construction (2011-2014)	Random Forest and Stochastic Gradient Tree Boosting	Injury type, energy type, body part, and injury severity
Yang et al. [15]	Construction (2015)	Support Vector Machine	Near miss falls
Kang and Ryu [16]	Construction (2008-2014)	Random Forest	Occupational accident types
Ayhan and Tokdemir [17]	Construction	Clustering, Artificial Neural Networks, Case-Based Reasoning	Accident outcome
Lee et al. [18]	Construction (2015-2020)	Clustering, Cramer's V, Chi-square test, Support Vector Machine, Principal Comp. Analysis	Injury severity level
Choi et al. [19]	Construction (2011-2016)	Logistic regression, Decision tree, Random Forest, Adaptive boosting	Risk of fatality accidents
Recal and Demirel [20]	Construction (2013-2016)	Support Vector Machines, Multinomial logistic, C5.0 decision tree, Stochastic Gradient Boosting, Neural Network	Fatal and non-fatal accidents as two-class and multi-class outcomes
Tetik et al. [21]	Construction (2010-2012)	C5.0 Decision tree algorithm	Determination of associations between the degree of injury and several variables
Koc et al. [22]	Construction (2011-2020)	Random Forest, Naïve Bayes, K-Nearest neighbor, Neural Networks	Fatal vs. non-fatal accident in İstanbul

complete separation problem which occurs when some observations of independent (explanatory) accident variables with a dependent (outcome) variable have values of zero (0) and causes one or some of the independent variables can perfectly or nearly perfectly predict the dependent accident variable [23-25]. In the presence of a quasi-complete separation problem, the model may fail to converge, predictions become biased, inaccurate prediction results are obtained, the standard errors can be very large values, and one or more regression coefficients become infinite. In this case, standard regression models are not suitable to fit the accident data when there are above-mentioned problems since estimated parameters become unstable and the fitting performances of the models are reduced in standard models. Therefore, imbalanced, and multicollinearity issues should be carefully considered before estimating regression parameters. Hoewever, machine learning algorithms consider a balanced assumption in a data set, and they are focused on increasing accuracy. Under these circumstances, integrating some alternative techniques and pre-processing (i.e., handling missing values, choosing and encoding variables, splitting the dataset, etc.) tasks that transform the data into a consistent, complete, and valid format before it is used [23] to address and handle these problems mentioned is necessary to obtain high-quality and accurate prediction results. For these reasons, in recent years, regularized prediction models as alternative solution approaches have been widely used and successfully applied in order to address imbalanced and multicollinearity challenges in literature [24], [26-28].

Regularization is a way to overcome the drawbacks of the standard regression models by modifying standard models. In this process, unstable regression parameters are penalized using a tuning parameter to the prediction functions. There are several regularized regression models and Firth, Ridge, Lasso, and Elastic net regression models are well-known in literature. Regularized prediction models help to diminish the variance and sample errors in the model and estimate robust regression parameters against imbalanced and multicollinearity, and improve model performances [24], [27].

In literature, to our knowledge, only three studies have been conducted using regularized prediction models to analyze occupational accidents. Gavanji [24] analyzed occupational injury data from the year 2007 to 2016 using Firth, Lasso, Elastic Net logistic regression models to predict fatal injury claims. Gonzalez-Delgado [26] studied occupational injuries that occurred in 2012 to predict the accident outcome (fatal/non-fatal) using Firth's logistic regression model. Gallego et al. [28] examined accident data between 1995 and 2017 to predict frequency rate, lost workdays, and severity rate in terms of the labour market, economy and productive structure-related variables using Lasso, Elastic Net and Adaptive Lasso linear regression models. In these studies, accident data for a particular industry are not investigated, and regularized logistic regression models are only used to predict injury cases being fatal. Besides, the number of independent variables is less and model performances are only assessed in terms of Bayesian information criteria (BIC) and Akaike's information criterion (AIC) in these studies.

The construction industry has hazardous environments and Law No. 6331 states construction works are a very dangerous occupation. Thus, construction safety is one of the broader fields of research in occupational safety literature. Safety-related studies are mostly conducted in terms of risk assessment and precautions in the construction industry. In addition, although machine learning and data mining applications for prediction and discovering patterns purposes have been conducted in the area of construction safety, to the best of our knowledge, the application of regularized prediction models has not been investigated in this field in the literature. Moreover, the applications of regularized prediction models in the context of occupational accident data and the use of large and variable-rich accident datasets

have thus far been quite limited in the literature. These remaining gaps form the main objectives of this study.

The facts that mentioned above highlighted a need for careful analysis of occupational accidents data and integration of new approaches to standard techniques. Thus, the aim of this study is to cover the issue of occupational accidents in the construction industry taking into account the defined challenges arising from analyzing accident data, and to investigate whether new approaches improve the prediction success as compared to the using standard binary logistic regression. For this reason, the present study will attempt to focus on several research questions for predicting the lost workdays (LWD) as an indicator of severity of occupational accidents in the construction industry:

(1) How can occupational accidents that occurred in the construction industry be analyzed to account for large and categorical variable-rich data?
(2) How can the imbalanced classification problem be solved when construction occupational accident data is analyzed in standard binary logistic regression?
(3) How can the multicollinearity problem be solved when construction occupational accident data is analyzed in standard binary logistic regression?
(4) How can the best-performing model be determined for occupational accident data used?
(5) What is the association between the LWD that resulting from occupational accidents and the accident variables?

This study contributes to the limited safety literature and analyses construction occupational accidents using the proposed approaches for the first time. The following section describes the prediction models, performance metrics and materials used in the study. Section 3 presents the results of the applications, and Section 4 contains the conclusions of the study.

2 Material and method

This section first presents the prediction methods applied and the criteria used to compare classification performance to achieve the aims of this study. The formal definitions of five different supervised machine learning algorithms, which are used to predict a categorical outcome, are provided in the following subsections. These algorithms consist of binary logistic regression as the standard model and the newer, regularized binary logistic regression techniques namely Firth, Ridge, Lasso, and Elastic net logistic regression models. In this study, binary logistic regression is used to model categorical dependent variable, the regularized binary logistic regression models are utilized to eliminate imbalanced and multicollinearity problems while modelling, and enhance the prediction performances of the models. Then, the data set and data pre-processing in this study are described. The data were manipulated by MS Excel 2016 and analyzed using glmnet and brglm packages in Rstudio software version 1.3.1093.

2.1 Binary logistic regression model

Let $V = \{(x_i, y_i): i = 1, 2, \dots, n\}$ be a data set used in the analysis. Here, $x_i = (x_{i1}, \dots, x_{ir})$ is the input vector of r independent variables and y_i is the output (dependent variable) measured on the ith observation, and n is the size of the V. If the output is categorical and takes two possible values which are coded as 1 (outcome present, class of interest) and 0 (outcome absent), the problem is considered as a binary logistic regression problem [29].

The binary logistic regression model is widely used to analyze the relationship between independent variables and a dependent variable in classification problems. This model with multiple independent variables is formally expressed as follows [30]:

$$\pi_i(x_i) = \mathrm{P}(y_i = 1|x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir})} \quad (1)$$

where $\pi_i(x_i) = \mathrm{P}(y_i = 1|x_i)$ represents the conditional probability that y_i is equal to 1 given x_i under ith observation. Additionally, the unknown regression parameters β_0 and $\beta_1 - \beta_r$ indicate the constant term and slope coefficients of independent variables, respectively. After manipulating Equation (1) by applying logit transformation, the model transforms into a linear model as follows [30]:

$$logit\,[\pi_i(x_i)] = ln\left(\frac{\pi_i(x_i)}{1 - \pi_i(x_i)}\right) = \beta_0 + \beta_1 x_{i1} \dots + \beta_r x_{ir} \quad (2)$$

The maximum likelihood method is generally used to determine the association between variables by constructing the likelihood function. The maximum likelihood approach tries to acquire the log-likelihood function maximum while determining unknown regression parameters. The likelihood and log-likelihood functions are expressed as follows, respectively [31]:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$
$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln(\pi_i(x_i)) + (1 - y_i)\, ln(1 - \pi_i(x_i))] \quad (3)$$

After the unknown regression parameters are estimated, the model is tested with the likelihood ratio test. If the estimated parameters are different from zero, the significance of the parameters is examined with p-values. The threshold for p-value is selected as 0.05 or less in this study.

There are some assumptions that should be checked for logistic regression. These basic assumptions including independence, linearity between the logit of the dependent and continuous independent variables, and the absence of multicollinearity must be met [32].

The binary logistic regression can suffer from multicollinearity problem. If the multicollinearity problem occurs, the Wald statistic, which confirms whether corresponding independent variable is significant or not by dividing estimated parameter by standard error gets smaller since the standard errors of the logistic regression parameters

are overestimated [29], [30]. Besides, a variable that contributes to the model is found to be statistically insignificant. Therefore, the presence of the multicollinearity is generally tested using correlation coefficients values in binary logistic regression. If this problem is detected, the relevant variables can be removed from the model, or the number of observations can be increased [33].

Moreover, in the regression analysis, categorical variables need to be digitized using dummy encoding. In this process, if a categorical variable has m categories, $m-1$ indicator columns, in other words dummy variables are introduced. The rule of creating $m-1$ dummy variables is to avoid falling into the "dummy variable trap", that is, perfect collinearity or multicollinearity if there is a perfect relationship between the variables [34]. The remaining m^{th} category of the categorical variable is treated as a reference variable.

In binary logistic regression, the estimated conditional probabilities for each observation are usually compared with the cut-off point of 0.5 [33]. If $\pi_i(x_i) > 0.5$, the value of y_i corresponding to this observation is 1; otherwise, it is classified as 0.

2.2 Regularized binary logistic regression models

When there is an imbalanced distribution between the levels of the dependent variable, and the frequency of events belonging to the class of interest is low, the unknown regression parameters can be estimated with deviations [35]. In other words, these features of the data may lead to quasi-complete or complete separation problems which produce biased or infinite estimates of the unknown parameters. Additionally, the presence of multicollinearity is another challenge in working with data since this problem can reduce the efficiency of the estimations and cause misclassification. Regularized (shrinkage) regression models are used as alternative techniques to classic estimation or prediction approaches for solving separation and multicollinearity problems. The popularity of regularized regression models among machine learning algorithms has been increasing in recent years. Firth, Ridge, Lasso, and Elastic net logistic regression models have been widely used as regularized models in the literature for different fields [28], [36-38]. In regularized logistic regression models, the unknown regression parameters are estimated by adding a penalty term to the log-likelihood functions of related models.

2.2.1 Firth's logistic regression model

Firth's logistic regression model introduced by Firth [39] is used as a possible solution to maximum likelihood estimation for the issues of imbalanced distribution and separation. Firth's logistic regression model is based on the regularized logistic likelihood estimator. The regularized likelihood function in Firth's logistic model can be shown in Equation (4).

$$L^{Firth}(\beta) = L(\beta) \times |I(\beta)|^{\frac{1}{2}} = L(\beta) \times |X^T W X|^{1/2} \quad (4)$$

where β represent the vector of unknown parameters, $L(\beta)$ indicates likelihood function, and $I(\beta)$ is Fisher information matrix in which X is the model matrix and W is the diagonal matrix that is subject to $diag\left(\pi_i(x_i)\left(1-\pi_i(x_i)\right)\right)$ which indicates the impact of each observation on the model [39].

The log-likelihood function $(\ell(\beta))$ is penalized by Jeffrey's invariant prior in Firth's logistic model. After taking the natural log of the corresponding likelihood function, in this case, the regularized log-likelihood function can be written as [24]:

$$\ell^{Firth}(\beta) = \ell(\beta) + (1/2)\, ln|I(\beta)| \quad (5)$$

The second term of the regularized log-likelihood function is maximized at $\pi_i(x_i) = 0.5$ when $\beta = 0$. Therefore, the values of regression parameters are shrunk towards zero (0) [40].

