Journal of Engineering, Project, and Production Management
2025, 15(1), 0002

A Multi-Output Regression Approach to Predicting Gender-Centric Workplace Fatal Accidents

Deepanshu Mahajan[1], Emel Sadikoglu[2], Uttam Kumar Pal[3], Saksham Timalsina[4], Chengyi Zhang[5], and Sevilay Demirkesen[6]

[1]Research Intern, Department of Civil Engineering, Gebze Technical University, Turkey, Email: mahajandeepanshu495@gmail.com

[2]Research Assistant, Department of Civil Engineering, Gebze Technical University, Turkey, Email: esadikoglu@gtu.edu.tr

[3]Graduate Research Assistant, Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, USA, Email: upal@uwyo.edu

[4]Graduate Research Assistant, Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, USA, Email: stimals1@uwyo.edu

[5]Associate Professor, Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, USA, Email: chengyi.zhang@uwyo.edu (corresponding author).

[6]Associate Professor, Department of Civil Engineering, Gebze Technical University, Turkey, E-mail: demirkesen@gtu.edu.tr

Project Management
Received January 31, 2024; received revision November 6, 2024; accepted November 7, 2024
Available online December 22, 2024

Abstract: The prevalence of accidents that raise substantial concerns, due to their potential to cause severe injuries, loss of life, and long-term health implications for workers across various industries highlights the urgent need for effective accident prevention measures. It is imperative to analyze accident characteristics and different features leading to accidents, hence, this study investigates the involvement of men and women in different industry accidents and the contributing factors. The number of men and women involved in accidents is examined through the application of linear regression and tree-based methods, including Random Forest, Extra Trees, and XGBoost Regressors, within the context of constrained multi-output regression based on the data from the year 1992 to 2015. A model that predicts gender-related accidents for the year 2016 to 2021 has been developed using historical data and its accuracy is compared to the actual data from that time. The error calculation has been determined using the Mean Average Percentage Error (MAPE) method, the results showing an error of 2.57%, 9.73%, and 2.86% in the prediction of the number of accidents for males, females, and total workers, respectively. Moreover, the feature importance analysis is conducted to identify the key factors influencing the number of accidents for both genders. The analysis reveals that the number of wage and salary workers have the highest importance for both genders, followed by the number of working-age and under-age workers for accident prediction of male and female workers, respectively. This study aims to gain insights into the underlying factors contributing to accident occurrence and enhance the understanding of accident involvement among men and women. The study results can contribute to the development of effective strategies for accident mitigation, worker protection, and the promotion of safer working environments by predicting accidents using machine learning models.

Keywords: Workplace accident, accident prediction, occupational safety, gender-based accident, mean average percentage error approach, feature importance

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2025-0002

1. Introduction

Accidents in the workplace pose significant risks to individuals and can have substantial implications for industries and society as a whole. According to the U.S. Bureau of Labor Statistics (BLS) data from 2022, 2.8 million non-fatal workplace injuries and illnesses were reported, and the injury case rate was stated as 2.3 cases per 100 full-time equivalent workers (BLS, 2022). Underreporting of occupational incidents, including fatalities, gives a false picture of the real problem. Given the significant impacts, numerous strategies for enhancing occupational health and safety are required. To develop proper

occupational health and safety strategies, it is crucial to examine the fundamental aspects linked to occupational injuries and accidents.

There are several factors associated with occupational injuries and accidents including individual-related, job-related, workplace-related, and organization-related (Khanzode et al., 2012; Park et al., 2019). The most investigated factors are sociodemographic, including age, sex, and occupational experience (Kriebel, 1982). According to Bena et al. (2013), the risk of injury begins high for young workers with limited experience and when sufficient work experience is gained, the risk decreases, however, the risk rises again for older workers (Jones et al., 2013). Some common job-related factors contributing to injuries include occupation type, activity type, work location, workplace attributes, shift of working, and the nature of worker-production interactions (Gonzalez-Delgado et al., 2015). Considering organization-related factors, higher management and coworker support, higher safety commitment, perception of a safer workplace, and smaller workgroup sizes are recognized to be associated with low injury risk (Khanzode et al., 2012).

This study focuses on individual and job-related factors. Understanding the factors that contribute to accidents and identifying their impact on different gender groups is crucial for developing effective safety measures and mitigating potential hazards (Park et al., 2019; Yosef et al., 2023). A gender-specific analysis can help define the problems and improve occupational health and safety management (Forssberg et al., 2022). This research study aims to investigate the involvement of men and women in various industry accidents and analyze the contributing factors that influence such accidents. Further, referring to the study conducted by Kriebel (1982), considerations such as age and employment status have been identified as potential contributors to accidents. By examining the relationship between these factors and accidents, this study seeks to gain insights into their respective impacts and provide valuable information for accident prevention strategies.

Machine learning techniques have been effective in determining the factors contributing to accidents (Lee et al., 2020). It is essential to recognize that men and women may face distinct workplace challenges and exhibit varying vulnerability to accidents. This study focuses on analyzing the involvement of men and women in accidents separately to identify any gender-specific patterns and differences using machine learning. In the subsequent sections of the paper, the methodology employed for data collection and analysis is presented. The results of the study are provided, highlighting the key findings related to the association of men and women in industry accidents. Finally, the implications of the findings and recommendations for future research and safety interventions are mentioned.

2. Research Procedures and Approach

The selection of an appropriate data set and its preprocessing are crucial steps in model development. In this study, a constrained multi-output linear regression model has been developed using the constraint equation. Tree-based regression models including Random Forest, XGBoost Regressor, and Extra Tree Regressor models have been employed to enhance the accuracy of the model. Furthermore, an optimization process based on the Sequential Least Squares Quadratic Programming (SLSQP) algorithm has been utilized to ensure that the mean squared error is minimized while satisfying the custom constraint function. The model has been evaluated using standard regression metrics on both the training and test datasets.

