RESEARCH

A Data-driven Comprehensive Analysis of Occupational Health and Safety in Turkey: Application of Statistical and Machine Learning Approaches

Meltem Aksoy · Aylin Adem · Ibrahim Yilmaz · Metin Dağdeviren

Received: 28 November 2025 / Revised: 2 March 2026 / Accepted: 22 April 2026

Abstract

Occupational Health and Safety (OHS), which holds substantial importance due to its profound human implications and far-reaching economic consequences, requires effective management to minimize workplace accidents and occupational diseases through data-driven risk assessment and proactive safety strategies. Achieving this necessitates the availability of historical data and its examination through appropriate analytical methodologies, which is critically important for determining evidence-based preventive measures. Thus, this study analyses Turkey's national OHS records from 2010 to 2022 using statistical analysis and machine learning approaches to evaluate sectoral risks, predict accident trends, and identify key determinants of workplace hazards. Results indicate substantial sector differences, with coal mining and heavy industry showing the highest accident and fatality rates. Machine learning models demonstrate strong predictive capability, with gradient boosting providing the best work accident prediction performance and random forests achieving the best performance for occupational disease prediction. Clustering analysis identifies three distinct industrial risk groups, while Principal Component Analysis (PCA) reveals regional disparities, particularly in highly industrialized provinces such as Istanbul, Kocaeli, and Izmir. Classification models further achieve over 98% accuracy in identifying high-risk groups, highlighting the potential of machine learning for proactive OHS management. The findings of this paper provide actionable insights for policymakers and industry leaders to optimize safety regulations and develop targeted interventions and strategies to reduce workplace risks and identify which sectors and worker groups should be prioritized.

Keywords Occupational Health and Safety · Statistical Analysis · Machine Learning · OHS Management · Predictive Modeling · Workplace Accidents

1 Introduction

Occupational Health and Safety (OHS) can be broadly defined as a management approach encompassing the measures taken to ensure employees' safety, allowing them to remain healthy and productive in the workplace [1, 2]. More specifically, all activities undertaken at both the employee and organizational levels to prevent occupational accidents or diseases can be seen, in a sense, as the definition of OHS. Thousands of workers lose their

Published online: 02 May 2026

lives or become incapacitated due to workplace accidents and occupational diseases that occur every year worldwide [3]. In addition to the human and psychological impacts, such as the loss of human resources and reduced productivity and motivation in the workplace, this situation also leads to significant economic losses, including healthcare costs and compensation payments [4–7]. One of the most fundamental indicators for measuring the performance of OHS activities, which are the primary means of preventing serious occupational accidents and diseases, is the number of workplace accidents resulting in death or injury, as well as the number of individuals at risk of contracting an occupational disease [8].

From this perspective, according to the published statistics for 2023, the number of work accidents in Turkey has been reported as 681,401 [9]. This number represents the overall total of reported work accidents (fatal + non-fatal) aggregated at the national level. When interpreting the changes in this ratio over the years, the impact of industrialization, the influence of technology on industrialization, and the resulting reduction in the use of blue-collar workers as active labor should be considered. Even when considering the reducing effect of these factors on workplace accidents, it is still quite high.

When a work accident occurs or an occupational disease is diagnosed in workplaces operating in Turkey, whether in the public or private sector, employers are required to report all relevant information to the official institutions and create an official record [9]. Based on these mandatory notifications, official institutions have compiled a comprehensive database containing detailed information on OHS activities over the years and have made this data available to the public.

Building on this rich and systematically collected dataset, the main objective of this study is to analyse long-term OHS data from multiple perspectives. The goal is to present the results holistically, assess how OHS is perceived, implemented, and prioritized in Turkey based on sectoral, regional, and economic patterns observed in workplace accident and occupational disease data, and map Turkey's position on this issue. To achieve these objectives and support future-oriented insights, classical statistical techniques are combined with Machine Learning (ML) methods, enabling both descriptive analyses of historical trends and data-driven projections.

Although the use of ML for OHS analysis is not new [10–12], the novelty of this study lies not only in its nationwide scope, longitudinal structure, and multidimensional analytical design, but also in its use of rigorously collected real-world data spanning multiple years in Turkey. The findings derived from this comprehensive dataset provide policy-relevant insights with strong potential to guide policymakers and support evidence-based improvements in the national OHS system, while also contributing to the safety science literature through a large-scale, context-specific empirical analysis.