2.2.2 Ridge logistic regression model

The Ridge logistic regression model was initially introduced by Schaefer et al. [41] and later by Duffy and Santer [42] is used when there is multicollinearity between independent variables. This model solves log-likelihood function of the binary logistic regression model using L_2-norm penalty ($\|\beta_j\|_2^2$) with tuning parameter λ that controls the amount of shrinkage. The ridge logistic regression model is expressed as follows [42]:

$$\ell^{Ridge}(\beta) = \ell(\beta) + \lambda \|\beta_j\|_2^2 = \ell(\beta) + \lambda \sum_{j=1}^{r} \beta_j^2 \quad (6)$$

In the ridge logistic regression model, the value of λ shrinks the estimated values of regression parameters towards 0, but none of the estimated parameters becomes exactly 0 [43], [44]. It estimates coefficients for all of the independent variables included in the model. This case is considered as a disadvantage of ridge regularization since it reduces the interpretability of the model. If regularization and coefficient estimation for all of the variables are needed, this model can be utilized [45].

2.2.3 Lasso logistic regression model

The Lasso model proposed by Tibshirani [46] is another regularized regression approach. The Lasso regression model overcomes the disadvantage of ridge regression since it performs variable selection process. This property provides an advantage to interpret the results of model more easily than ridge model. Therefore, if the regularization is required and the weights of less important variables need to be reduced to 0, the lasso model can be applied.

Lasso logistic regression model is obtained by adding different penalty term called L_1-norm ($\|\beta_j\|_1$) with tuning parameter λ to the negative log-likelihood function. This model has the following form [36]:

$$\ell^{Lasso}(\beta) = -\ell(\beta) + \lambda \left\|\beta_j\right\|_1 = -\ell(\beta) + \lambda \sum_{j=1}^{r} |\beta_j| \quad (7)$$

There are some drawbacks that make model less stable in lasso logistic regression model. This model randomly chooses any variable among highly correlated variables and ignores the rest ones. Besides, it chooses at most n independent variables in high dimensional data ($r > n$). However, there may be more variable parameters than n without 0 values in the last model. Another drawback of this model is that lasso model function is not exactly convex; thus, different estimates can be obtained according to the order of different independent variables when fitting the model.

2.2.4 Elastic net logistic regression model

The Elastic net model introduced by Zou and Hastie [47] is used as another regularization and variable selection approach. This model addresses the drawbacks of Lasso regression. The Elastic net model is a combination of lasso and ridge models. In other words, it uses a mixture of L_1-norm and L_2-norm penalties that conveys the features of both lasso and ridge, respectively. The Elastic net approach uses another parameter to tune, and provides a balance association between reducing the size of parameters and shrinking them to 0.

The elastic net logistic regression model adds two regularization terms to the log-likelihood function with a mixing parameter α that indicates the degree of balance between lasso and ridge approaches. This model can be formulated as follows [47]:

$$\ell^{Elastic\ net}(\beta) = \ell(\beta) + \lambda \left[\frac{1}{2}(1-\alpha) \sum_{j=1}^{r} \beta_j^2 + \alpha \sum_{j=1}^{r} |\beta_j| \right] \quad (8)$$

The elastic net logistic regression model is equivalent to ridge logistic regression when $\alpha = 1$ and to lasso logistic regression when $\alpha = 0$. The value of α is usually taken as 0.5 to perform an equal combination of ridge and lasso models.

2.3 Choice of tuning parameter

Determining the value of tuning parameter λ (lambda) is critical since it controls the parameter estimates. The λ parameter is both a monotonically decreasing function of the variance of parameter estimation and a monotonically increasing function of the bias of this parameter [27]. When the value of λ increases, the variance decreases and the bias increases. The tuning parameter λ is generally determined by data-driven approaches such as k-fold cross validation.

In k-fold cross-validation, the data set is split into to k groups, where one group is used as a test data set and other k-1 groups form the training data set [48]. 5-fold or 10-fold cross-validation is commonly used. In 10-fold cross validation, the data is splitting into 10 sub-samples of equal size [48], [49]. A training set consisting of 9 sub-samples are used to fit the model and test set consisting of the rest of one sub sample is utilized to assess the model's validity. This process is repeated until each sub-sample is used once as a test set. Then, the value of λ is calculated.

In literature, there are two main λ parameters which are known as $lambda.min$ and $lambda.1se$. The values of $lambda.min$ and $lambda.1se$ indicate the value of minimum misclassification error and the most regularized model of the misclassification error within one standard error of the minimal error, respectively [50]. The values of estimated parameters depend on the amounts of λ. It is important to choose right value of λ to avoid overfitting and underfitting issues for the model. Cross validation is widely utilized to choose the proper λ values that provide a proper balance between variance and bias and diminish the misclassification error. For this reason, $cv.glmnet()$ from glmnet package is used to determine the best value of λ in this study.

2.4 Prediction performance comparison criteria

2.4.1 Akaike's information criterion (AIC)

The AIC is a measure of fit to assess the different models. It adds a penalized term with the number of estimated parameters included in the model to the value of the log-likelihood. The value of AIC is computed as follows [30]:

$$AIC = -2\big(\ell(\beta)\big) + 2d \quad (9)$$

where d is the number of estimated parameters. Smaller AIC values mean better model fit.

2.4.2 Confusion matrix and related metrics

The confusion matrix and several metrics obtained with the help of this matrix are used to measure the prediction performance of the logistic regression models. The general structure of the confusion matrix for the binary classification with actual values on one axis and predicted values calculated with the classification algorithm on another are given in Table 2.

Table 2. Confusion matrix of binary classification problem [51]

	Predicted values	
Actual values	Negative	Positive
Negative (N)	True Negative (TN)	False Positive (FP)
Positive (P)	False Negative (FN)	True Positive (TP)

True positive (TP) and True negative (TN) show outcomes that are correctly predicted as positive and negative, respectively. Contrarily, False positive (FP) and False negative (FN) indicate the negative and positive outcomes that are wrongly predicted as positive and negative, respectively. After constructing the confusion matrix, four basic rates are utilized to describe the predictive quality of a model. These rates are given below [51-53]:

$$True\ positive\ rate = \frac{TP}{TP + FN} \quad (10)$$

$$True\ negative\ rate = \frac{TN}{TN + FP} \quad (11)$$

$$Positive\ predictive\ value = \frac{TP}{TP + FP} \quad (12)$$

$$Negative\ predictive\ value = \frac{TN}{TN + FN} \quad (13)$$

The true positive rate is also known as $sensitivity$ or $recall$. The true negative rate is also called $specificity$ and the positive predictive value is synonyms with $precision$.

Other confusion matrix metrics are a combination of the basic rates. Besides, if there is an imbalanced distribution among classes, balanced accuracy, F1 and G-means metrics are used to avoid the misleading results rather than others [54]. Metrics used in this study are defined with their related formulations below.

Balanced accuracy is a widely used metric to deal with imbalanced data. It is the arithmetic mean of the true positive and true negative rates. In other words, it adds sensitivity and specificity scores and divides their addition by two. This score is computed as follows [51]:

$$Balanced\ Accuracy = \frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} \quad (14)$$

$F1$ is another appropriate metric in the case of imbalanced class distribution. $F1$ is defined as the harmonic mean of the precision and sensitivity [52]. It can be formulated as follows:

$$F1 = \frac{2 \times \left(\frac{TP}{TP + FN} \times \frac{TP}{TP + FP}\right)}{\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP}\right)} \quad (15)$$

$G-$mean is also used as an evaluation metric for the imbalance data learning. $G-$mean is the geometric mean of sensitivity and specificity [54]. It can be defined as follows:

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (16)$$

The values of comparison metrics are between 0 (the worst possible) and 1 (the best). The higher the comparison metrics scores, the better result. There are no standard range rules for all the metrics. These scores strongly depend on how imbalanced and high-dimensional the data is. Therefore, they are interpreted based on particular prediction problems in the literature [29], [55-56].

2.4.3 *Areas under the curves*

The receiver operating characteristic (ROC) curve and precision-recall (PR) curve with areas under these curves have also been widely used to compare the performance of models [29], [52]. The ROC curve plots the proportion of true positive rate (sensitivity or recall) and false positive rate $(1 - specificity)$ of the model. The area under the ROC curve $(ROC\ AUC)$ represents the performance of the model by the rate of correct classification of positive and negative samples. The random prediction value which is a baseline measure for $ROC\ AUC$ is 0.50. If the value of $ROC\ AUC$ is smaller than 0.50, it indicates that the model performance is worse than the random prediction. A $ROC\ AUC$ value close to 1 means better performance for the model. However, if this value is 1, it implies that the model has the perfect skill to classify observations into classes. Alternatively, a PR curve which plots the precision against the recall and the area under the PR curve $(PR\ AUC)$ can be used to evaluate the model performance. The baseline random prediction value for $PR\ AUC$ is the rate of actual positives (P) to total of actual positives and actual negatives (N) which is $(P/(P + N))$ [55]-[56]. The $PR\ AUC$ value greater than this rate represents better model performance. The PR curve is more appropriate for imbalanced data since it ignores the true negatives.

2.5 Data set, data pre-processing and data analysis process

The national data set used for this study is the construction occupational non-fatal injury data from January 2013 to December 2017 for workers in the Central Anatolia region which accounts for approximately 20% of the occupational accidents in the construction industry in Türkiye. The data was obtained from the Turkish SSI. A total of 25,944 non-fatal occupational accidents were extracted from the data set after omitting missing values and "no information" entries as a first stage of data pre-processing.

All variables were categorized into five groups as victim-related, workplace-related, accident time-related, accident and sequence of events-related and post-accident state-related variables. Additional independent variables are included in this study using the time and date variables of the data set and these are work experience, accident year, accident month, accident day and hour of the accident. Some numerical variables; age, work experience and size of workplace were required to be converted to categorical variables after checking logistic regression assumptions since the assumption of linearity between the logit of the dependent and continuous independent variables is not satisfied. Geographical locations were grouped by taking into account the insured working population of the cities. In this study, the categories of the independent variables followed the European Statistics on Accidents at Work (ESAW) methodology.

Age group, education level, marital status, last work experience, OSH and vocational training status of the victim, and occupation of the victim based on International Standard Classification of Occupations (ISCO-08) groups are included in the victim-related variables. In addition, local unit where accident occurs, economic activity of the project based on Nomenclature of Economic Activities (NACE) codes which group organizations according to the their business activities, total number of workers working at the

workplace of the victim, the place/post occupied by the victim, and the type of workplace, working area or location where the victim was present or working just before the accident are the variables within workplace-related group. Besides, accident-time related group provide information about year, month, day, and hour which accident occurs. Accident and sequence of events-related group addresses information regarding the process where the accident take place, main type of work/task being performed by the victim at accident time, the activity being performed by the victim just before the accident, the contact that injured the victim, and the event that triggers the accident, the tool, object, or instrument involved in the abnormal event. Moreover, post-accident state-related group reveal information regarding the physical results for the victim, the part of the body injured, and the number of days where the victim is unfit for work due to an accident at work.

The number of LWD was taken as the dependent variable in the analysis of accident data. The variable LWD refers to the number of days away from work where the workers who suffered an occupational accident are unfit for work [57]. In ESAW methodology, only cases of accidents with more than three days of absence are considered. Therefore, the variable LWD was converted into a categorical variable based on ESAW criteria in this study to predict the probability of non-fatal occupational accidents with more than and less than or equal to three LWDs. The dependent variable was coded as "1" for more than 3 days lost accidents, and '0' for less than or equal to 3 days lost accidents. Table 3 presents all variables considered in this study with their brief definitions and descriptive statistics.