It is noteworthy that the model's purpose is to predict the number of men and women involved in accidents, considering the workers' age and employment status, and to quantify the importance of each feature variable contributing to the accident. Fig. 1. shows the flowchart of steps involved in machine learning model development outlining actions involved in the constrained multi-output regression process for predicting the number of men and women involved in accidents.

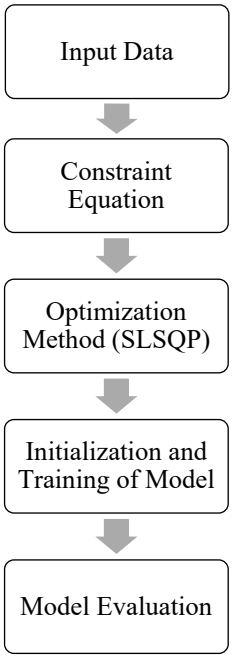
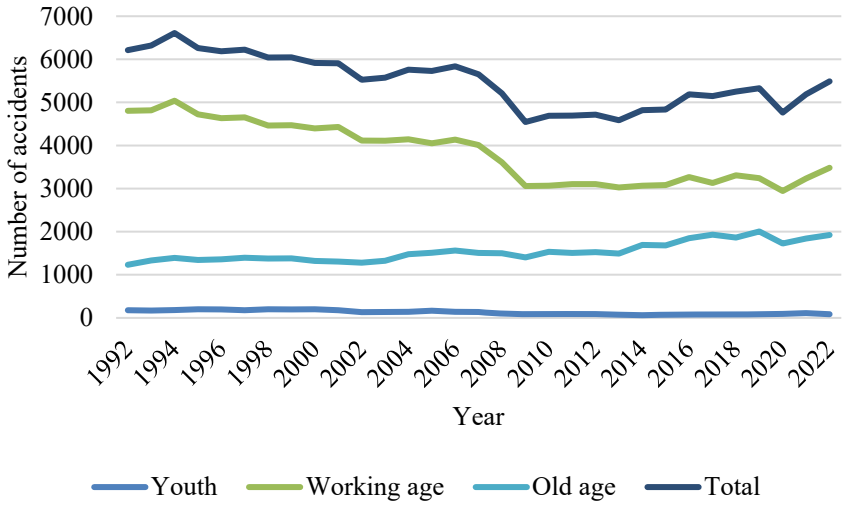
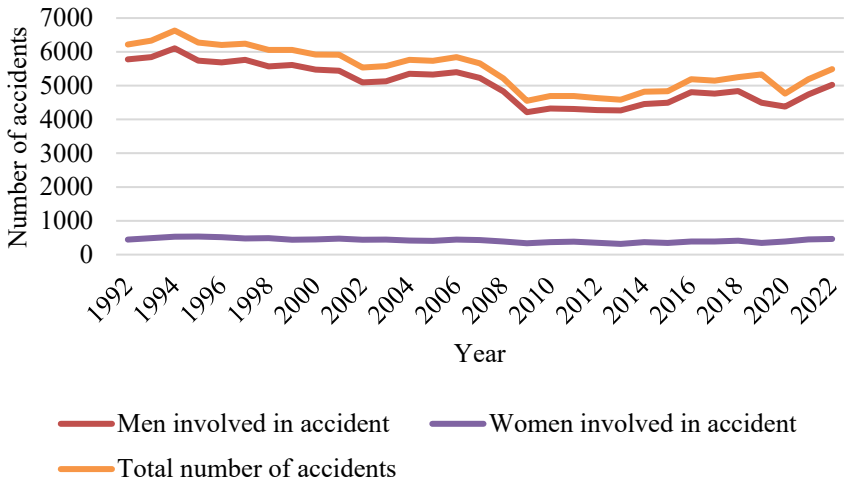


Fig. 1. Flowchart of steps involved in model development and evaluation

(a)



(b)



(c)

Fig. 2. (a) Accident distribution based on employment status, **(b)** Accident distribution based on the working age, and **(c)** Accident distribution based on gender

2.1. Dataset Description

The data was retrieved from the US Bureau of Labor Statistics website, this data spanning 1992 to 2022 presents information related to industrial fatal accidents, covering a wide range of industries including construction, transportation and

warehousing, manufacturing, agriculture, forestry, and others. Fig. 2 (a), Fig. 2 (b), and Fig. 2 (c) show the trend and accident distribution based on employment status, age group, and gender.

Data preprocessing was conducted before model development, and there are no missing values in the data set. The age variable was considered under three different categories, people below the age of 18 are considered youth or young age people, people older than 18 years, up to 60 years are classified as working age people, and people above 60 are classified as older age people. Table 1 shows the feature (independent) variables and the target (dependent) variables used in machine learning models.

Table 1. Feature and target variables

Number	Feature variable	Target variable
1	Wage and salary workers	Number of men involved
2	Self-employed	Number of women involved
3	Young age people	
4	Working age people	
5	Older age people	
6	Total number of accidents	

2.2. Model Development

Machine learning is useful in providing statistical models and algorithms to train and create models that can be used to make predictions and classifications (Maheronnaghsh et al., 2023). Machine learning models have gained significant popularity in various real-world applications, including document analysis and computer vision. Further, it has been utilized in health and safety studies and some application areas including identification, classification, prediction, monitoring, controlling, and prevention of safety risks (Dobrucalı et al., 2023). The multi-output linear regression model is a powerful statistical tool commonly used in predictive analysis to understand the relationships between multiple dependent variables and a set of explanatory variables (Xu et al., 2023). The general Machine Learning model is formulated using the Eq. (1) (Natarajan, 2014).

$$Y^i = f(X^i) \tag{1}$$

Where Yi is the number of either men or women and f(Xi) is the function of either gender those are involved in the accidents.