This research provides a comprehensive examination of Turkey's OHS landscape over a 13-year period (2010–2022) by integrating statistical techniques and multiple ML algorithms. The analysis incorporates several innovative elements, including examining the frequency and severity of workplace accidents and occupational diseases across economic sectors; identifying key factors contributing to incident occurrence; grouping sectors into distinct risk categories based on accident characteristics; reducing the dimensionality of provincial data to uncover regional safety patterns; classifying worker groups into high- and low-risk categories based on incident severity; analysing demographic differences across age and gender; and investigating temporal trends, including monthly and seasonal variations in both workplace accidents and occupational diseases.

2 Literature Review

The literature review section was organized by classifying the techniques used to analyze OHS data. The studies cited in this section were selected to illustrate the main methodological approaches used in the analysis of OHS data. Rather than presenting a general literature review, the literature was structured by classifying studies according to the analytical techniques employed, namely classical statistical methods and ML–based approaches. The selected studies cover different countries, sectors, time periods, and data characteristics, thereby reflecting the diversity of applications and methodological trends in the OHS literature. In addition, particular emphasis was placed on studies that involve data structures similar to those considered in this research and that employ analytical techniques closely aligned with the methods used in this paper.

Several studies have employed classical statistical techniques to explore patterns, trends, and risk factors in OHS data. A summary of studies using statistical analysis is presented in Table 1. Zhou et al. [13] conducted a statistical analysis of hazardous chemical accidents in China from 2008 to 2021. Their analysis focused on specific business categories, industry distribution, business scales, provincial location distribution, time variability characteristics, and causes of accidents. Soykan [14] conducted statistical analyses and revealed trends in occupational accidents in the Turkish fisheries and aquaculture sectors between 2006 and 2020. Costa da Silva et al. [15] analyzed data with a statistical approach for the Brazilian mining sector, using occupational accident and disease notification data from 2017 to 2021. Gil and Pelon [16] analyzed data on occupational accidents in the Polish mining industry from 2007 to 2018 using statistical techniques and estimated the number of occupational accidents in the short term. Bilim and Bilim [17] analyzed accidents in Turkish coal mines from 2014 to 2019 using Logistic Regression (LogReg). Based on this analysis, they developed an equation to estimate workday loss based on the relationship between workers' demographic characteristics and the accidents that caused this loss. Similarly, Eravcı and Yılmaz [18] conducted a statistical analysis on the economic impact of permanently disabled workers resulting from occupational accidents in Turkey.

Małysa and Gajdzik [19] analyzed work accidents in the Polish steel industry, and based on these analyses, worked on modeling the number of work accident victims. They used various prediction models, such as the Winter and Holt models. They reported a decreasing trend in the number of people injured in work accidents in the steel industry. Tanaya et al. [20] conducted statistical tests to determine whether a relationship exists between occupational accidents in the South Borneo coal industry and the age, length of service, and education level of those involved in the accidents. Their findings indicated no significant relationship between these variables. Jafari et al. [21] collected data from a company to identify risk factors associated with workplace accidents and aimed to determine the groups requiring precautions through statistical analyses. Data were gathered on variables such as body mass index, education level, work experience, smoking habits, use of sedative drugs, and sleep disorders from employees in an Iranian company. They inferred that employees with a high body mass index, low education level, and limited work experience were more prone to workplace accidents.

Table 1 Studies using statistical techniques for OHS data analysis

Study	Country	Sector	Data Period	Methods	Purpose
Machfudiyanto et al. [11]	Indonesia	Construction	-	PLS-SEM	Analysis of factors affecting safety design
Zhou et al. [13]	China	Chemical industry	2008–2021	Categorization of OHS characteristics using histograms	Analysis of accident characteristics and causes
Soykan [14]	Turkey	Fisheries and aquaculture	2006–2020	Kruskal–Wallis and Mann–Whitney U tests	Trend analysis and subsectors comparison of occupational accidents
Costa da Silva et al. [15]	Brazil	Mining	2017–2021	Retrospective observational time-series	Analysis of occupational accident and disease notifications
Gil and Pelon [16]	Poland	Mining	2007–2018	Econometric models	Short-term estimation of occupational accidents
Bilim and Bilim [17]	Turkey	Coal mining	2014–2019	LogReg	Estimation of workday loss
Eravcı and Yılmaz [18]	Turkey	Multiple sectors	2013–2022	Descriptive statistics	Economic impact analysis of permanently disabled workers
Małysa and Gajdzik [19]	Poland	Steel industry	2009–2018	Holt and Winter models	Modeling the number of work accident victims
Tanaya et al. [20]	Indonesia	Coal industry	2017–2019	Descriptive statistics	Analysis of relationships between accidents and demographic variables
Jafari et al. [21]	Iran	Manufacturing	2016	LogReg and chi-square	Identification of workplace accident risk factors