Table 3. Variable set used for the study

Variable Group	Variable	Subcategory of Variable (n)
Victim-related	Age Group (AGE_GR)	AGE.GR.1: 14-24 age (n = 6089)
		AGE.GR.2: 25-34 age (n = 8265)
		AGE.GR.3: 35- 44 age (n = 6184)
		AGE.GR.4: 45-54 age (n = 4296)
		AGE.GR.5: 55-64 age (n = 1046)
		AGE.GR.6: 65 age and above (n = 64)
	Education (EDU)	EDU.1: Not literate (n = 100)
		EDU.2: Literate (n = 3785)
		EDU.3: Primary school (n = 8181)
		EDU.4: Secondary school (n = 7889)
		EDU.5: High school (n =5077)
		EDU.6: Undergraduate level (n = 912)
	Marital status (MAS)	MAS.1: Single (n = 8164)
		MAS.2: Married (n = 16906)
		MAS.3: Other (n = 874)
	Experience (EXP)	EXP.1: $\leq$ 1 month (n = 9894)
		EXP.2: 1-6 months ($n = 10863$)
		EXP.3: 6-12 months ($n = 2516$)
		EXP.4: 12-18 months ($n = 931$)
		EXP.5: 18-24 months ($n = 481$)
		EXP.6: > 24 months ($n = 1259$)
	OSH education (OSHEDU)	OSHEDU.1: Yes ($n = 23740$), OSHEDU.2: No ($n = 2204$)
	Vocational education (VOCEDU)	VOCEDU.1: Yes ($n = 20482$), VOCEDU.2: No ($n = 5462$)
	Occupation (OCC)	OCC.1: Managers ($n = 18$)
		OCC.2: Professional ($n = 214$)
		OCC.3: Technicians and associate professionals ($n = 2143$)
		OCC.4: Clerical support workers ($n = 117$)
		OCC.5: Service and sales workers ($n = 342$)
		OCC.6: Skilled agricultural, forestry and fishery workers ($n = 48$)
		OCC.7: Craft and related trades workers ($n = 12688$)
		OCC.8: Plant and machine operators, and assemblers ($n = 2127$)
		OCC.9: Elementary occupations ($n = 8247$)
Workplace-related	Geographical Location (GEOL)	GEOL.1: Ankara ($n = 14261$)
		GEOL.2: Eskisehir- Sivas ($n = 3047$)
		GEOL.3: Konya-Kayseri ($n = 4990$)
		GEOL.4: Others ($n = 3646$)
	Project Type (PRT)	PRT.1: Construction of buildings ($n = 16424$)
		PRT.2: Construction of roads and railways ($n = 2932$)
		PRT.3: Construction of utility projects ($n = 2273$)
		PRT.4: Construction of other civil engineering projects ($n = 468$)
		PRT.5: Demolition and site preparation ($n = 426$)

Table 3. Variable set used for the study (continued)

Variable Group	Variable	Subcategory of Variable (n)
		PRT.6: Electrical, plumbing and other construction installation
		PRT.7: Building completion and finishing ($n = 923$)
		PRT.8: Other specialised construction activities ($n = 943$)
	Size of the Workplace (SIZW)	SIZW.1: < 10 workers ($n = 2164$)
		SIZW.2: 10-20 workers ($n = 2198$)
		SIZW.3: 21-49 workers ($n = 3348$)
		SIZW.4: 50-99 workers ($n = 3050$)
		SIZW.5: 100-199 workers ($n = 3097$)
		SIZW.6: 200-249 workers ($n = 1015$)
		SIZW.7: 250-499 workers ($n = 3114$)
		SIZW.8: 500-999 workers ($n = 2736$)
		SIZW.9: 1000 and above workers ($n = 5222$)
	Workstation (WOR)	WOR.1: Usual workstation or within the usual local unit of work ($n = 5919$)
		WOR.2: Occasional or mobile workstation or in a journey on behalf of the employer ($n = 17607$)
		WOR.3: Other workstation ($n = 2418$)
	Working environment (WOE)	WOE.1: Industrial site ($n = 1977$)
		WOE.2: Construction site, construction, opencast quarry, opencast mine ($n = 19324$)
		WOE.3: Farming, breeding, fish farming, forest zone ($n = 27$)
		WOE.4: Tertiary activity area, office, amusement area, miscellaneous ($n = 154$)
		WOE.5: Health establishment ($n = 134$)
		WOE.6: Public area ($n = 1088$)
		WOE.7: In the home ($n = 99$)
		WOE.8: Sports area ($n = 23$)
		WOE.9: In the air, elevated, excluding construction sites ($n = 37$)
		WOE.10: Underground, excluding construction sites ($n = 238$)
		WOE.11: On /over water, excluding construction sites ($n = 16$)
		WOE.12: In high pressure environments, excluding construction sites ($n = 29$)
		WOE.13: Other ($n = 2798$)
Accident time-related	Year (YEAR)	YEAR.1: 2013 ($n = 3595$), YEAR.2: 2014 ($n = 4302$), YEAR.3: 2015 ($n = 4955$), YEAR.4: 2016 ($n = 6220$), YEAR.5: 2017 ($n = 6872$)
	Month (MONTH)	MTH.1: January ($n = 1259$), MTH.2: February ($n = 1452$), MTH.3: March ($n = 1930$), MTH.4: April ($n = 2171$), MTH.5: May ($n = 2334$), MTH.6: June ($n = 2213$), MTH.7: July ($n = 2326$), MTH.8: August ($n = 2808$), MTH.9: September ($n = 2203$), MTH.10: October ($n = 2396$), MTH.11: November ($n = 2664$), MTH.12: December ($n = 2188$)
	Day (DAY)	DAY.1: Monday ($n = 4185$), DAY.2: Tuesday ($n = 4038$), DAY.3: Wednesday ($n = 4026$), DAY.4: Thursday ($n = 4116$), DAY.5: Friday ($n = 4022$), DAY.6: Saturday ($n = 3395$), DAY.7: Sunday ($n = 2162$)
	Hour (HOUR)	HOUR.1: 00:00-01:59 ($n = 174$), HOUR.2: 02:00-03:59 ($n = 210$), HOUR.3: 04:00- 05:59 ($n = 160$), HOUR.4: 06:00-07:59 ($n = 372$), HOUR.5: 08:00-09:59 ($n = 4613$), HOUR.6: 10:00-11:59 ($n = 7022$), HOUR.7: 12:00-13:59 ($n = 2918$), HOUR.8: 14:00-15:59 ($n = 5393$), HOUR.9: 16:00-17:59 ($n = 3787$), HOUR.10: 18:00-19:59 ($n = 663$), HOUR.11: 20:00-21:59 ($n = 374$), HOUR.12: 22:00-23:59 ($n = 258$)
Accident and sequence of events-related	Process of accident (PRA)	PRA.1: At work ($n = 23463$)
		PRA.2: Rest break ($n = 877$)
		PRA.3: Busy with occupational activity ($n = 1422$)
		PRA.4: Commuting from work to home ($n = 182$)
	General activity (GENAC)	GENAC.1: Production, manufacturing, processing, storing ($n = 3076$)
		GENAC.2: Excavation, Construction, Repair, Demolition ($n = 14577$)
		GENAC.3: Agricultural type work, forestry, horticulture, fish farming, work with live animals ($n = 33$)
		GENAC.4: Service provided to enterprise and/or to the general public; intellectual activity ($n = 250$)
		GENAC.5: Other work related to above first four tasks ($n = 294$)
		GENAC.6: Movement, sport, artistic activity ($n = 399$)
		GENAC.7: Other working processes not listed in the above classification ($n = 7315$)

Table 3. Variable set used for the study (continued)

Variable Group	Variable	Subcategory of Variable (n)
	Specific Activity (SPECAC)	SPECAC.1: Operating machine ($n = 719$)
		SPECAC.2: Working with hand-held tools ($n = 4777$)
		SPECAC.3: Driving/being on board a means of transport or handling equipment ($n = 989$)
		SPECAC.4: Handling of objects ($n = 2900$)
		SPECAC.5: Carrying by hand ($n = 3331$)
		SPECAC.6: Movement ($n = 3426$)
		SPECAC.7: Presence ($n = 1615$)
		SPECAC.8: Other specific physical activities ($n = 8187$)
	Mode of Injury (MODI)	MODI.1: Contact with electrical voltage, temperature, hazardous substances ($n = 493$)
		MODI.2: Drowned, buried, enveloped ($n = 23$)
		MODI.3: Horizontal or vertical impact with or against a stationary object (the victim is in motion) ($n = 4025$)
		MODI.4: Struck by object in motion, collision with ($n = 3088$)
		MODI.5: Contact with sharp, pointed, rough, coarse material agent ($n = 3579$)
		MODI.6: Trapped, crushed, etc. ($n = 2516$)
		MODI.7: Physical or mental stress ($n = 180$)
		MODI.8: Bite, kick, etc. (animal or human) ($n = 107$)
		MODI. 9: Other contacts ($n = 11933$)
	Deviation (DEV)	DEV.1: Deviation due to electrical problems, explosion, fire ($n = 439$)
		DEV.2: Deviation by overflow, overturn, leak, flow, vaporisation, emission ($n = 1267$)
		DEV.3: Breakage, bursting, splitting, slipping, fall, collapse of material agent ($n = 3520$)
		DEV.4: Loss of control of machine, means of transport or handling equipment, handheld tool, object, animal ($n = 4160$)
		DEV.5: Slipping, stumbling and falling, fall of persons ($n = 6081$)
		DEV.6: Body movement without any physical stress ($n = 1519$)
		DEV.7: Body movement under or with physical stress ($n = 846$)
		DEV.8: Shock, fright, violence, aggression, threat, presence ($n = 121$)
		DEV.9: Other Deviations ($n = 7991$)
	Material agent of the Deviation (MAT_DEV)	MAT_DEV.1: No material agent ($n = 2217$)
		MAT_DEV.2: Buildings, structures, surfaces (at ground level) ($n = 2155$)
		MAT_DEV.3: Buildings, structures, surfaces (above ground level) ($n = 2010$)
		MAT_DEV.4: Buildings, structures, surfaces (below ground level) ($n = 183$)
		MAT_DEV.5: Systems for the supply and distribution of materials, pipe networks ($n = 208$)
		MAT_DEV.6: Motors, systems for energy transmission and storage ($n = 57$)
		MAT_DEV.7: Hand tools ($n = 4003$)
		MAT_DEV.8: Machines and equipment ($n = 1590$)
		MAT_DEV.9: Conveying, transport and storage systems ($n = 554$)
		MAT_DEV.10: Land and other transport vehicles ($n = 1288$)
		MAT_DEV.11: Materials, objects, products, machine or vehicle components, debris, dust ($n = 3581$)
		MAT_DEV.12: Chemical, explosive, radioactive, biological substances ($n = 128$)
		MAT_DEV.13: Safety devices and equipment ($n = 24$)
		MAT_DEV.14: Office equipment, personal equipment, sports equipment, weapons, domestic appliances ($n = 66$)
		MAT_DEV.15: Living organisms and human-beings ($n = 101$)
		MAT_DEV.16: Bulk waste ($n = 92$)
		MAT_DEV.17: Physical phenomena and natural elements ($n = 287$)
		MAT_DEV.18: Other material agents ($n = 7400$)
Post-accident state-related	Type of injury (TINJ)	TINJ.1: Wounds and superficial injuries ($n = 12322$)
		TINJ.2: Bone fractures ($n = 3416$)
		TINJ.3: Dislocations, sprains and strains ($n = 3563$)
		TINJ.4: Traumatic amputations ($n = 67$)
		TINJ.5: Concussion and internal injuries ($n = 202$)

Table 3. Variable set used for the study (continued)

Variable Group	Variable	Subcategory of Variable (n)
		TINJ.6: Burns, scalds and frostbites ($n = 322$)
		TINJ.7: Poisonings and infections ($n = 207$)
		TINJ.8: Effects of sound, vibration and pressure ($n = 38$)
		TINJ.9: Shock ($n = 61$)
		TINJ.10: Multiple injuries ($n = 219$)
		TINJ.11: Others ($n = 5527$)
	Part of Body Injured (PBINJ)	PBINJ.1: Head ($n = 4194$)
		PBINJ.2: Neck, inclusive spine and vertebra in the neck ($n = 248$)
		PBINJ.3: Back, including spine and vertebra in the back ($n = 865$)
		PBINJ.4: Torso and organs ($n = 802$)
		PBINJ.5: Upper Extremities ($n = 8739$)
		PBINJ.6: Lower Extremities ($n = 5432$)
		PBINJ.7: Whole body and multiple sites ($n = 608$)
		PBINJ.8: Other Parts of body injured ($n = 5056$)
	Lost Workdays (LWD)	Less than or equal to 3 days (≤ 3 days) ($n = 19010$)
		More than 3 days (> 3 days) ($n = 6934$)

The distribution and relationship of each independent variable with the dependent variable were investigated by cross-tabulation approach. It was determined that there was no significant relationship between only the independent variables of WOE, DAY and HOUR and the dependent variable of LWD. However, these three variables were also included in the analysis to reveal the combined effect between LWD and all independent variables. In this way, the prediction models were created for LWD, and the validity of these models was tested without initial variable selection in the present study.