Constrained machine learning refers to a specific type of problem where the learned models are required to meet high accuracy standards and adhere to predefined constraints or limitations (Perez et al., 2021). Since the data includes some constraints, such as the sum of the number of men and women involved in the accident is equal to the total number of accidents, a constrained machine learning model is utilized. This constraint is introduced to enhance the accuracy and reliability of the predictive outcomes, ensuring that the model aligns with the inherent relationship between the total accident count and the respective contributions of men and women. Incorporating constraints during the process of learning patterns and rules can be highly beneficial. A variety of specialized systems and techniques have been created to address constraint-based problems, based on the study conducted by (De Raedt et al., 2010). The constraint equation enforces the constraint that the predicted values for each output variable should satisfy certain conditions as given in Eq. (2).

$$(Y_{imen} + Y_{iwomen}) - X_{itotal} = 0, \forall \ \mathrm{i} = 1, 2, \ldots, \mathrm{n} \tag{2}$$

Where:

Y_{imen}: Number of men involved in accident i,

Y_{iwomen}: Number of women involved in accident i,

X_{itotal}: Total number of accidents

The optimization method employed to solve the constrained regression problem in this study is the Sequential Least Squares Programming (SLSQP) algorithm. SLSQP is a widely used optimization algorithm for problems with constraints (Fu et al., 2019), and it starts by initializing the weights with some initial values. The gradient and Hessian of the objective function concerning the weights are then calculated. These calculations depend on the specific objective function being minimized.

The objective function for the constrained multi-output linear regression model can be defined as the mean squared error (MSE) between the actual and predicted values of men and women involved in accidents. The formula for MSE can be written as Eq. (3) (Chai and Draxler, 2014).

$$MSE = \frac{1}{n}\sum_i \left(\left(Y_{i_{men\_predicted}} - Y_{i_{men\_actual}} \right)^2 + \left(Y_{i_{women\_predicted}} - Y_{i_{women\_actual}} \right)^2 \right) \tag{3}$$

Where:

n: the number of observations

$Y_{imen\_predicted}$: the predicted values of men involved in accidents for the ith observation,

Y_{imen_actual}: the predicted values of men involved in accidents for the ith observation,

$Y_{iwomen\_predicted}$: the predicted values of women involved in accidents for the ith observation,

Y_{iwomen_actual}: the predicted values of women involved in accidents for the ith observation,

The gradient is a vector of partial derivatives of the objective function for the weights. In the SLSQP algorithm, the gradient is estimated numerically by the optimization solver for each weight using Eq. (4) (Perez et al., 2012).

$$\frac{\partial MSE}{\partial \beta_i} = \frac{2}{n}\sum_i \left(\left(Y_{i_{men\_predicted}} - Y_{i_{men\_actual}} \right) \frac{\partial Y_{i_{men\_predicted}}}{\partial \beta_i} + \left(Y_{i_{women\_predicted}} - Y_{i_{women\_actual}} \right) \frac{\partial Y_{i_{women\_predicted}}}{\partial \beta_i} \right) \quad (4)$$

Where,

$\frac{\partial Y_{i_{men\_predicted}}}{\partial \beta_i}$ and $\frac{\partial Y_{i_{women\_predicted}}}{\partial \beta_i}$: the partial derivatives of the predicted values of men involved in accidents to the weight βi for the ith observation,

The Hessian is a matrix of second partial derivatives of the objective function for the weights. In the SLSQP algorithm, the Hessian is also estimated numerically by the optimization solver. The elements of the Hessian matrix can be approximated using Eq. (5) (Kazeev and Tyrtyshnikov, 2010).

$$\frac{\partial^2 MSE}{\partial \beta_i \partial \beta_j} = \frac{2}{n}\sum_i \left(\frac{\partial Y_{i_{men\_predicted}}}{\partial \beta_i} \frac{\partial Y_{i_{men\_predicted}}}{\partial \beta_j} + \frac{\partial Y_{i_{women\_predicted}}}{\partial \beta_i} \frac{\partial Y_{i_{women\_predicted}}}{\partial \beta_j} \right) \quad (5)$$

The system of equations is solved using the gradient and Hessian information by updating the weights iteratively. The weights are updated using the solution obtained from the system of equations. The optimal values that minimize the objective function while satisfying the constraint condition are calculated using Eq. (6) (Shen et al., 2019).

$$\theta^{(i+1)} = \theta^{(i)} - \left(H^{-1} \nabla MSE\left(\theta^{(i)}\right) \right) \quad (6)$$

Where:

$\theta^{(i)}$ is the weight vector at iteration i,

$H^{(-1)}$ is the inverse of the Hessian matrix,

∇ MSE($\theta^{(k)}$) is the gradient vector of the MSE objective function to the weights,

After updating the weights, the constraint equation is checked if it is satisfied, and the process is repeated until the constraint is satisfied. This calculation is repeated until the convergence criteria are met, such as reaching a maximum number of iterations, or the weights no longer significantly change.

After the linear regression model, tree-based regression models are also employed. These models help to address the statistical significance of the parameters rather than to estimate, present, discuss, or examine their signs or respective marginal contributions (Favero et al., 2023). The ensemble technique of tree-based regression models i.e. Random Forest, Extra Trees, and XGBoost combines multiple decision trees to improve prediction performance (Djarum et al., 2021). Thus, the ensemble nature of these models allows them to model interactions between variables, providing a more flexible representation of the underlying structure. They can handle both numerical and categorical features and are robust to outliers in the data (Cutler et al., 2009). Additionally, tree-based models provide feature importance measures, which can help identify the most influential features in the prediction process. However, predictive data is prone to errors. Data whose value is not close to zero can be evaluated using the mean absolute percentage error (MAPE) which is the most used forecast accuracy model given its interpretability and scale dependency (Kim and Kim, 2016). Thus, MAPE has been adopted to test the accuracy of the model. If At and Ft denote the actual and forecast values at the data point t respectively, then MAPE for n number of years is calculated using Eq. (7).

$$MAPE = \frac{1}{n}\sum_{t=1}^{N} \left| \frac{A_t - F_t}{A_t} \right| \quad (7)$$

2.3. Model Evaluation

A combination of metrics including RMSE, R2 score, and explained variance score is used for the evaluation of models (Zhou et al., 2021). By considering all three metrics, a comprehensive understanding of the model's performance in terms of accuracy, fit, and explained variability can be evaluated. Using a combination of these metrics is beneficial for evaluating the prediction of men and women involved in accidents in multi-output regression.