In recent years, ML techniques have been increasingly adopted in OHS research to enhance predictive accuracy, automate risk assessment, and support data-driven decision-making. These studies mainly focus on accident cause prediction, severity classification, risk forecasting, and pattern recognition using supervised and hybrid learning models. An overview of representative ML-based studies is presented in Table 2. Özkan and Ulas [10] predicted the causes of accidents in the Turkish metal industry using data from 2013 to 2018, employing Random Forest (RF), k-Nearest Neighbor (KNN), Gradient Boosting (GB), and Recursive Partitioning and Regression Trees (RPART) methods, and emphasized that the RF algorithm performed the best in this framework. da Silva et al. [22] examined neural and non-neural methods for classifying the severity of occupational accidents in the shoe manufacturing industry and stated that RF and Extreme Gradient Boosting (XGBoost) are the two most effective non-neural methods. They concluded that these methods could classify new accidents and simulate scenarios. Santurian et al. [23] explored the relationship between occupational accidents and heat waves in Madrid, Barcelona, and Valencia. Heat waves were defined using daily occupational accident data collected between 2012 and 2021, and the Generalized Additive Model (GAM) and Distributed Lag Nonlinear Model (DLNM) were applied to determine the relationship between occupational accidents and heat waves. Machfudiyanto et al. [11] analyzed the factors affecting the implementation of safety design using Partial Least Squares - Structural Equation Modeling (PLS-SEM) and Artificial Neural Network (ANN). They applied the proposed approach in the Indonesian construction industry. Mammadov et al. [24] proposed an artificial intelligence-based model to predict the consequences of occupational accidents by analyzing a dataset containing injuries, near misses, and asset-product damage from a pipeline construction project using ML algorithms. Koç et al. [25] focused on identifying the vulnerable body parts of construction workers. For the Turkish construction sector, they used a multi-stage ensemble model with ML algorithms to determine which factors influence the vulnerability of specific body parts and suggested precautions accordingly. Koç et al. [12] analyzed occupational accidents recorded in the Turkish construction sector to predict construction workers who may face a risk of fatality. In this analysis, they used a combination of the RF algorithm and particle swarm optimization for prediction. Koç and Gurgun [26] developed a scenario-based automatic preprocessing model that determines the best scenario for predicting the severity of

Table 2 Studies using ML–based approaches for OHS data analysis

Study	Country	Sector	Data Period	Methods	Purpose
Özkan and Ulas [10]	Turkey	Metal industry	2013–2018	RF, KNN, GB, RPART	Prediction of accident causes
da Silva et al. [22]	Brazil	Shoe manufacturing	2016–2022	RF, XGBoost, ANN	Classification of accident severity
Santurian et al. [23]	Spain	Multiple sectors	2012–2021	GAM, DLNM	Analysis of relationship between heat waves and occupational accidents
Machfudiyanto et al. [11]	Indonesia	Construction	-	ANN	Analysis of factors affecting safety design
Mammadov et al. [24]	-	Pipeline construction	-	ANN, Deep Learning, Functional Linear Model, Decision Tree (DT), GB, RF, Support Vector Machines (SVM)	Prediction of accident consequences
Koç et al. [25]	Turkey	Construction	2012–2020	RF, KNN, Naïve Bayes (NB), ANN	Identification of vulnerable body parts
Koç et al. [12]	Turkey	Construction	2007–2020	RF, Particle Swarm Optimization	Prediction of fatality risk
Koç and Gurgun [26]	Turkey	Construction	2012–2020	XGBoost	Prediction of accident severity
Koç et al. [27]	Turkey	Construction	2012–2020	Wavelet transform, ANN, SVM, Multivariate Adaptive Regression Splines	Accident number estimation
Goldberg [28]	USA	Metal and mining	2015–2020	Text mining, DT, RF, NB, SVM, ANN	Classification of occupational accident records
Koklonis et al. [29]	Greece	Health services	2014–2019	NB, Bayesian Network, KNN, Multilayer Perceptron	Prediction and classification of workplace accidents
Sadeghi et al. [30]	Malaysia	Construction	-	ANN and fuzzy inference systems	OHS risk assessment

construction accidents. XGBoost analysis shows that year, cause of material, age, past accidents, experience, and salary are the most effective variables. Koç et al. [27] estimated the number of accidents for short, medium, and long-term periods based on occupational accidents recorded in Turkey between 2012 and 2020, using time series data. They combined the discrete wavelet transform and ML methods to estimate the number of occupational accidents and applied these methods to data in the field of construction safety. Goldberg [28] applied text mining to analyze occupational accident data in textual form. He used ML models to classify accident records from different perspectives, such as the injured body part, source of injury, type of event causing the injury, whether hospitalization occurred, and whether limb loss occurred. He applied his proposed algorithms to accident data from the metal and mining sectors and stated that the performance was satisfactory. Koklonis et al. [29] aimed to develop a decision support system to predict and classify workplace accidents using ML algorithms. The data used consisted of 476 incident reports from Metaxa Cancer Hospital in Greece, collected between 2014 and 2019. Sadeghi et al. [30] developed a model based on the integration of neural networks with fuzzy inference systems to assess OHS risks for workers in the Malaysian construction industry and applied it to several Malaysian construction projects to demonstrate its effectiveness.