All categorical variables were handled as another step of data pre-processing. 157 dummy variables were created from 24 independent variables. Additionally, data were split into training set and test set data by using a stratified random sampling approach with a ratio of 70:30. The purpose of using the stratified random sampling method is to give an equal chance of being selected for each data element in the data set and to reduce the prediction variance. In this process, the training set was used to develop and tune the models and the test set was used to test the models' predictive performance.

3 Results and discussions

Out of a total of 25,944 accidents considered in the study, 18,161 accidents were used in the training set and 7,783 accidents were used in the test set. Table 4 provides the distribution details of the LWD variable in both the training and test data sets.

Table 4. Distribution of the variable of LWD on training and test data

Data set	≤ 3 days	> 3 days	Total
Train set	13,307	4,854	18,161
Test set	5,703	2,080	7,783

In our analysis, as the event of interest, the number of accidents with more than 3 workdays lost was low (< 27%). It appears that there is an imbalanced distribution between the two classes of the dependent variable, since the event of interest, in other words the positive class (> 3 days), occurred less frequently than the negative class (≤ 3 days).

The reference categories of the dependent variable and independent variables were determined through the normative approach which considers the most interesting comparisons in the logical sense. The presence of multicollinearity between variables in the training data set was checked through pearson and spearman's rho correlation analyses. It was observed that many independent variables were highly correlated with each other in the training data.

Due to the mentioned reasons, regularized logistic regression models are needed to deal with the imbalanced distribution and multicollinearity issues. Thus, in this study, five different prediction models, which are explained in detail in Section 2, were used to predict more than 3 workdays lost on the training data and to compare their prediction abilities. Moreover, an attempt to optimize the tuning parameters of the prediction models was done using a using 10-fold cross-validation algorithm to increase the prediction success of the models.

3.1 Application of prediction models

The results obtained from the application of Binary logistic, Firth, Ridge, Lasso and Elastic Net logistic regression methods to develop the corresponding regression models for the LWD indicator are presented in this section.

To run Ridge, Lasso and Elastic net logistic models, glmnet package in Rstudio is used. For fitting these models, cv.glmnet function from glmnet package is utilized. This function conducts 10-fold cross-validations to choose the proper value of tuning parameter λ. Besides, the brglm package in Rstudio is used to run Firth's model. To develop the Elastic Net model, the mixing parameter α-value was chosen as 0.5 to give equal weight to Ridge and Lasso logistic regression models.

In Figure 1, the plots depict the variation of coefficients and the misclassification error against tuning parameter log lambda in Ridge, Lasso and Elastic Net logistic regression models obtained for the accident data set of 70% training data set.

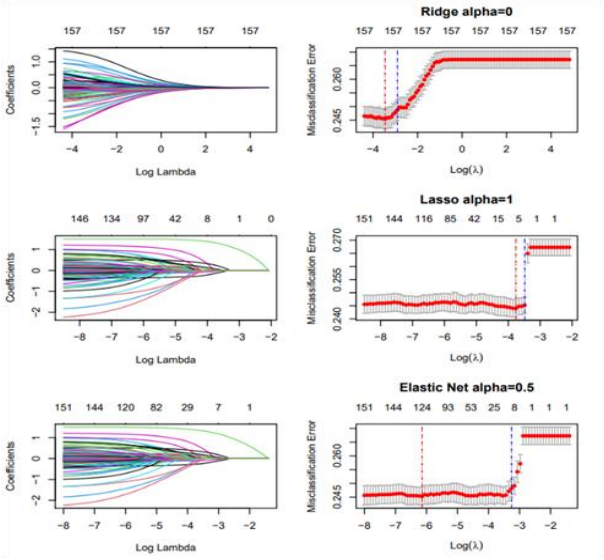


Figure 1. The variation of coefficients in Ridge, Lasso and Elastic net logistic regression models (on the left) and 10-fold cross-validated estimate of the misclassification prediction errors (on the right) as a function of log lambda showing the regularization of coefficients

In Figure 1, there are two vertical lines on the right-side plots. The red and blue dashed lines show the logarithmic transformation values of the $lambda.min$ and $lambda.1se$ values calculated for each model by cross-validation algorithm, respectively. Additionally, the numbers on the top of the plots indicate nonzero coefficient estimates, in other words, the independent variables selected by the relevant models for a given log lambda. As can be seen from the Ridge logistic plot, this number is the same as the number of variables in the data and is constant for all the lambda values. On the other hand, in Lasso and Elastic Net plots, the values of some coefficients are zero, which means that the relative variables are insignificant and removed in the models.

The Ridge logistic yields $lambda.min$ =0.0314 and $lambda.1se$ =0.0549 with 157 dummy variable of coefficient estimates. The number of variables in the lasso logistic model is 6 with $lambda.min$ =0.0232 and 4 with $lambda.1se$ =0.0307. Besides, the number of variables remaining in the model via Elastic net logistic is 124 if $lambda.min$ =0.0022 while it is 8 if $lambda.1se$ =0.0385.

Standard binary and Firth's logistic regression can indicate if an independent variable is statistically significantly related with the dependent variable based on p-values or confidence intervals. Nevertheless, Ridge, Lasso and Elastic net models do not generate meaningful p-values or confidence intervals [58]. Thus, the coefficient estimates of Ridge, Lasso and Elastic Net logistic models for each of the lambda values are provided. The estimated coefficients in the standard and regularized logistic regression models are shown in Table 5.

According to the results of the coefficient estimate, the absolute coefficient values of the variables computed mostly appears to be decreased and approach zero via Ridge logistic compared to the standard binary and Firth's logistic regression models. Moreover, many variables are removed by Lasso and Elastic net logistic models. However, there may be some insignificant variables in Lasso and Elastic net logistic models since both models can eliminate less important variables from the model and perform better. The results also showed that the relationship between some variables and LWD is negative. This implies that a one-unit increase in the value of the variables that have a positive relationship with LWD increases the probability of occupational accidents resulting in more than three LWD, while those that are negatively correlated have a decreasing effect on this probability and less susceptible to LWD.

In the standard binary logistic regression model, the variables determined to be significant were AGE_GR.4, EDU.3, EDU.4, EDU.5, VOCEDU.2, SIZW.9, PRT.2, PRT.3, WOE.2, WOE.5, WOE.10, MTH.8, MTH.9, YEAR.2, YEAR.3, YEAR.4, YEAR.5, GENAC.6, SPECAC.7, SPECAC.8, MAT_DEV.8, TINJ.2, TINJ.3, TINJ.4, TINJ.5, TINJ.6, TINJ.7, TINJ.9, TINJ.10, PBINJ.2, PBINJ.3, PBINJ.4, PBINJ.5, PBINJ.6, PBINJ.7 and PBINJ.8 according to the p-values ($p<0.05$). All independent variables found to be significant with the binary logistic regression model were also significant in Firth's logistic model except the variables EDU.3 and YEAR.3.

The most important variable in Ridge, Lasso and Elastic net logistic models was "TINJ.2" since it had the greatest absolute value of the coefficient. It also indicates that there is a positive relationship between bone fractures (TINJ.2) and LWD. Additionally, "PBINJ.5" is another most important variable in Lasso lambda.1se model and there is a positive relationship between upper extremity injuries (PBINJ.5) and LWD.

Overall, "TINJ" and "PBINJ" have the most influence on LWD in all models. Other variables were chosen differently in each model. Besides, the effects of each variable used in the data set in consideration of the reference category can be interpreted by using the $Exp(\beta_i)$ for positive and $\frac{1}{Exp(\beta_i)}$ for negative coefficient estimates. For example, "TINJ.1" was chosen as reference category in the variable of "TINJ". According to the "TINJ.1", the probability of having more than three LWD is 4.4915 ($Exp(1.5022)$) times higher in accidents with "TINJ.2" in the standard binary logistic model (if all other variables are held constant). This probability value is equal to 4.4233 in Firth's logistic, 3.6987 in Ridge lamda.min, 3.7889 in Lasso lambda.min, 4.4812 in Elastic net lambda.min, 3.2851 in Ridge lambda.1se, 3.4542 in Lasso lambda.1se and 3.5233 in Elastic net lambda.1se logistic model. Additionally, occupational accidents with "PBINJ.5" have the highest probability of exposure to more than three LWD (4.4812 times) compared to the reference category "PBINJ.1" by considering Lasso lambda.1se model. Using the same calculation approach, the relationship between LWD and other independent variables can be discussed for each model. In this study, the results were also interpreted based on the best-performing model in the following subsection.

Table 5. Coefficient estimates of the training data for different logistic regression models