RMSE provides a measure of the average error between predicted and actual values, where a lower RMSE indicates better accuracy and a closer fit of the model to the actual data (Chicco et al., 2021), and is computed using Eq. (8).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (8)$$

R^2 Score evaluates the goodness of fit of the model, which is calculated using Eq. (9) as recommended by (Kvålseth, 1985). A value of 1 indicates that the model perfectly predicts the dependent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

Where:

y^i is the actual value,

$\bar{y}^i$ is the predicted value,

$\bar{y}$ is the mean of actual value,

The explained variance score is calculated by employing Eq. (10) (O'Grady, 1982) and assesses the proportion of variance explained by the model. A higher Explained Variance Score indicates that the model captures a larger portion of the variability in the data.

$$Explained\ Variance = 1 - \frac{Var(y - \hat{y})}{Var(y)} \quad (10)$$

Where:

$\mathrm{Var}(y - \hat{y})$ and Var(y) is the variance of prediction errors and actual values, respectively.,

Looking at the training and testing scores, we can observe that the models achieved higher scores on the training dataset compared to the testing dataset. This suggests that the models might be overfitting the training data, as they perform better on the seen data during training but show slightly lower performance on unseen test data. Among all the models used, the Random Forest model is performing better in all three model evaluation indicators. Hence the Feature Importance using random forest model is used for the further analysis of the contribution of different parameters in the accident involvement.

Table 2. Training and Test Performance of Machine Learning Models

Model		R2 Score	Explained Variance Score	RMSE
Linear Regression	Training	0.95	0.95	19.97
	Test	0.51	0.59	33.86
Random Forest Regressor	Training	0.99	0.99	28.52
	Test	0.63	0.77	62.93
XGBoost Regressor	Training	0.99	0.99	0.001
	Test	0.36	0.60	87.43
Extra Tree Regressor	Training	0.97	0.98	0.01
	Test	0.69	0.77	22.32

2.4. Feature Importance

The Feature Importance values in Random Forest represent the relative importance of each feature in making predictions. These values indicate the contribution of each feature towards the overall performance of the model. Higher values indicate a stronger influence of the feature on the predictions, while lower values indicate less importance (Palczewska et al., 2013).

For each decision tree, a node's importance, assuming only two child nodes, is calculated using Gini Importance which can be determined using Eq. (11) (Ronaghan, 2018).

$$GI_j = W_j \cdot C_j - W_{left(j)} \cdot C_{left(j)} - W_{right(j)} \cdot C_{right(j)} \quad (11)$$

Where:

GI_j is the Gini Importance of node j, representing the importance of the node,

W_j is the weighted number of samples reaching node j, indicating the sum of sample weights assigned to the node,

$W_{right(j)}$ is the weighted number of samples reaching the right child node of j,

$W_{left(j)}$ is the weighted number of samples reaching the left child node of j,

C_j is the impurity value of node j, typically calculated using the Gini index or another impurity measure,

$C_{right(j)}$ is the impurity value of the left node of j,

$C_{left(j)}$ is the impurity value of the left node j.

This equation quantifies the importance of a node by considering the reduction in impurity achieved by splitting on that node, considering the weighted impurity contributions from its child nodes. The importance of each feature on a decision tree is calculated applying Eq. (12) (Fu et al., 2019).

$$f_i = \frac{\sum_{(j:node\ j\ split\ on\ feature\ i)} n_j}{\sum_{(all\ k\ belonging\ to\ nodes)} n_k} \quad (12)$$

Where:

f_i is the importance of feature i,

n_j is the importance of node j,

nk is the importance of node k,

$\sum_{(j:node\ j\ split\ on\ feature\ i)} n_j$ is the sum of importance values for all nodes that split on feature i,

$\sum_{(all\ k\ belonging\ to\ nodes)} n_k$ is the sum of importance values for all nodes.

The feature importance values fi calculated for each feature can be further normalized to a value between 0 and 1 utilizing Eq. (13) (Ronaghan, 2018).

$$norm = \frac{f_i}{\sum_{j\ belonging\ to\ all\ features} f_j} \quad (13)$$

Where:

f_i is the importance of feature i,

$\sum_{j\ belonging\ to\ all\ features} f_j$ is the sum of importance values for all features.

The final feature importance, at the Random Forest level, is calculated as the average over all the trees (Ronaghan, 2018), following Eq. (14).

$$RFfi_i = \frac{\sum_{j=1}^{T} normfi_{ij}}{T} \quad (14)$$

Where:

$RFfi_i$ is the final feature importance for feature i in the Random Forest model,

$Normfi_{ij}$ is the normalized feature importance for feature i in tree j,

T is the total number of trees in the Random Forest

The sum of the feature importance scores on each tree is calculated and divided by the total number of trees. The feature importance score ranges from 0 to 1. The sum of all feature importance values is equal to 1.

3. Results and Discussion

First, the number of accidents of men and women is predicted using the model developed from the historical data and is compared with the real accident data from the year 2016 to 2021 that has been acquired from the BLS (2022). The graph in Fig. 3 shows the relationship between the predicted and actual number of accidents of men, women, and total for the different periods, 2016 to 2021, from which, it can be said that the model is performing well for this study.