A review of recent studies shows that although statistical and ML methods are increasingly used in the OHS domain, existing research remains fragmented in terms of scope, methodology, and analytical depth. First, a large portion of the literature is sector-specific rather than national in scope, focusing primarily on construction [12, 25–27], metal manufacturing [10, 28], mining [17], or health services [29]. While these studies provide valuable insights at the micro level, they do not allow cross-sectoral benchmarking or broader policy implications. Second, many research efforts employ single-method analytical designs, either using only descriptive statistics or regression-based models [14, 17, 18], or applying isolated supervised ML methods such as RF, GB, ANN, or XGBoost solely for prediction or classification [10, 11, 22, 24]. Third, most existing ML studies examine only one analytical dimension of OHS data, such as accident causes [10, 25], severity prediction [12, 22], or short-term forecasting [27], without extending the analysis to occupational diseases, demographic trends, or regional and sectoral risk patterns.

This study makes several original and significant contributions to the OHS literature:

- Methodological Contribution: This study proposes a comprehensive multi-method analytical framework that integrates descriptive statistics, ML–based prediction and classification, time-series forecasting, dimensionality reduction, and clustering analyses. To the best of our knowledge, such an integrated and holistic analytical approach has not previously been applied to OHS data.
- Data Scope and Coverage: Unlike prior studies that rely on limited samples or focus on single industries, this research utilizes nationwide longitudinal OHS records covering the period 2010–2022 and encompassing all economic activities classified under NACE Rev. 2.1. This extensive coverage enhances the robustness, generalizability, and representativeness of the findings.
- Cross-Sectoral and Regional Insights: The study enables systematic cross-sector comparisons and regional profiling of OHS risks, providing insights that are largely absent in the existing literature. This allows for the identification of sector-specific and region-specific risk patterns at a national scale.
- Demographic Risk Assessment: By incorporating demographic dimensions into the analysis, the study offers a detailed assessment of OHS risks across different worker groups, contributing to a more nuanced understanding of occupational safety dynamics.
- Practical and Policy Relevance: The findings derived from real-world data and advanced analytical techniques provide evidence-based insights that can support policymakers, regulators, and practitioners in designing targeted and data-driven OHS interventions.

3 Methodology

This section describes the dataset, the variables used in the study, and the methods applied for risk assessment, including descriptive analysis, predictive modeling, forecasting techniques, clustering, and dimensionality reduction techniques.

3.1 Dataset Description

The raw data on work accidents and occupational diseases in Turkey was obtained from the Social Security Institution and covers the years 2010 to 2022 [9]. The analysis period was limited to 2010–2022 for methodological and data consistency reasons. First, systematic and standardized nationwide reporting of occupational accidents and occupational diseases in Turkey became fully operational after 2010, following regulatory and institutional improvements in the Social Security Institution's data collection framework. Earlier records are either incomplete or not fully comparable in terms of variable definitions and reporting practices. Second, 2022 represents the most recent year for which finalized and publicly available national OHS statistics were accessible at the time of analysis. Restricting the study to this period therefore ensures longitudinal consistency, data reliability, and comparability across years.

The primary goal of this study is to predict sector-level accident risks and assess the impact of economic activities on workplace safety using statistical and ML approaches. To achieve this, we extracted specific variables from the dataset, including workplace incident attributes such as the number of work accidents, the number of occupational diseases, total incapacity days, permanent incapacity cases, fatal accidents, fatal occupational diseases, work accident duration, working period of insurance, and type of injury. Additionally, demographic attributes like age and gender, along with sector-based classifications following the NACE Rev. 2.1 economic activity codes, were included. These variables and their explanations are provided in Table 3. Since the data originate from mandatory formal reporting systems, the dataset contains no missing values, and therefore no missing data completion procedures were required. Duplicate records were removed before analysis to ensure data integrity.

3.2 Data Analysis Framework

Our research employed a comprehensive, multi-method analytical framework to assess OHS risks. The analysis proceeded in multiple stages as shown in Fig. 1.