Variable	Subcategory of	Binary	Firth's	Ridge logistic	Lasso logistic	Elastic net	Ridge logistic	Lasso logistic	Elastic net
	Constant	-1.0157	-0.9613	-1.4153	-1.2448	-1.6338	-1.3453	-1.2151	-1.2528
AGE_GR	AGE_GR.1					Reference			
	AGE_GR.2	0.0272	0.0275	-0.0113			-0.0170		
	AGE_GR.3	0.0772	0.0769	0.0422		0.0425	0.0361		
	AGE_GR.4	0.2230	0.2217	0.1645		0.1817	0.1472		
	AGE.GR.5	0.0360	0.0372	0.0129			0.0147		
	AGE.GR.6	0.5695	0.5827	0.4316		0.4055	0.3710		
EDU	EDU.1	-0.3989	-0.3736	-0.4741		-0.4985	-0.4224		
	EDU.2	0.1362	0.1306	-0.0734		-0.0906	-0.0807		
	EDU.3	0.2407	0.2343	0.0302			0.0177		
	EDU.4	0.2864	0.2797	0.0665		0.0421	0.0496		
	EDU.5	0.3437	0.3362	0.1199		0.1028	0.0980		
	EDU.6					Reference			
MAS	MAS.1	-0.0598	-0.0588	-0.0887		-0.0800	-0.0895		
	MAS.2					Reference			
	MAS.3	-0.0807	-0.0785	-0.0659		-0.0395	-0.0570		
OCC	OCC.1					Reference			
	OCC.2	-0.5835	-0.5938	0.0565			0.0427		
	OCC.3	-0.7863	-0.7974	0.0075			0.0092		
	OCC.4	-1.2572	-1.2365	-0.3709		-0.3232	-0.3262		
	OCC.5	-0.8742	-0.8797	-0.0712		-0.0118	-0.0655		
	OCC.6	0.2617	0.2439	0.8269		0.8615	0.7198		
	OCC.7	-0.8305	-0.8416	-0.0337		-0.0192	-0.0340		
	OCC.8	-0.8647	-0.8744	-0.0229			-0.0130		
	OCC.9	-0.7729	-0.7839	0.0352		0.0354	0.0366		
EXP	EXP.1	-0.0754	-0.0743	-0.0281		-0.0042	-0.0212		
	EXP.2	-0.0961	-0.0948	-0.0505		-0.0293	-0.0436		
	EXP.3	-0.0320	-0.0308	0.0077		0.0065	0.0127		
	EXP.4	0.1244	0.1246	0.1173		0.1376	0.1042		
	EXP.5	0.0763	0.0786	0.0978		0.0929	0.0951		
	EXP.6					Reference			
OSHEDU	OSHEDU.1					Reference			
	OSHEDU.2	-0.1273	-0.1261	-0.0481		-0.0485	-0.0250		
VOCEDU	VOCEDU.1					Reference			
	VOCEDU.2	0.0974	0.0966	0.0872		0.0753	0.0818		
GEOL	GEOL.1					Reference			
	GEOL.2	-0.0178	-0.0168	-0.0133			-0.0107		
	GEOL.3	0.0938	0.0930	0.0985		0.0940	0.0959		
	GEOL.4	-0.0014	-0.0008	0.0048			0.0082		
SIZW	SIZW.1					Reference			
	SIZW.2	0.0694	0.0693	0.1179		0.0841	0.1181		
	SIZW.3	-0.0059	-0.0060	0.0566		0.0218	0.0622		
	SIZW.4	-0.0049	-0.0051	0.0553		0.0142	0.0598		
	SIZW.5	-0.0502	-0.0496	0.0222			0.0303		
	SIZW.6	-0.0795	-0.0781	-0.0102		-0.0030	-0.0034		
	SIZW.7	-0.0516	-0.0513	0.0137			0.0186		
	SIZW.8	-0.1587	-0.1573	-0.0764		-0.0985	-0.0635		
	SIZW.9	-0.4091	-0.4063	-0.2931	-0.1452	-0.3577	-0.2610	-0.0487	-0.1839
PRT	PRT.1					Reference			
	PRT.2	0.2140	0.2127	0.1460		0.1644	0.1214		
	PRT.3	0.1711	0.1701	0.1255		0.1247	0.1092		
	PRT.4	0.1736	0.1741	0.1395		0.1174	0.1243		
	PRT.5	0.1605	0.1601	0.1488		0.1067	0.1361		
	PRT.6	0.0917	0.0919	0.0793		0.0512	0.0680		
	PRT.7	0.0488	0.0503	0.0590		0.0250	0.0553		
	PRT.8	0.1313	0.1331	0.1445		0.1098	0.1367		
WOR	WOR.1					Reference			
	WOR.2	0.0658	0.0648	0.0296		0.0124	0.0197		
	WOR.3	0.0535	0.0538	0.0266			0.0182		
WOE	WOE.1					Reference			
	WOE.2	-0.1977	-0.1952	-0.0991		-0.0852	-0.0808		
	WOE.3	-0.9006	-0.7954	-0.6249		-0.5436	-0.5295		
	WOE.4	-0.1828	-0.1710	-0.0790			-0.0586		
	WOE.5	-1.9123	-1.8124	-1.2082		-1.4822	-1.0003		
	WOE.6	-0.1571	-0.1540	-0.0240			0.0049		
	WOE.7	-0.5308	-0.5021	-0.3451		-0.2764	-0.2904		
	WOE.8	-0.4056	-0.3705	-0.2378		-0.1253	-0.1607		
	WOE.9	0.2243	0.2254	0.3943		0.2731	0.4101		
	WOE.10	-1.3909	-1.3513	-0.9601		-1.1265	-0.8084		
	WOE.11	-0.7494	-0.5837	-0.4099		-0.2184	-0.3236		
	WOE.12	-0.9307	-0.7365	-0.5717		-0.4394	-0.4725		
	WOE.13	-0.1631	-0.1607	-0.0284			-0.0045		
DAY	DAY.1	0.0004	0.0000	-0.0026			-0.0036		
	DAY.2	-0.0387	-0.0388	-0.0317		-0.0181	-0.0273		
	DAY.3	0.0625	0.0612	0.0536		0.0566	0.0478		
	DAY.4	0.0767	0.0757	0.0652		0.0677	0.0590		

Table 5. Coefficient estimates of the training data for different logistic regression models (continued)

Variable	Subcategory of Variable	Binary logistic	Firth's logistic	Ridge logistic lambda.min	Lasso logistic lambda.min	Elastic net logistic lambda.min	Ridge logistic lambda.1se	Lasso logistic lambda.1se	Elastic net logistic lambda.1se
	DAY.5	-0.0287	-0.0291	-0.0229		-0.0024	-0.0214		
	DAY.6	-0.0733	-0.0729	-0.0582		-0.0448	-0.0508		
	DAY.7					Reference			
MONTH	MTH.1					Reference			
	MTH.2	0.0451	0.0445	0.1200		0.1049	0.1177		
	MTH.3	0.0278	0.0273	0.0959		0.0855	0.0919		
	MTH.4	-0.0586	-0.0584	0.0258		0.0033	0.0283		
	MTH.5	-0.0435	-0.0435	0.0371		0.0206	0.0397		
	MTH.6	-0.0817	-0.0812	0.0111			0.0179		
	MTH.7	-0.1725	-0.1713	-0.0600		-0.0603	-0.0463		
	MTH.8	-0.3056	-0.3035	-0.1772		-0.1998	-0.1535		
	MTH.9	-0.2576	-0.2553	-0.1313		-0.1416	-0.1101		
	MTH.10	-0.1009	-0.1004	-0.0052			0.0032		
	MTH.11	-0.1395	-0.1386	-0.0369		-0.0300	-0.0271		
	MTH.12	-0.1692	-0.1676	-0.0682		-0.0618	-0.0555		
YEAR	YEAR.1					Reference			
	YEAR.2	-0.1810	-0.1793	-0.0503		-0.0799	-0.0194		
	YEAR.3	-0.1251	-0.1241	-0.0074		-0.0328	0.0167		
	YEAR.4	-0.3551	-0.3520	-0.2068		-0.2598	-0.1662		
	YEAR.5	-0.3434	-0.3405	-0.2046		-0.2548	-0.1669		
HOUR	HOUR.1					Reference			
	HOUR.2	0.3119	0.2989	0.1563		0.1207	0.1371		
	HOUR.3	0.3107	0.3010	0.1306		0.0806	0.1082		
	HOUR.4	0.3119	0.2972	0.1543		0.1143	0.1363		
	HOUR.5	0.1432	0.1267	0.0170			0.0156		
	HOUR.6	0.1622	0.1455	0.0290		0.0189	0.0241		
	HOUR.7	0.0934	0.0773	-0.0259		-0.0185	-0.0243		
	HOUR.8	0.1159	0.0995	-0.0106		-0.0037	-0.0124		
	HOUR.9	0.1287	0.1124	0.0024			0.0008		
	HOUR.10	-0.0223	-0.0352	-0.1194		-0.1165	-0.1065		
	HOUR.11	0.0714	0.0608	-0.0578		-0.0125	-0.0570		
	HOUR.12	0.2333	0.2213	0.0970		0.0496	0.0875		
PRA	PRA.1					Reference			
	PRA.2	-0.1177	-0.1139	-0.1442		-0.0994	-0.1405		
	PRA.3	-0.0577	-0.0563	-0.0423		-0.0142	-0.0305		
	PRA.4	-0.2428	-0.2353	-0.1939		-0.1365	-0.1625		
GENAC	GENAC.1					Reference			
	GENAC.2	-0.0336	-0.0339	-0.0627		-0.0719	-0.0588		
	GENAC.3	-1.0275	-0.8903	-0.8028		-0.7573	-0.6904		
	GENAC.4	-0.1769	-0.1700	-0.1980		-0.1837	-0.1708		
	GENAC.5	0.0243	0.0269	0.0041			0.0075		
	GENAC.6	0.3482	0.3434	0.2663		0.2363	0.2451		
	GENAC.7	0.0647	0.0637	0.0187			0.0144		
SPECAC	SPECAC.1					Reference			
	SPECAC.2	-0.2140	-0.2127	-0.0238			-0.0112		
	SPECAC.3	-0.1531	-0.1535	0.0191			0.0297		
	SPECAC.4	-0.2072	-0.2060	-0.0170		-0.0030	-0.0069		
	SPECAC.5	-0.1832	-0.1820	0.0166			0.0272		
	SPECAC.6	-0.1916	-0.1906	0.0026			0.0150		
	SPECAC.7	-0.3771	-0.3735	-0.1524		-0.1568	-0.1233		
	SPECAC.8	-0.3296	-0.3273	-0.1087		-0.1203	-0.0876		
MODI	MODI.1					Reference			
	MODI.2	0.6744	0.6878	0.5999		0.4827	0.5497		
	MODI.3	0.0203	0.0154	-0.0188			-0.0171		
	MODI.4	-0.0237	-0.0281	-0.0770		-0.0262	-0.0788		
	MODI.5	-0.0180	-0.0228	-0.0785		-0.0204	-0.0891		
	MODI.6	0.2306	0.2239	0.2156	0.0061	0.2222	0.2079		0.0553
	MODI.7	0.0544	0.0541	0.0561		0.0001	0.0625		
	MODI.8	-0.3863	-0.3510	-0.3427		-0.1980	-0.3052		
	MODI.9	0.0745	0.0690	0.0396		0.0461	0.0343		
DEV	DEV.1					Reference			
	DEV.2	0.0684	0.0645	-0.0406		-0.0268	-0.0416		
	DEV.3	0.1326	0.1267	0.0136			0.0066		
	DEV.4	0.1585	0.1525	0.0547		0.0305	0.0484		
	DEV.5	0.2415	0.2352	0.1415		0.1191	0.1397		
	DEV.6	0.0207	0.0165	-0.0899		-0.0777	-0.0907		
	DEV.7	0.2490	0.2430	0.1228		0.1049	0.1102		
	DEV.8	0.2740	0.2782	0.0441			0.0043		
	DEV.9	0.0113	0.0063	-0.1068	-0.0426	-0.1082	-0.1107		-0.0749
MAT_DEV	MAT_DEV.1					Reference			
	MAT_DEV.2	-0.0691	-0.0685	-0.0778		-0.0651	-0.0730		
	MAT_DEV.3	0.0264	0.0266	0.0187			0.0216		
	MAT_DEV.4	0.0957	0.0977	0.0325			0.0135		
	MAT_DEV.5	-0.1715	-0.1599	-0.1741		-0.1042	-0.1595		
	MAT_DEV.6	0.3715	0.3801	0.2447		0.1989	0.2214		
	MAT_DEV.7	0.1238	0.1222	0.0834		0.0878	0.0669		
	MAT_DEV.8	0.2347	0.2326	0.2214		0.2400	0.2044		
	MAT_DEV.9	0.1262	0.1264	0.1077		0.0881	0.1052		
	MAT_DEV.10	-0.0134	-0.0118	0.0138			0.0235		
	MAT_DEV.11	-0.0080	-0.0083	-0.0386		-0.0002	-0.0430		
	MAT_DEV.12	-0.1314	-0.1051	-0.1771		-0.0216	-0.1751		
	MAT_DEV.13	0.4908	0.5203	0.3596		0.2958	0.3089		
	MAT_DEV.14	-0.5279	-0.4768	-0.4116		-0.3604	-0.3637		

Table 5. Coefficient estimates of the training data for different logistic regression models (continued)