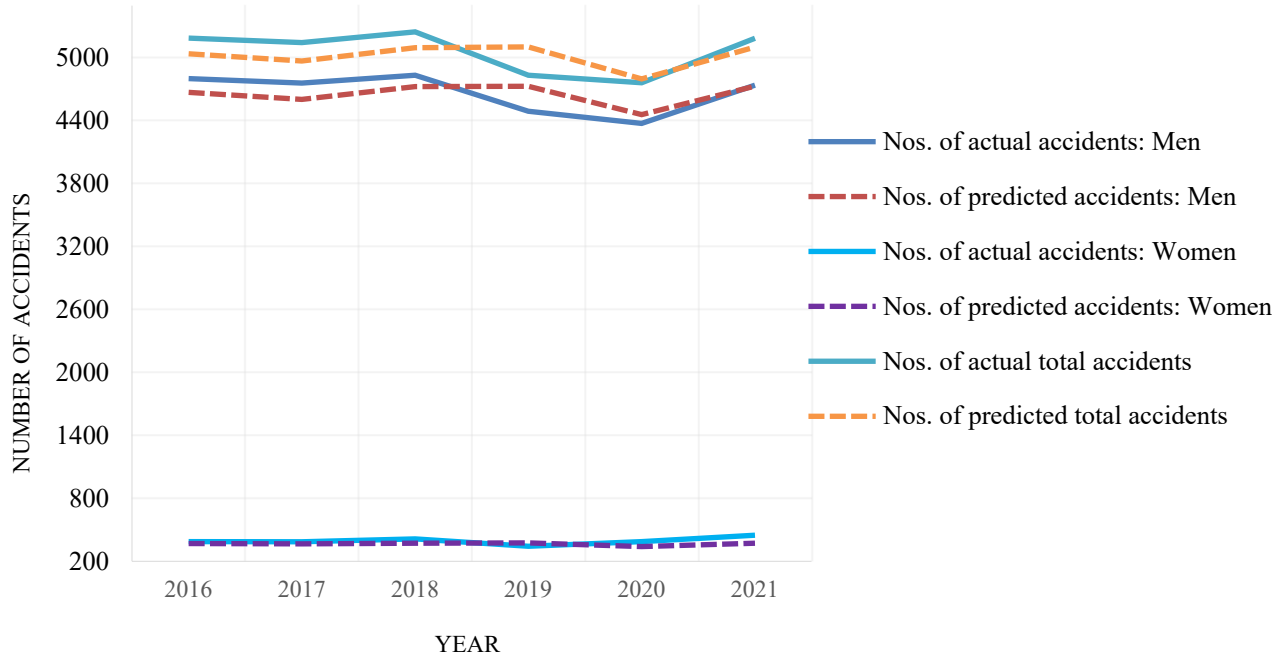


Fig. 3. Accident distribution actual vs predicted

Secondly, according to analysis results, feature importance scores are calculated for both the men and women involved in the fatal accidents. Based on these feature importance scores of men involved in accidents as can be seen in Fig. 4, it can be inferred that the number of wage and salary workers, the working-age workers, and the total number of accidents are the most influential factors for the involvement of male workers in accident. The features: the number of self-employed workers and older age worker's accident history have relatively lower influence. The feature: younger age workers fall in between with moderate importance.

Wage and Salary Workers: The feature importance value of 0.29 suggests that the variable "wage and salary workers" has the strongest influence on the number of men involved in accidents. This suggests the employment status of individuals classified as wage and salary workers is a significant factor in determining the likelihood of their involvement in accidents. It could indicate that men who work as wage and salary workers may be more prone to accidents compared to the self-employed category.

Working Age: With an importance value of 0.28, the variable "working age" also has an equal impact on the accident rates. This suggests the age range considered as working age plays an important role in determining the number of men involved in accidents. It implies that men within the working age range may have a higher likelihood of being involved in accidents compared to other age groups.

Total number of accidents: This variable has a significant importance score of 0.24. This implies that the overall frequency or occurrence of accidents, regardless of gender, has a moderate importance in the number of men involved in accidents. It indicates that a higher number of accidents overall might lead to a higher number of men being involved in accidents.

Youth: The feature importance value of 0.17 indicates that the variable "youth" has a moderate influence among the other variables considered. This implies that the age group classified as youth may have relatively less impact on the number of men involved in accidents compared to wage and salary workers and the working age group.

Older age and Self-employed workers: These features have the lowest importance in the rate of accidents among all others. Accidents involving men in this category may be less frequent or have different contributing factors compared to the other factors.