Descriptive analysis. Descriptive statistical analysis was first conducted to examine patterns, trends, and distributions in workplace incidents. This assessment provided an initial understanding of how work accidents and occupational diseases vary across sectors, regions, and demographic groups.

ML-*based Prediction.* For predictive modeling, we focused on forecasting two key occupational safety indicators: the number of work accidents and the number of occupational diseases. We constructed a robust dataset with carefully selected predictor variables, including occupational incidents, incapacity days, worker demographics, economic activities, and fatality records. These variables were specifically chosen to capture meaningful correlations and patterns in workplace safety trends. We implemented multiple ML algorithms to develop our predictive models. Decision Tree (DT) [31] provided interpretable hierarchical structures that revealed influential variables affecting accident outcomes. RF [32] enhanced prediction accuracy by aggregating multiple DT while delivering critical feature importance metrics. GB [33] further improved our predictive capability by sequentially building models that systematically corrected previous errors, effectively capturing complex relationships within safety data. Linear Regression (LR) [34] served as our baseline statistical approach, establishing a comparative benchmark against more sophisticated techniques. As the ML models were formulated as regression problems with continuous target variables (i.e., the number of work accidents and occupational diseases), model performance was evaluated using standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2), which are more appropriate for assessing predictive accuracy

Table 3 Description of variables

Variable	Description
Year	Year in which the work accident or occupational disease occurred (2010–2022).
Month	The month in which the work accident or occupational disease occurred.
Number of Work Accidents	Total number of recorded work accidents in a given sector or region.
Number of Occupational Diseases	Total number of recorded occupational diseases in a given sector or region.
Total Incapacity Days (Work Accidents)	The sum of incapacity days resulting from work accidents.
Total Incapacity Days (Occupational Diseases)	The sum of incapacity days due to occupational diseases.
Number of Permanent Incapacities (Work Accidents)	The number of cases where work accidents resulted in permanent disabilities.
Number of Permanent Incapacities (Occupational Diseases)	The number of cases where occupational diseases resulted in permanent disabilities.
Number of Fatal Work Accidents	Total number of work-related fatalities.
Number of Fatal Occupational Diseases	Total number of fatalities resulting from occupational diseases.
Work Accident Duration	The number of lost workdays due to temporary incapacity.
Working Period of Insured	The total period the insured worker has been working before the accident.
Type of Injury	Type of injury sustained (fracture, sprain, burn, etc.) based on ESAW classification.
Economic Activity (NACE Rev.2.1)	Sector-based classification of workplaces using NACE Rev.2.1 codes.
Gender	Gender distribution of workers affected by work accidents and occupational diseases.
Age	Age of workers affected by work accidents and occupational diseases.
Province	The geographical province where the incident occurred.

in continuous outcome settings. For the ML-based prediction, the dataset was randomly split into training and testing subsets using a 70/30 ratio, where 70% of the observations were used for model training and 30% were held out for testing.

Because occupational disease cases constitute a very small proportion of the dataset, the prediction target is highly sparse and rare. To examine whether this imbalance could be mitigated through resampling, both SMOTE-based oversampling and random undersampling of the majority class were applied during the model training process. SMOTE was used to synthetically increase the representation of minority cases, while random undersampling was applied to reduce the dominance of the majority class. However, neither approach resulted in a substantial improvement in predictive performance. In particular, random undersampling led to information loss by discarding a large number of majority-class observations, which likely weakened the models' ability to learn stable patterns. Overall, these findings indicate that the limited predictive performance primarily stems from a weak predictive signal in the available features rather than class imbalance alone.

Workplace Risk Classification. In this study, we aimed to classify aggregated workplace incident data into high-risk and low-risk categories using a two-stage analytical approach. Each observation in this classification task represents a unique combination of year, age group, and gender, with incident counts (e.g., work accidents, fatal accidents, occupational diseases) aggregated across all workers within that demographic stratum. In the first stage, a weighted risk scoring framework was applied to derive risk labels for each stratum, followed by a ML–based classification stage to automatically distinguish between high-risk and low-risk groups.