Variable	Subcategory of Variable	Binary logistic	Firth's logistic	Ridge logistic lambda.min	Lasso logistic lambda.min	Elastic net logistic lambda.min	Ridge logistic lambda.1se	Lasso logistic lambda.1se	Elastic net logistic lambda.1se
	MAT_DEV.15	0.0981	0.1122	0.0436			0.0253		
	MAT_DEV.16	-0.3660	-0.3499	-0.2488		-0.1956	-0.1982		
	MAT_DEV.17	0.0987	0.1000	0.0747		0.0280	0.0685		
	MAT_DEV.18	-0.0431	-0.0436	-0.0563		-0.0425	-0.0569		
TINJ	TINJ.1					Reference			
	TINJ.2	1.5022	1.4869	1.3080	1.3321	1.4999	1.1894	1.2396	1.2594
	TINJ.3	0.4988	0.4943	0.4073	0.2574	0.5001	0.3490	0.1289	0.2573
	TINJ.4	0.9954	0.9830	0.8841		0.9465	0.8080		
	TINJ.5	0.9945	0.9893	0.5784		0.8566	0.4482		
	TINJ.6	0.7725	0.7640	0.5619		0.6587	0.4739		
	TINJ.7	-2.3372	-2.1740	-1.2040		-1.7795	-0.9727		
	TINJ.8	-0.7712	-0.5640	-0.7011		-0.4920	-0.6314		
	TINJ.9	-1.3893	-1.2226	-1.0243		-1.0357	-0.8826		
	TINJ.10	1.2233	1.2100	0.9930		1.1707	0.8764		0.2058
	TINJ.11	-0.0164	-0.0163	-0.0796			-0.1030		
PBINJ	PBINJ.1					Reference			
	PBINJ.2	0.4130	0.4180	0.0491		0.1852	-0.0165		
	PBINJ.3	0.6203	0.6158	0.2517		0.4455	0.1787		
	PBINJ.4	0.4497	0.4474	0.1118		0.2753	0.0546		
	PBINJ.5	0.8363	0.8289	0.4440	0.0732	0.6858	0.3534	1.2396	0.1220
	PBINJ.6	0.8077	0.8003	0.4193		0.6503	0.3332		0.0631
	PBINJ.7	0.7690	0.7629	0.3762		0.5959	0.2875		
	PBINJ.8	0.5778	0.5725	0.1826		0.4060	0.0976		

3.2 Prediction performance comparison

A comparison analysis was conducted between the standard and regularized logistic regression models that were used in this study to determine the best model to predict the severity of non-fatal construction accident data. Several model performance comparison criteria were utilized to compare the models. Table 6 depicts model comparison AIC values for these prediction models.

Table 6. AIC values for the prediction models

Model	AIC
Binary logistic	19,020
Firth's logistic	19,022
Ridge logistic $lambda.min$	19,126
Ridge logistic $lambda.1se$	19,206
Lasso logistic $lambda.min$	19,660
Lasso logistic $lambda.1se$	19,822
Elastic net logistic $lambda.min$	18,998
Elastic net logistic $lambda.1se$	19,598

According to the AIC values, Elastic net logistic with lambda.min has the lowest AIC score. Besides, it was also observed that there was a very minimal difference between the AIC values of the models. Therefore, it is understood that different model comparison criteria should also be considered.

Confusion matrices were used as another approach to evaluate the models' performances. The cut-off value was taken as 0.50, which is commonly used, to construct confusion matrices on the training and test data and to obtain the performance values of the standard and regularized logistic regression models. Table 7 provides the accuracy, balanced accuracy, F1-score, G-mean, ROC AUC and PR AUC values of each model for both the training and test dataset. These metrics are selected because they are appropriate for binary classification [51-56].

The rates of correctly predicting the case of losing more than 3 working days are 27.38% $(1329/1329+3525=27.38)$ and 25.14% $(523/523+1557=25.14)$ on the training and test data, respectively. However, when it comes to 3 or fewer working days, the rates are 93.48% $(12439/12439+868=93.48)$ on the training set and 93.48% $(5331/5331+372=93.48)$ on the test set. Thus, although the accuracy rates of the standard binary logistic regression model are high in the confusion matrices obtained on both training and test data, the rates of correctly predicting the positive class are lower than those of negative ones. This result is due to the imbalanced distribution among the categories of the dependent variable "LWD". It is, thus, preferable to assess models through balanced accuracy, F1-score, G-mean and PR AUC values instead of the accuracy and ROC AUC values commonly used.

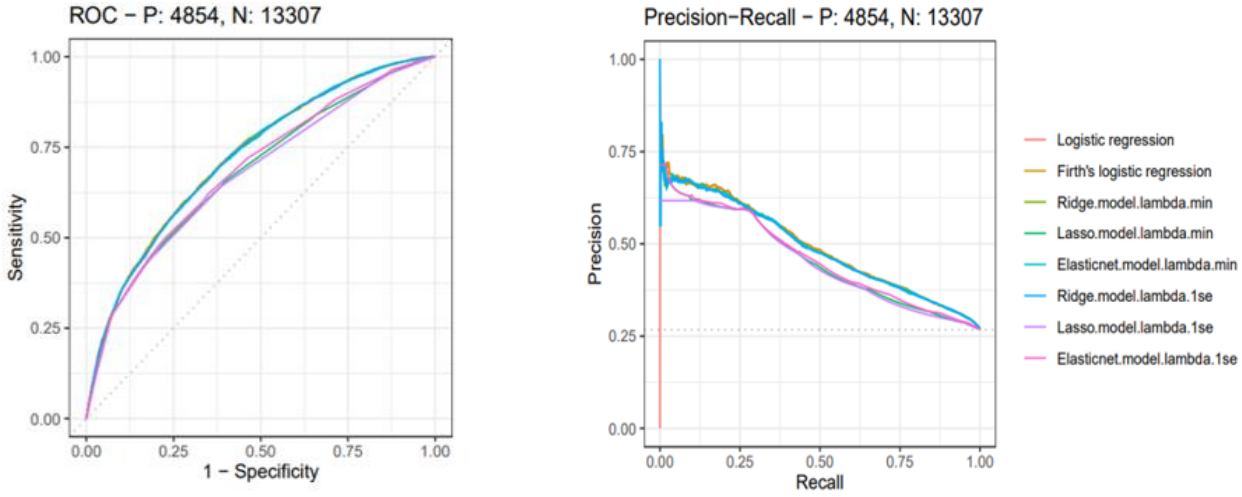
The balanced accuracy value obtained is lower than the accuracy value since the weights of the positive and negative classes of the dependent variable are taken to be equal. Besides, F1, G-mean and PR AUC values focus on the positive class that is less frequently in the dependent variable and thus, they give a more unbiased result like the balanced accuracy metric.

It was observed that the performance results on training and test data sets were close to each other and were good. This means that there is no overfitting or underfitting problem and a successful prediction can be obtained when new data are used in the prediction models. According to the performance results on the training data, it is determined that Firth's logistic regression model had better performance values (bolded) in predicting the LWD. However, this case changes on the test data based on the performance metrics. The performance results of the test set show that while the Lasso logistic with lambda.min fits the data better in terms of balanced accuracy, F1-score, and G-mean values, the Elastic net logistic with lambda.min gives better result based on the PR AUC performance metric.

Furthermore, the ROC AUC and PR AUC curves for comparing all models visually in the training and test data sets are depicted in Figure 2 and Figure 3.

Table 7. Performance measurements for the prediction models

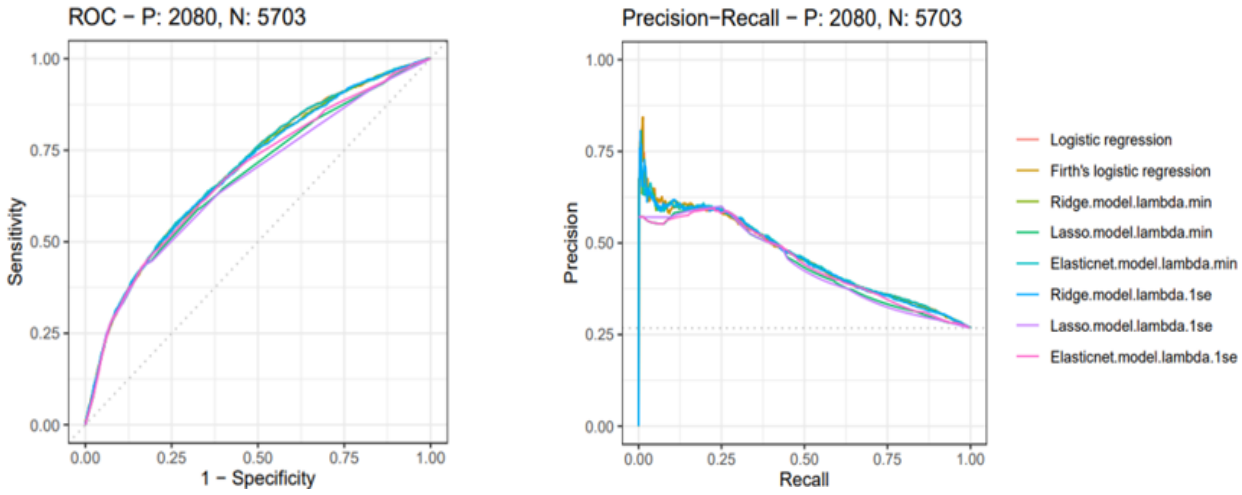
Data set	Model	Confusion matrix TN / FN	Confusion matrix FP / TP	Accuracy	Balanced Accuracy	F1-score	G-mean	ROC AUC	PR AUC
Train set	Binary logistic	12439	3525 / 868 1329	0.7581	0.6043	0.3770	0.5059	0.7235	0.4934
	Firth's logistic	12438	3523 / 869 1331	**0.7582**	**0.6045**	**0.3774**	0.5063	**0.7236**	**0.4936**
	Ridge logistic. $lambda.min$	12597	3697 / 710 1157	0.7573	0.5925	0.3443	0.4750	0.7217	0.4899
	Ridge logistic. $lambda.1se$	12703	3840 / 604 1014	0.7553	0.5818	0.3134	0.4466	0.7199	0.4881
	Lasso logistic. $lambda.min$	12380	3511 / 927 1343	0.7556	0.6035	0.3771	**0.5074**	0.6813	0.4575
	Lasso logistic. $lambda.1se$	12431	3578 / 876 1276	0.7547	0.5985	0.3643	0.4956	0.6748	0.4500
	Elastic net logistic. $lambda.min$	12471	3568 / 836 1286	0.7575	0.6010	0.3687	0.4983	0.7226	0.4907
	Elastic net logistic. $lambda.1se$	12504	3682 / 803 1172	0.7530	0.5906	0.3433	0.4764	0.6885	0.4622
Test set	Binary logistic	5331	1557 / 372 523	0.7522	0.5931	0.3516	0.4848	0.6969	0.4622
	Firth's logistic	5329	1557 / 374 523	0.7519	0.5929	0.3514	0.4847	0.6970	0.4623
	Ridge logistic. $lambda.min$	5394	1627 / 309 453	0.7513	0.5818	0.3188	0.4539	0.6957	0.4613
	Ridge logistic. $lambda.1se$	5435	1679 / 268 401	0.7498	0.5729	0.2917	0.4286	0.6942	0.4607
	Lasso logistic. $lambda.min$	5322	1539 / 381 541	0.7533	**0.5966**	**0.3604**	**0.4927**	0.6739	0.4427
	Lasso logistic. $lambda.1se$	5353	1557 / 350 523	**0.7550**	0.5950	0.3542	0.4858	0.6667	0.4398
	Elastic net logistic. $lambda.min$	5352	1574 / 351 506	0.7527	0.5909	0.3446	0.4778	**0.6993**	**0.4631**
	Elastic net logistic. $lambda.1se$	5377	1603 / 326 477	0.7522	0.5861	0.3309	0.4650	0.6831	0.4469



(a) (b)

Figure. 2. Comparison of (a) ROC curve (b) PR curve on the training data set

As it can be seen from the graphs, the ROC AUC values of all models are higher than the baseline random prediction value of 0.50. Additionally, the baseline random prediction value for the PR curve was computed as 0.2672 by dividing the actual positive events by the sum of actual positive and negative events ($[4854/(4854+13307)]$) for training and $[2080/(2080+5703)]$ for test sets) in this study. The PR AUC values of all models are greater than 0.2672 on both the training and test sets. Thus, the results indicate that the performance of the models constructed for LWD is good.



(a) (b)

Figure. 3. Comparison of (a) ROC curve (b) PR curve on the test data set

All in all, using the training data to construct models, and test data to validate the models built, Firth's logistic model can be proposed as the best alternative for prediction purposes after all the results have been considered and interpreted.