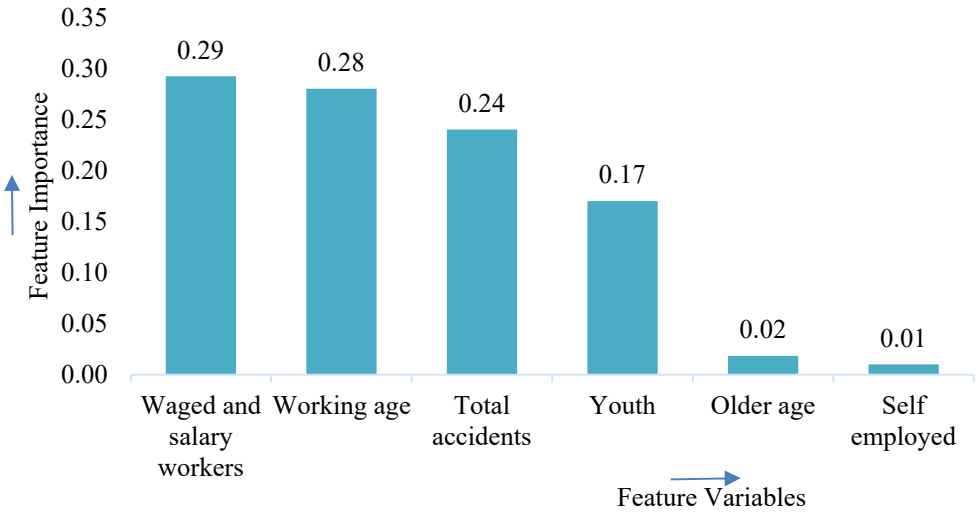


Fig. 4. Feature Importance Scores for Men Predictions

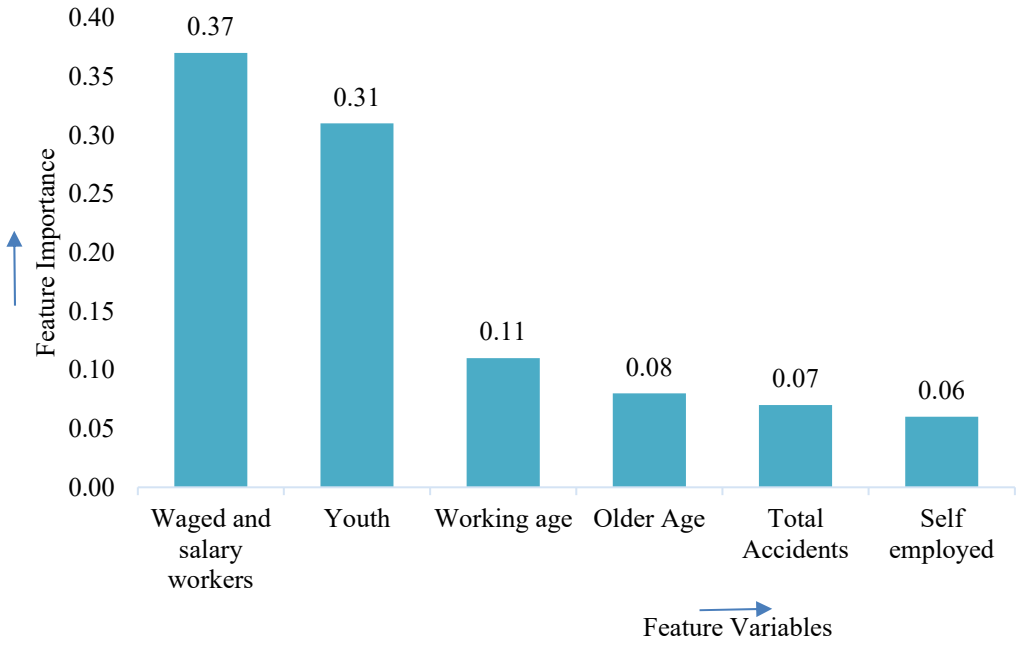


Fig. 5. Feature Importance Scores for Women Predictions

The graph in Fig. 5 shows the importance of each feature variable regarding the number of female workers involved in the accidents. Based on the feature importance scores, it can be implied that wage and salary workers have the highest importance value of 0.37, followed by younger age workers with an importance value of 0.31, and working age with an importance value of 0.11 in the number of women involved in accidents. The number of older age workers, the total number of accidents, and the presence or absence of self-employment also contribute, although to a lesser extent.

Wage and salary workers: The feature importance value of 0.37 indicates that the variable of this feature has a significant influence on the number of women involved in accidents. Based on the nature of work performed by a woman, one can determine the level of risk they are in.

Younger age workers: With an importance value of 0.31, this variable also has a noticeable impact. This suggests that the age group classified as youth plays a crucial role in determining the likelihood of women being involved in accidents. It implies that young women may have a higher probability of being involved in accidents compared to other age groups.

Working age workers: The feature importance value of 0.11 suggests that this factor has a relatively lower influence compared to the other two variables.

Other variables: The feature importance values of 0.08, 0.07, and 0.06 suggest that these features have a relatively lower influence compared to the other three variables.

However, the analysis should consider the specific context and limitations to draw accurate conclusions. The feature importance scores obtained from the random forest model are based on the specific data and methodology used. These scores may not capture the complete picture or may be influenced by various limitations. For instance, the Random Forest model assumes certain assumptions and may be sensitive to the choice of hyperparameters, sample size, or feature representation. Additionally, feature importance scores can be influenced by the presence of correlated features or the quality of the data. Therefore, it is crucial to acknowledge these limitations and understand that the feature importance scores provide insights within the specific constraints and assumptions of the model and data used.

The accuracy of the model based on the actual and predicted data between 2016 and 2021 is depicted in Table 3, with error calculation done using the MAPE approach. The global COVID-19 pandemic, which began in 2019, may have contributed to certain random errors observed thereafter.

Table 3. Forecast Error Evaluation

Year	Actual number of accidents			Predicted number of accidents			Error			Mean Absolute Percentage Error MAPE		
	Men	Women	Total	Men	Women	Total	Men	Women	Total	Men	Women	Total
2016	4803	387	5190	4673	368	5041	130	19	-149	2.43	4.91	2.87
2017	4761	386	5147	4605	366	4971	156	20	-176	3.28	5.18	3.42
2018	4837	413	5250	4727	372	5099	110	41	-151	2.00	9.93	2.88
2019	4492	344	4836	4731	375	5106	-239	-31	270	5.32	9.01	5.58
2020	4377	387	4764	4461	340	4801	-84	47	37	2.23	12.14	0.78
2021	4741	448	5189	4733	371	5104	8	77	-85	0.17	17.19	1.64
Average			5063			5020	13.5	28.83	-42.33	2.57	9.73	2.86