The risk scoring process was based on three primary metrics: the number of fatal work accidents, the number of work accidents, and the number of occupational diseases. Each metric threshold was determined as follows:

 https://doi.org/10.1007/s44196-026-01351-7

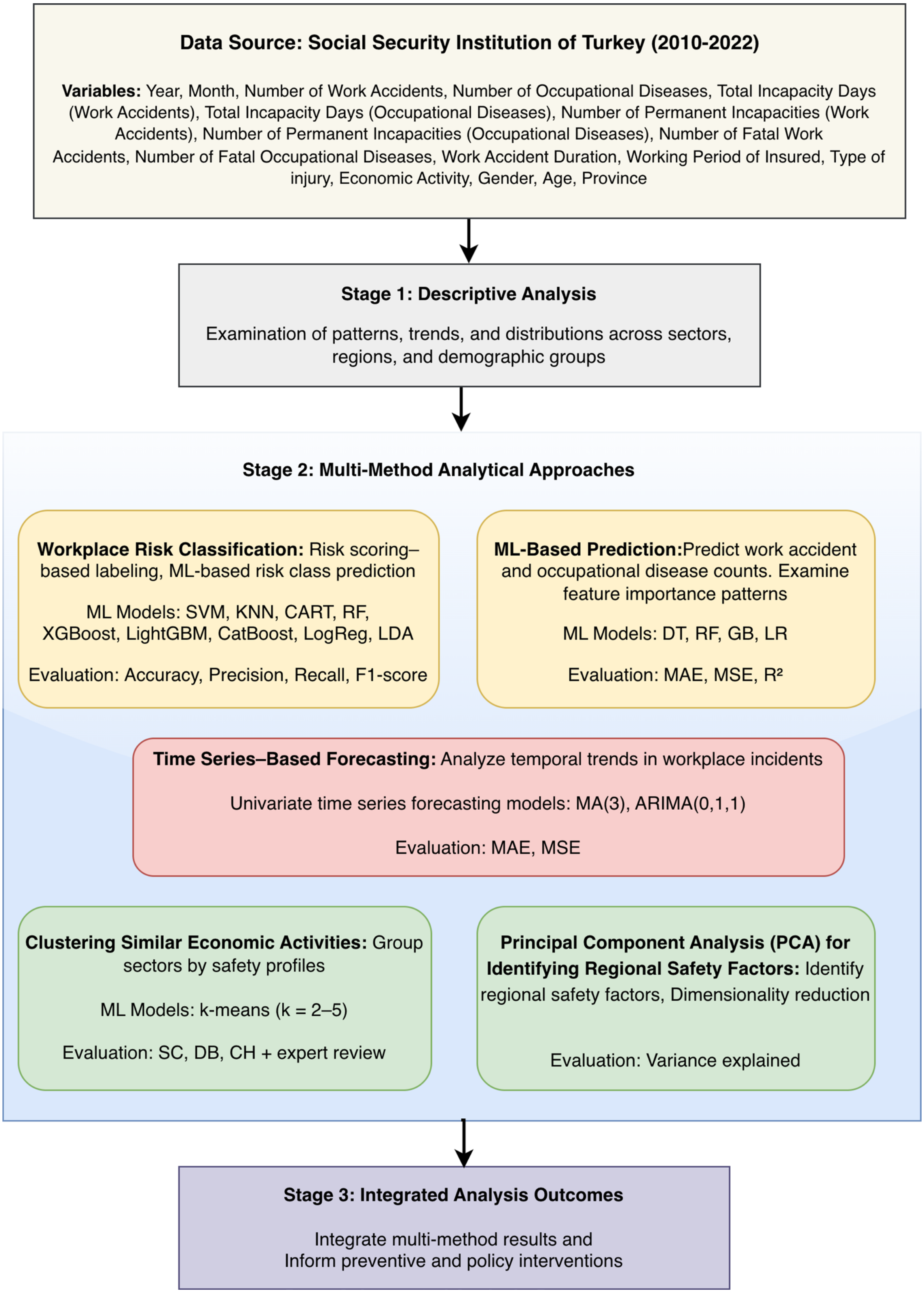


Fig. 1 Multi-method analytical framework

- Number of fatal work accidents: The mean value was chosen as the threshold, reflecting a standard baseline for identifying higher-risk cases.
- Number of work accidents: The 75th percentile (upper quartile) was selected, capturing the top 25% of data as high-risk.
- Number of occupational diseases: The 75th percentile was used to focus on cases with a higher prevalence.

Each metric was assigned an importance value: 0.5, 0.3, and 0.2, respectively. When a metric exceeded its threshold, it contributed to the overall weighted risk score. Observations corresponding to a specific year, age group, and gender combination with a total risk score above 0.5 were labelled as High-Risk, while the remaining observations were labelled as Low-Risk. This labelling strategy enabled a comprehensive evaluation of both accident severity and frequency.