In the "Victim-related variable group, within the age (AGE_GR), education (EDU), and vocational education (VOCEDU) subcategories, the relationship between LWD and the variables of age groups of 45-54 (AGE_GR.4), secondary school (EDU.4) and high school graduates (EDU.5), and lack of vocational education (VOCEDU.2) are positive, respectively. It means that if one of related variable is chosen and others held constant (taking the value of 0) in the model, the probability of having more than three LWD accidents will increase with the chosen variable. In the AGE_GR subcategory, the probability of having more than three LWD accidents in AGE_GR.4 was 1.2482 times higher than in the age range between 14 and 24 (AGE_GR.1). For the EDU subcategory, EDU.4 and EDU.5 were 1.3227 and 1.3996 times, respectively, higher than workers that had an undergraduate degree (EDU.6). Besides, VOCEDU.2 was 1.1014 times more likely than for vocational training (VOCEDU.1). The fact that some physiological states seen in advanced ages negatively affect the healing time of injuries, most construction workforce in Türkiye has high school diploma or less and starts to work without completing their vocational training or has problems in understanding and applying the vocational education given due to low education level may cause an increase in the LWD.

In the "Workplace-related" variable group, within the size of the workplace (SIZW) and working environment (WOE) subcategories, there is a negative relationship between the LWD and the variables of having 1000 or above workers in the workplace (SIZW.9), the working environments of construction site, construction, opencast quarry and mine (WOE.2), health establishment (WOE.5) and underground (WOE.10). Thus, the probability of having greater than three LWD will reduce with SIZW.9, WOE.2, WOE.5, and WOE.10. On the other hand, within the project type (PRT) subcategory, there is a positive relationship between LWD and the variables of construction of roads and railways (PRT.2) and construction of utility projects (PRT.3). The probability of having more than three LWD will increase with PRT.2 and PRT.3. In the SIZW subcategory, the probability of having more than three LWD accidents in SIZW.9 vs. the reference category of 10 or fewer workers (SIZW.1) was 1.5013 $(1/Exp(-0.4063))$ times lower. For the PRT subcategory, PRT.2 and PRT.3 were 1.2370 and 1.1854 times, respectively, higher than construction of building (PRT.1). This probability in the WOE.2, WOE.5 and WOE.10 were 1.2156 $(1/Exp(-0.1952))$, 6.1252 $(1/Exp(-1.8124))$, and 3.8636 $(1/Exp(-1.3513))$ times, respectively, less likely than for industrial site (WOE.1). The fact that most of construction companies operating in the construction industry are small and medium-sized and there is an increase in roads and railways and utility projects in the region compared to other projects can be causes of these results.

Under the "Accident time-relate" variable group, within the month (MTH) and year (YEAR) subcategories, the relationship between LWD and the variables of month of August (MTH.8), the month of September (MTH.9), 2014 (YEAR.2), 2016 (YEAR.4) and 2017 (YEAR.5) are negative, respectively. The probability of experiencing more than three LWD will decrease with MTH.8, MTH.9, YEAR.2, YEAR.4, and YEAR.4. In the MTH subcategory, the probability of experiencing more than three LWD accidents in MTH.8 and MTH.9 were 1.3546 $(1/Exp(-0.3035))$ and 1.2909 $(1/Exp(-0.2553))$ times, respectively, less than January (MTH.1). In the YEAR subcategory, the probability for YEAR.2, YEAR.4 and YEAR.5 was 1.1964 $(1/Exp(-0.1793))$, 1.4219 $(1/Exp(-0.3520))$, and 1.4057 $(1/Exp(-0.3405))$ times, respectively lower than the year of 2013 (YEAR.1). These outcomes may have resulted from the low number of construction projects in the relevant months and the effect

of safety precautions taken in construction workplaces due to sanctions and occupational safety awareness over time.

In the "Accident and sequence of events-related" variable group, within the general activity (GENAC) and material agent of the deviation (MAT_DEV) subcategories, there is positive a relationship between LWD and the variables of movement-related activity (GENAC.6) and machines and equipment (MAT_DEV.8). The probability of having more than three LWD will increase with GENAC.6 and MAT_DEV.8. On the contrary, within the specific acitiviy (SPECAC) subcategory, there is a negative relationship between LWD and the variables of activity of presence (SPECAC.7) and other specific physical activities (SPECAC.8). In the GENAC subcategory, the probability of three LWD accidents in GENAC.6 was 1.4098 times more likely than the production, manufacturing, processing, and storing activity (GENAC.1). Besides, this probability in SPECAC.7 and SPECAC.8 vs. the operating machine activity (SPECAC.1) were 1.4528 ($1/Exp(-0.3735)$), and 1.3872 ($1/Exp(-0.3273)$) times lower, respectively. For the MAT_DEV subcategory, MAT_DEV.8 had 1.2618 times riskier than without having material agent (MAT_DEV.1). This result may have arisen from the occurrence of more frequent accidents in movement and operationing machines and equipments.

In the "Post-accident state-related" variable group, within the type of injury (TINJ) and part of body injured (PBINJ) subcategories, the relationship between LWD and the variables of having bone fractures (TINJ.2), dislocations, sprains and strains (TINJ.3), traumatic amputations (TINJ.4), concussion and internal injuries (TINJ.5), burns, scalds and frostbites (TINJ.6), multiple injuries (TINJ.10), injury in neck part (PBINJ.2), back part (PBINJ.3), torso and organs (PBINJ.4), upper extremities (PBINJ.5), lower extremities (PBINJ.6), whole body and multiple sites (PBINJ.7) and other parts of the body injured (PBINJ.8) are positive, respectively. The probability of having more than three LWD will increase with TINJ.2, TINJ.3, TINJ.4, TINJ.5, TINJ.6, TINJ.10, PBINJ.2, PBINJ.3, PBINJ.4, PBINJ.5, PBINJ.6, PBINJ.7 and PBINJ.8. Conversely, within the TINJ subcategory, the relationship between LWD and the variables of poisoning and infections (TINJ.7), getting shocked (TINJ.9) are negative. The probability of experiencing more than three LWD will reduce with TINJ.7 and TINJ.9. In the TINJ subcategory, the probability of experiencing more than three LWD accidents occasioned by accidents related to TINJ.2, TINJ.3, TINJ.4, TINJ.5, TINJ.6 and TINJ.10 were 4.4234, 1.6394, 2.6725, 2.6894, 2,1468 and 3.3533 times, respectively, more likely than those related to wounds and superficial injuries (TINJ.1). On the other hand, the probability for TINJ.7 was 8.7936 ($1/Exp(-2.1740)$) and for TINJ.9 ($1/Exp(-1.2226)$) was 3.3958 times lower than TINJ.1. For the PBINJ subcategory, the probability of experiencing more than three LWD accidents increased in all subcategories compared with the head (PBINJ.1). This probability in PBINJ.2, PBINJ.3, PBINJ.4, PBINJ.5, PBINJ.6, PBINJ.7 and PBINJ.8 were 1.5189,1.8512, 1.5642, 2.2907, 2.2263, 2.1445 and 1.7728 times, respectively, higher than for PBINJ.1. The injury type and its location on the body has a great impact on the LWD since they affect the healing time and back to work status.

In general, increasing age, low education level, problems in obtaining vocational education due to low education level, increasing utility and road and railway projects in recent years, moving with vehicles, and use of machines and equipments are considered as cases that increase the occurrence risk of occupational accidents in the literature. Compared to the findings of this study, it is not surprising that in the presence of the stated cases, the probability of having more than three LWD accidents is increased. Besides, considering the effect of the safety precautions taken, it is expected that the probability of LWD decreases as time passes. The top determinants for having more than three days LWD at construction industry due to construction accidents are related to post-accident state. It can be concluded that the greater the impact of the occupational accident, the higher the loss of working days. However, in this study, contrary to what is known, it has been determined that despite the increases in the number of workers and construction projects, increasing workplace size, working in the summer months and around the construction site reduce the probability of having more than three LWD accidents. The severity of non-fatal injuries would be diminished in construction industry by focusing on the related variables that significantly trigger to the occurrence of LWD. In this direction, we can apply proactive precautions such as providing and using right and sufficient personal protective equipment, enhancing the content of the training, and teaching style.

4 Conclusions

Construction activities remain one of the most hazardous industries worldwide. Therefore, standard and regularized logistic regression approaches as machine learning classification algorithms are applied to national construction accident data and the results obtained are compared in this study. Based on the model results, it can be concluded that the type of injury and the body part injured have a significant impact on the occurrence of occupational accidents resulting in more than LWD. The prediction performance of all models for non-fatal accidents with more than three LWD is good, with only slight differences in their performance values. The Ridge logistic regression model does not reduce any coefficients to zero, making it difficult to interpret the constructed model. However, Lasso and Elastic net estimators have variable selection capabilities. When all the models are compared, Firth's logistic model was found to display the best performance in predicting the LWD resulting from non-fatal occupational accidents in the training set. However, this shouldn't imply the overall superiority of the model in all cases of occupational accidents phenomenon, and additional care should be given to each dataset to understand its nature and identify any problems in the data. When there is no regularization requirement, standard binary logistic regression can be used. However, some form of regularization is usually necessary, particularly with large

sample sizes and categorical data. Therefore, other prediction approaches should be considered. Firth's logistic regression model is a good alternative for reducing bias caused by imbalance problems, while the Ridge logistic regression is useful in the event of a multicollinearity problem, and all variables are necessary for the determination of presence of a relationship with LWD. Lasso and Elastic net logistic become the tools of choice in cases of multicollinearity issues and when there is a need to eliminate irrelevant variables. In this way, the best model which fits the data well can be determined.

This study analyzes, for the first time, occupational accidents that occurred in the "construction of buildings", "civil engineering" and "specialized construction activities" sectors in the Central Anatolia region. The study considers accident variables that have not been previously covered in the literature. Moreover, this study demonstrates how standard binary logistic regression and regularized logistic regression models can be applied in a machine learning classification context in a large categorical occupational accident dataset. The study has the potential to advance the current knowledge of data analysis techniques for predicting the severity of non-fatal construction accidents using more innovative and interpretable machine learning tools. The use of regularized models on occupational accidents in the study opens new doors for researchers working in this field. Additionally, this study also shines a spotlight to the OSH professionals responsible for the implementation of OSH activities in workplaces in performing such analysis to model their historical accident records. Furthermore, the findings of this study provide vital information for assessing the occurrence of LWD risk at construction industry. These findings can be used to develop more appropriate safety precautions in the construction industry.

As mentioned above, Firth's logistic decreases bias, while Ridge, Lasso and Elastic net logistic models stabilizes the prediction in case of multicollinearity. However, none of the models deal with both problems. Future studies could develop a double regularized model integrating Firth's logistic regression with a ridge, lasso, or elastic net parameter. As a potential next step, synthetic data could be generated to increase the performance of the models applied. In this study, dummy-coded categorical variables were used to build all models, but different coding strategies, such as one-hot encoding and contrast coding could be used to reveal how the coding choices affect the prediction results. Additionally, the occupational accident data used in this study could be subjected to analysis for prediction purposes using other machine learning methods, and the models in this study could also be applied in different fields.

Acknowledgement

Authors would like to thank SSI for providing the data. This research is supported by the Scientific Research Projects Committee of Eskisehir Technical University under the Doctor of Philosophy dissertation grant 1705F427. This work is also a part of Ph.D. thesis, which is named as 'Prediction of occupational accidents, risk analysis and prioritization of precautions to be taken by house of quality in construction industry' [59] prepared by Şura Toptancı and supervised by Nihal Erginel and Ilgın Acar.

Conflict of interest

The authors declare that there is no conflict of interest.

Similarity rate (iThenticate): 19%

References

[1] J.P. Leigh, S.B. Markowitz, M. Fahs and P. Landrigan, Costs of Occupational Injuries and Illnesses. University of Michigan Press, Ann Arbor, Mich., 2000.