4. Conclusion

In this paper, constrained machine learning algorithms using the SLSQP optimization technique have been presented to predict the number of men and women involved in accidents. Also, the influence of various independent variables on such predictions has been quantified and discussed. Then various conclusions have been drawn looking at the feature importance of the Random Forest model. Based on the feature importance values, the employment status of wage and salary workers appears to be a strong influence regarding men involved in accidents, followed by the working age group and the total number of accidents. Similarly, for women, the same feature, that is the wage and salary of workers appear to be the most important factor responsible for the number of women involved in accidents. The younger age workers also play a significant role, followed by the number of working-age workers, while the older age workers, total number of accidents, and employment status of self-employed workers have a relatively lower influence. The creation of a model that can forecast gender-specific accident rates with a comparatively higher degree of precision can be beneficial for both the prediction as well as research in the future. Overall, this research contributes to the understanding of gender-specific patterns in accidents and highlights the importance of considering various factors in accident prediction models. The findings can support organizations and policymakers in implementing targeted interventions to reduce accidents and improve workplace safety for both men and women. By identifying the most important features contributing to accidents involving men and women, strategies can be implemented to address specific risk areas. This research contributes to the existing body of knowledge on occupational safety and provides practical insights for creating safer work environments.

It is important to note that the limited dataset posed challenges in achieving robust predictions. The research findings suggest that the model's performance could be further improved with a larger and more diverse dataset including more features. Therefore, future research should focus on gathering a larger and more representative dataset to enhance the analysis and prediction of accidents. This would enable a better understanding of the underlying patterns and provide more reliable insights into the factors influencing accidents involving men and women. Moreover, future research may focus on developing different machine learning models, the consideration of alternative optimization algorithms, and the extension of the methodology. By addressing these issues, more robust and reliable accident prevention strategies can be developed to ensure workplace safety.

Author Contributions

Deepanshu Mahajan and Emel Sadikoglu contributed to conceptualization, methodology, analysis, investigation, data collection, and draft preparation. Uttam Kumar Pal and Saksham Timalsina performed the data analysis and manuscript editing. Chengyi Zhang and Sevilay Demirkesen contributed to the supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

Bena, A., Giraudo, M., Leombruni, R., and Costa, G. (2013). Job tenure and work injuries: A multivariate analysis of the relation with previous experience and differences by age. *BMC Public Health*, 13(1). https://doi.org/10.1186/1471-2458-13-869

Chai, T., and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Geosci. *Model Dev. Discuss*, 7, 1525–1534. https://doi.org/10.5194/gmdd-7-1525-2014

Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623. https://doi.org/10.7717/peerj-cs.623

Cutler, A., Cutler, D. R., and Stevens, J. R. (2009). Tree-Based Methods. In *High-Dimensional Data Analysis in Cancer Research* (pp. 1–19). Springer New York. https://doi.org/10.1007/978-0-387-69765-9_5

De Raedt, L., Guns, T., and Nijssen, S. (2010). Constraint Programming for Data Mining and Machine Learning. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*. www.aaai.org

Djarum, D. H., Ahmad, Z., and Zhang, J. (2021). River Water Quality Prediction in Malaysia Based on Extra Tree Regression Model Coupled with Linear Discriminant Analysis (LDA). In *Computer Aided Chemical Engineering* (Vol. 50, pp. 1491-1496). Elsevier.

Dobrucali, E., Sadikoglu, E., Demirkesen, S., Zhang, C., Tezel, A., and Kiral, I. A. (2023). A bibliometric analysis of digital technologies use in construction health and safety. *Engineering, Construction and Architectural Management*, ahead-of-print(ahead-of-print). https://doi.org/10.1108/ECAM-08-2022-0798

Favero, L. P., Belfiore, P., and de Freitas Souza, R. (2023). *Data science, analytics and machine learning with R.* Academic Press.

Forssberg, K. S., Vänje, A., and Parding, K. (2022). Bringing in gender perspectives on systematic occupational safety and health management. *Safety Science*, 152. https://doi.org/10.1016/j.ssci.2022.105776

Fu, Zhengqing and Liu, Goulin and Guo, Lanlan. (2019). Sequential Quadratic Programming Method for Nonlinear Least Squares Estimation and Its Application. *Mathematical Problems in Engineering*. 2019. 1-8. 10.1155/2019/3087949.

Gonzalez-Delgado, M., Gómez-Dantés, H., Fernández-Niño, A., Robles, E., Borja, V. H., and Aguilar, M. (2015). *Factors Associated with Fatal Occupational Accidents among Mexican Workers: A National Analysis.* https://doi.org/10.1371/journal.pone.0121490

Jones, M. K., Latreille, P. L., Sloane, P. J., and Staneva, A. V. (2013). Work-related health risks in Europe: Are older workers more vulnerable? *Social Science and Medicine, 88,* 18–29. https://doi.org/10.1016/j.socscimed.2013.03.027

Kazeev, V. A., and Tyrtyshnikov, E. E. (2010). Structure of the Hessian Matrix and an Economical Implementation of Newton's Method in the Problem of Canonical Approximation of Tensors. Computational Mathematics and Mathematical Physics, 50(6), 979–998. https://doi.org/10.1134/S0965542510060011

Khanzode, V. V., Maiti, J., and Ray, P. K. (2012). Occupational injury and accident research: A comprehensive review. In *Safety Science* (Vol. 50, Issue 5, pp. 1355–1367). https://doi.org/10.1016/j.ssci.2011.12.015

Kim, S., and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting,* 32(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Kriebel, D. (1982). Occupational Injuries: Factors Associated with Frequency and Severity. *International Archives of Int Arch Occup Environ Health*, 50, 209–218.