These risk labels were subsequently used as the target variable for the supervised ML models, while demographic attributes (age group, gender, and year) and workplace incident indicators (number of work accidents, number of occupational diseases, number of fatal work accidents, number of fatal occupational diseases and permanent incapacity outcomes) served as input features to train the classification models. To perform the classification, we employed a variety of algorithms, including classical ML models such as SVM [35], KNN [36], Classification and Regression Trees (CART) [31], and RF [32]. In addition, ensemble methods like RF [32], XGBoost [37], LightGBM [38], and CatBoost [39] were used to capture complex interactions and improve classification accuracy. We also utilized traditional classification methods such as LogReg [40] and Linear Discriminant Analysis (LDA) [41] to establish performance benchmarks. These models were chosen to ensure that both linear and non-linear relationships in the data were adequately addressed. These classification models were systematically evaluated using accuracy, precision, recall, and F1-score metrics to identify the most effective approaches. To prevent overfitting and underfitting and reduce evaluation bias associated with a single data split, a 10-fold cross-validation strategy was implemented across all supervised models, as recommended in empirical ML research [42, 43].

Time Series-based Forecasting. To complement the ML approaches, two univariate time series forecasting techniques were applied to analyse the temporal dynamics of workplace incidents. Unlike the ML models, which rely on multiple predictors, these time series models use only historical values of the target variables, offering an alternative analytical perspective. Combining both feature-based ML and temporal modelling enhanced the robustness of the analysis and contributed to a more comprehensive risk assessment framework. In this study, Moving Average (MA) [44] was selected for its simplicity and interpretability, smoothing short-term fluctuations to reveal broader patterns. In addition, Auto Regressive Integrated Moving Average (ARIMA) [44] was employed to capture more complex temporal dependencies. Specifically, we implemented the MA (3) and the ARIMA (0,1,1) models for each target variable. The ARIMA (0,1,1) configuration was selected based on its lowest Akaike Information Criterion (AIC) among candidate models, indicating the best fit for capturing the underlying structure of the time series after differencing. This model is particularly effective for non-stationary data with a consistent trend, as observed in the work accident data. On the other hand, the 3-year moving average (MA (3)) was chosen for its simplicity and interpretability, providing a smoothed representation of short-term trends without requiring model fitting, making it a practical baseline for comparison. Performance of these time series models was evaluated using MAE and MSE.

*Clustering Similar Economic A*ctivities. For economic activities categorization, the k-means clustering algorithm [45] was applied to identify distinct risk groups based on safety profiles, enabling targeted policy interventions. As an unsupervised learning method, k-means requires the number of clusters (k) to be specified a priori. Since no ground-truth labels are available in this context, the optimal value of k cannot be determined directly and must be inferred using internal performance metrics. Accordingly, we evaluated multiple candidate solutions (k = 2, 3, 4, and 5) using complementary data-driven metrics that capture different aspects of clustering quality. Clustering performance was assessed using the Silhouette Coefficient (SC) [46], Davies–Bouldin Index (DB) [47], and Calinski–Harabasz Index (CH) [48]. These performance metrics emphasize distinct properties of cluster structure: SC reflects cohesion and separation, with higher values indicating more cohesive and well-separated clusters; DB measures average cluster compactness relative to separation, where lower values indicate better clustering quality; and CH evaluates between-cluster dispersion relative to within-cluster variance, with higher values favoring solutions that exhibit greater separation between clusters relative to their internal variability.

In addition to quantitative performance evaluation, an expert-based interpretability assessment was incorporated as part of the clustering analysis. Three academic experts with domain knowledge in OHS and industrial engineering participated in this process. The experts examined the clustering solutions highlighted by each performance metric and were asked to assess them independently with respect to interpretability, conceptual coherence, and practical relevance. This was followed by a structured discussion to reconcile perspectives and establish a shared qualitative assessment.

Principal Component Analysis (PCA) for Identifying Regional Safety Factors. PCA [49] was employed to identify key regional safety factors by reducing dimensionality and revealing underlying geographical patterns in workplace safety data. The dataset includes quantitative variables such as the number of work accidents, occupational diseases, incapacity days, and fatalities for each province. PCA was applied after standardizing the data to ensure all variables contribute equally to the analysis.

4 Results

This section presents descriptive analysis, predictions, sectoral risk categorization, regional safety assessments, and workplace risk classification.

4.1 Descriptive Analysis

Table 4 summarizes the descriptive statistics of work accidents and occupational diseases across different economic activities and provinces in Turkey, computed at the year–economic activities–gender and year–province–gender observation levels, respectively. The findings show that work accidents are considerably more frequent and severe than occupational diseases across both dimensions. The high standard deviations indicate substantial variability, particularly in accident frequency, incapacity days, and fatal incidents, suggesting uneven distribution of occupational risks across sectors and geographic regions. To identify distributional patterns in the occurrence of work accidents and occupational diseases, sector- and province-level visual analyses were conducted (see Appendix A, Figs. 11, 12, 13 and 14). The results indicate that work accidents are largely concentrated in construction- and manufacturing-related activities, whereas occupational diseases are more prevalent in mining and heavy industry sectors. At the regional level, work accidents are primarily clustered in industrial provinces such as Istanbul, Izmir, and Ankara, while occupational diseases are disproportionately observed in provinces with intensive mining activities, particularly Zonguldak.