[2] L.I. Boden, E.A. Biddle and E.A. Spieler, Social and economic impacts of workplace illness and injury: current and future directions for research. American Journal of Industrial Medicine, 40, 398-402, 2001. https://doi.org/10.1002/ajim.10013.

[3] S. Linacre, Australian social trends 2007. Australian Bureau of Statistics. ABS catalogue no. 4102. https://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/3550D34DA999401ECA25748E00126282?opendocument, Accessed 01 March 2021.

[4] D.L. Goetsch, Occupational Safety and Health for Technologists, Engineers, and Managers. 6th ed. Pearson Prentice Hall, 2008.

[5] International Labour Organization (ILO), Databases and subjects: labour force by sex and age-2020. https://ilostat.ilo.org/topics/population-and-labour-force/#, Accessed 23 May 2021.

[6] International Labour Organization (ILO), Safety and health at work. https://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.htm , Accessed 5 December 2020.

[7] P. Hämäläinen, J. Takala and T.B. Kiat, Global estimates of occupational accidents and work-related illnesses 2017. https://www.icohweb.org/site/images/news/pdf/Report%20Global%20Estimates%20of%20Occupational%20Accidents%20and%20Work-related%20Illnesses%202017%20rev1.pdf, Accessed 5 December 2020.

[8] Bureau of Labor Statistics (BLS), Census of fatal occupational injuries - 2018. https://www.bls.gov/iif/oshcfoi1.htm, Accessed 7 December 2020.

[9] SGK, SGK İstatistik Yıllıkları- 2018. http://www.sgk.gov.tr/wps/portal/sgk/tr/kurumsal/istatistik/sgk_istatistik_yilliklari, 8 December 2020.

[10] H. Cakan, Analysis and modeling of roofer and steel worker fall accidents. Ph.D. Thesis, Wayne State University, Michigan, USA, 2012.

[11] S. Onder, Evaluation of occupational injuries with lost days among opencast coal mine workers through logistic regression models. Safety Science, 59, 86-92, 2013, http://dx.doi.org/10.1016/j.ssci.2013.05.002.

[12] Ö. Akboga, Modeling of construction accident severity using logistic regression. Ph.D. Thesis, Ege University, İzmir, Türkiye, 2014.

[13] A. Bilim, Analysis and modeling of occupational accidents occurring in highway and railway constructions. Ph.D. Thesis, Konya Technic University, Konya, Türkiye, 2018.
[14] A.J. Tixier, M.R. Hallowell, B. Rajagopalan and D. Bowman, Application of machine learning to construction injury prediction. Automation in Construction, 69, 102-114, 2016. https://doi.org/10.1016/j.autcon.2016.05.016.
[15] K. Yang, C.R. Ahn, M.C. Vuran and S.S. Aria, Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit. Automation in Construction, 68, 194-202, 2016. https://doi.org/10.1016/j.autcon.2016.04.007.
[16] K. Kang and H. Ryu, Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science, 120:226-236, 2019. https://doi.org/10.1016/j.ssci.2019.06.034.
[17] B.U. Ayhan and O.B. Tokdemir, Safety assessment in megaprojects using artificial intelligence. Safety Science, 118:273-287, 2019. https://doi.org/10.1016/j.ssci.2019.05.027.
[18] J.Y. Lee, Y.G. Yoon, T.K. Oh, S. Park and S.I. Ryu, A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry. Applied Sciences, 10(21), 7949, 2020. https://doi.org/10.3390/app10217949.
[19] J. Choi, B. Gu, S. Chin and J.S. Lee, Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, 102974, 2020. https://doi.org/10.1016/j.autcon.2019.102974.
[20] F. Recal and T. Demirel, Comparison of machine learning methods in predicting binary and multi-class occupational accident severity. Journal of Intelligent & Fuzzy Systems, 40(6), 10981-10998, 2021. https://doi.org/10.3233/JIFS-202099.
[21] Y.Ö. Tetik, Ö. Akboğa Kale, I. Bayram and S. Baradan, Applying decision tree algorithm to explore occupational injuries in the Turkish construction industry. Journal of Engineering Research, 10(3), 59-70, 2022. https://doi.org/10.36909/jer.12209.
[22] K. Koc, Ö. Ekmekcioğlu and A.P. Gurgun, Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. Engineering, Construction and Architectural Management, (ahead-of-print), 2022. https://doi.org/10.1108/ECAM-04-2022-0305.
[23] J.M. Pereira, M. Basto and A.F. da Silva, The logistic lasso and ridge regression in predicting corporate failure. Procedia Economics and Finance, 39, 634-641, 2016. https://doi.org/10.1016/S2212-5671(16)30310-0.
[24] R. Gavanji, Penalized regression methods for modelling rare events data with application to occupational injury study. Master Thesis, University of Saskatchewan, Canada, 2019.
[25] S. Sarkar and J. Maiti, Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. Safety science, 131, 104900, 2020. https://doi.org/10.1016/j.ssci.2020.104900.
[26] M. Gonzalez-Delgado, H. Gómez-Dantés, J.A. Fernández-Niño, E. Robles, V.H. Borja and M. Aguilar, Factors associated with fatal occupational accidents among Mexican workers: a national analysis. PloS one, 2015. https://doi.org/10.1371/journal.pone.0121490.
[27] S.S. Uysal, Comparison of The Logistic Elastic Net Method with Alternative Methods. Master Thesis, Eskisehir Osmangazi University, Türkiye, 2020.
[28] V. Gallego, A. Sánchez, I. Martón and S. Martorell, Analysis of occupational accidents in Spain using shrinkage regression methods. Safety Science, 133, 105000, 2021. https://doi.org/10.1016/j.ssci.2020.105000.
[29] D.W. Hosmer, S. Lemeshow and R.X. Sturdivant, Applied Logistic regression. 3rd ed. Hoboken, New Jersey: Wiley, 2013.
[30] A. Agresti, An introduction to categorical data analysis. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2019.
[31] S.A. Czepiel, Maximum likelihood estimation of logistic regression models: theory and implementation. czep. net/stat/mlelr.pdf. Accessed 20 August 2022.
[32] B.G. Tabachnick and L.S. Fidell, Using multivariate statistics. 6th ed. Boston: Pearson, 2013.
[33] G. Kemalbay and B.N. Alkış, Borsa endeks hareket yönünün çoklu lojistik regresyon ve k-en yakın komşu algoritması ile tahmini. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 27(4), 556-569, 2020. https://doi.org/10.5505/pajes.2020.57383.
[34] D.N. Gujarati and D.C. Porter, Basic econometrics. 5th ed. New York: McGraw-Hill/Irwin, 2009.
[35] G. King and L. Zeng, Logistic regression in rare events data. Political Analysis, 9, 137-163, 2001. https://doi.org/10.1093/oxfordjournals.pan.a004868.
[36] C.F. İşçen, S.S. Uysal and A.A. Yavuz, Su kalitesi değişimine etki eden değişkenlerin lojistik regresyon, lojistik-ridge ve lojistik lasso yöntemleri ile tespiti. Biyoloji Bilimleri Araştırma Dergisi, 14(1), 1-12, 2021. https://bibad.gen.tr/ index.php/bibad/article/view/375.
[37] Z.Y. Algamal and M.H. Lee, Applying penalized binary logistic regression with correlation based elastic net for variables selection. Journal of Modern Applied Statistical Methods, 14(1), 168-179, 2015. https://doi.org/10.22237/jmasm/1430453640.
[38] S. Doerken, M. Avalos, E. Lagarde and M. Schumacher, Penalized logistic regression with low prevalence exposures beyond high dimensional settings. PLoS One, 14(5), e0217057, 2019. https://doi.org/10.1371/journal.pone.0217057.
[39] D. Firth, Bias reduction of maximum likelihood estimates. Biometrika, 80, 27–38, 1993. https://doi.org/10.2307/2336755.

[40] M.S. Rahman and M. Sultana, Performance of Firth- and logF-type penalized methods in risk prediction for small or sparse binary data. BMC Medical Research Methodology, 17(1), 1-15, 2017. https://doi.org/10.1186/s12874-017-0313-9.

[41] R.L. Schaefer, L.D. Roi and R.A. Wolfe, A ridge logistic estimator. Communications in Statistics - Theory and Methods, 13(1), 99-113, 1984. https://doi.org/10.1080/03610928408828664.

[42] D.E. Duffy and T.J. Santne, On the Small Sample Properties of Norm-Restricted Maximum Likelihood Estimators for Logistic Regression Models. Communications in Statistics - Theory and Methods, 18, 959-980, 1989. https://doi.org/10.1080/03610928908829944.

[43] S. Le Cessie and J.C. Van Houwelingen, Ridge estimators in logistic regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41(1), 191-201, 1992. https://doi.org/10.2307/2347628.

[44] W.J. Fu, Penalized regressions: the bridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3), 397-416, 1998. https://doi.org/10.2307/1390712.

[45] G. James, D. Witten, T. Hastie, and R. Tibshirani, Linear model selection and regularization. In: An Introduction to Statistical Learning, Springer Text in Statistics, , 225-288, Springer, New York, NY, 2021.

[46] R. Tibshirani, Regression Shrinkage and Selection via the LASSO. Journal of Royal Statistical Society B, 58, 267–288, 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[47] H. Zou and T. Hastie, Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301– 320, 2005. https://doi.org /10.1111/j.1467-9868.2005.00503.x.

[48] T. Hastie, R. Tibshirani and M. Wainwright, Statistical learning with sparsity. Monographs on statistics and applied probability, 143, CRC Press, 2015.

[49] B. Jason, A Gentle Introduction to k-fold Cross-Validation. https: //machinelearningmastery.com/k-fold-cross-validation/, Accessed 1 August 2021.

[50] T. Hastie, R. Tibshirani, and J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2001.

[51] M. Grandini, E. Bagli and G. Visani, Metrics for multi-class classification: an Overview. A white paper, https://arxiv.org/pdf/2008.05756.pdf, Accessed 8 May 2021.

[52] T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432, 2015. https://doi.org/10.1371/journal.pone.0118432.

[53] D. Chicco, N. Tötsch and G. Jurma, The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix. BioData Mining, 14(13), 1-22, 2021. https://doi.org/10.1186/s13040-021-00244-z.

[54] H. Guo, H. Liu, C. Wu, W. Zhi, Y. Xiao and W. She, Logistic discrimination based on G-mean and F-measure for imbalanced problem. Journal of Intelligent & Fuzzy Systems, 31(3), 1155-1166, 2016. https://doi.org/10.3233/IFS-162150.

[55] J. Davis and M. Goadrich, The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning, pp. 233-240, Pittsburgh, PA, 2006.

[56] H.R. Sofaer, J.A. Hoeting and C.S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events. Methods in Ecology and Evolution, 10(4), 565-577, 2019. https://doi.org/10.1111/2041-210X.13140.

[57] Eurostat. European Statistics on Accidents at Work (ESAW), Summary methodology 2013. https://ec.europa.eu/eurostat/documents/3859598/5926181/KS-RA-12-102-EN.PDF/56cd35ba-1e8a-4af3-9f9a-b3c47611ff1c, Accessed 1 August 2020.

[58] W. Jiang, P. Lakshminarayanan, X. Hui, P. Han, Z. Cheng, M. Bowers, I. Shpitser, S. Siddiqui, R.H. Taylor, H. Quon and T. McNutt, Machine learning methods uncover radiomorphologic dose patterns in salivary glands that predict xerostomia in patients with head and neck cancer. Advances in radiation oncology, 4(2), 401-412, 2019. https://doi.org/10.1016/j.adro.2018.11.008.

[59] Ş. Toptancı, Prediction of occupational accidents, risk analysis and prioritization of precautions to be taken by house of quality in construction industry (Doctoral dissertation). Available from Turkish Council of Higher Education Thesis Center (Thesis No: 694456), 2021.