Kvålseth, T. O. (1985). Cautionary Note about R^2. *American Statistician*, 39(4), 279–285. https://doi.org/10.1080/00031305.1985.10479448

Kvålseth, T. O. (1985). Cautionary Note aboutR2. *The American Statistician*, 39(4), 279–285. doi:10.1080/00031305.1985.10479448

Lee, J. Y., Yoon, Y. G., Oh, T. K., Park, S., and Ryu, S. Il. (2020). A Study on Data Pre-Processing and Accident Prediction Modelling for Occupational Accident Analysis in the Construction Industry. *Applied Science*, 10(21). https://doi.org/10.3390/app10217949

Maheronnaghsh, S., Zolfagharnasab, H., Gorgich, M., and Duarte, J. (2023). Machine learning in Occupational Safety and Health-a systematic review. *International Journal of Occupational and Environmental Safety*, 7(1), 14-32.

Natarajan, B. K. (2014). *Machine learning: A theoretical approach. Elsevier*

O'Grady, K. E. (1982). Measures of Explained Variance: Cautions and Limitations. *Psychological Bulletin*, 92(3), 766–777.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, 92(3), 766–777. doi:10.1037/0033-2909.92.3.766

Palczewska, A., Palczewski, J., Robinson, R. M., and Neagu, D. (2013). *Interpreting random forest classification models using a feature contribution method.*

Park, S. J., Jung, M., and Sung, J. H. (2019). Influence of Physical and Musculoskeletal Factors on Occupational Injuries and Accidents in Korean Workers Based on Gender and Company Size. *International Journal of Environmental Research and Public Health*, 16(345). https://doi.org/10.3390/ijerph16030345

Perez, Guillaume and Ament, Sebastian and Gomes, Carla and Lallouet, Arnaud. (2021). *Constrained Machine Learning: The Bagel Framework.*

Perez, R. E., Jansen, P. W., and Martins, J. R. R. A. (2011). pyOpt: a Python-based object-oriented framework for nonlinear constrained optimization. *Structural and Multidisciplinary Optimization*, 45(1), 101–118. doi:10.1007/s00158-011-0666-3

Perez, R. E., Jansen, P. W., and Martins, J. R. R. A. (2012). PyOpt: A Python-based object-oriented framework for nonlinear constrained optimization. *Structural and Multidisciplinary Optimization*, 45(1), 101–118. https://doi.org/10.1007/s00158-011-0666-3

Ronaghan, S. (2018, May 11). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. *Medium; Towards Data Science*. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Shen, Z., Ribeiro, A., Hassani, H., Qian, H. and Mi, C. (2019). Hessian Aided Policy Gradient. *Proceedings of the 36th International Conference on Machine Learning* 97:5729-5738 Available from https://proceedings.mlr.press/v97/shen19d.html.

U.S. Bureau of Labor Statistics. (2022). *IIF Home*. U.S. Bureau of Labor Statistics. https://www.bls.gov/iif/

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C., and Shen, X. (2023). A Survey on Multi-output Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.

Xu, Donna and Shi, Yaxin and Tsang, Ivor and Ong, Yew and Gong, Chen and Shen, Xiaobo. (2019). *Survey on Multi-Output Learning. IEEE Transactions on Neural Networks and Learning Systems*. PP. 1-21. 10.1109/TNNLS.2019.2945133.

Xu, Y., Zhou, Y., Sekula, P., and Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, 6. https://doi.org/10.1016/j.dibe.2021.100045

Yosef, T., Sineshaw, E., and Shifera, N. (2023). Occupational injuries and contributing factors among industry park construction workers in Northwest Ethiopia. *Front Public Health.* https://doi.org/10.3389/fpubh.2022.1060755

Zhou, J., Gandomi, A. H., Chen, F., Holzinger, A., and Balas, V. E. (2021). *Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics*. https://doi.org/10.3390/electronics10050593

Deepanshu Mahajan is currently an undergraduate student at the Indian Institute of Technology, Madras at the Department of Data Science and Applications. He worked as a research intern at Gebze Technical University from May 2023 to Aug 2023. He currently conducts research on Machine Learning and Deep Learning Applications.

Emel Sadikoglu received a B.Sc. degree in civil engineering from Bogazici University, Istanbul, Turkey, in 2018, and an M.Sc. degree in construction management from Bogazici University, Istanbul, Turkey, in 2021. She is currently a Research and Teaching Assistant with the Department of Civil Engineering, Construction Management Division, Gebze Technical University, Kocaeli, Turkey. Her research interests include lean construction, agility, green construction, occupational health and safety, and digital technologies.

Uttam Kumar Pal is pursuing his Ph.D. in Civil Engineering focusing on Construction Management at the University of Wyoming, USA, and working with a dynamic research group led by Prof. Dr. Charlie Zhang in the Construction Innovation and Research Lab. He has a bachelor's degree in civil engineering from Tribhuvan University, IOE Pulchowk Campus, Nepal. And, has experience working with the government as well as private sector contractors in the field of highway/pavement engineering. He is also interested and researching how the world can produce, supply, and utilize clean, net-zero carbon energy, especially through Nuclear Energy generation, focusing on the efficacy, economics, safety, resilience, environmental impacts, and more, particularly through the development and deployment of the advanced nuclear reactors.

Saksham Timalsina is a graduate research assistant at the Department of Civil and Architectural Engineering and Construction Management at the University of Wyoming, USA. He completed his undergraduate degree in Civil Engineering from Tribhuvan University, Nepal, and is currently pursuing a doctorate with the same major at the University of Wyoming. His research interests include Construction Safety, Automated Construction, Sustainable Construction, and Energy Infrastructure Safety.

Chengyi Zhang is currently an associate professor in the Department of Civil and Architectural Engineering and Construction Management at the University of Wyoming, USA. His research focuses on construction management, construction safety, construction workforce development, and emerging technology in construction.

Sevilay Demirkesen currently works as an associate professor at the Department of Civil and Environmental Engineering of Gebze Technical University. Sevilay is a former member of P2SL of the University of California, Berkeley. Her research focuses on construction management, construction safety, and lean construction.