Table 4 Descriptive statistics of work accidents and occupational diseases across economic activities and provinces (2010–2022)

	Economic Activities		Provinces	
	Mean	Standard Deviation	Mean	Standard Deviation
Number of Work Accidents	1675.95	3883.25	1825.56	6203.34
Number of Occupational Diseases	3.29	14.27	1.92	12.67
Total Incapacity Days (Work Accident)	16931.56	42,167.41	19,160.34	56,128.90
Total Incapacity Days (Occupational Disease)	15.07	194.88	16.50	234.76
Number of Fatal Work Accidents	8.55	29.53	5.86	20.31
Number of Fatal Occupational Diseases	0.03	0.59	-	-
Number of Permanent Incapacity (Work Accident)	18.65	60.36	5.76	22.92
Number of Permanent Incapacity (Occupational Disease)	0.98	4.17	16.05	48.83

Table 5 Work accidents and occupational diseases by age and gender (2010–2022)

	Female		Male	
Ages	# of Work Accidents	# of Occupational Diseases	# of Work Accidents	# of Occupational Diseases
-14	307	1	2104	23
15–17	15,111	0	40,436	0
18–24	214,580	9	709,299	64
25–29	136,101	15	587,510	188
30–34	120,283	23	533,775	336
35–39	125,492	28	468,526	560
40–44	110,497	49	370,625	709
45–49	74,012	54	264,341	573
50–54	34,296	47	133,693	310
55–59	11,715	23	60,813	149
60–64	2738	16	18,772	114
65+	705	17	5065	269
Total	845,837	282	3,194,959	3295

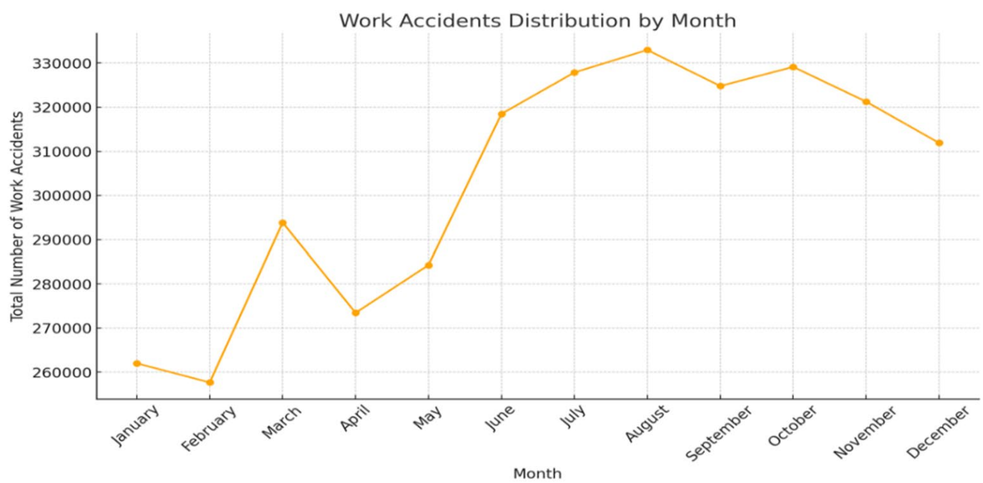


Fig. 2 Monthly distribution of work accidents (2010–2022)

Table 5 presents the distribution of work accidents and occupational diseases by age and gender. The results show that both incidents are considerably more frequent among male workers across all age groups. Work accidents peak in younger and middle-aged groups (particularly ages 18–39), whereas occupational diseases are more common in older age groups, especially after age 40. This pattern suggests that work accidents tend to occur earlier in working life, while occupational diseases accumulate over prolonged exposure to hazardous working conditions.

Figure 2 presents the monthly distribution of work accidents. Work accidents show a clear seasonal pattern, with lower counts in the early months of the year and a peak during summer (June–August), followed by a gradual decline toward the end of the year.

Figure 3 illustrates the monthly distribution of occupational diseases from January to December. Unlike work accidents, which display a clear seasonal pattern, occupational diseases do not follow a distinct seasonal trend and instead fluctuate throughout the year, likely reflecting delayed diagnosis and cumulative exposure rather than immediate incidents.

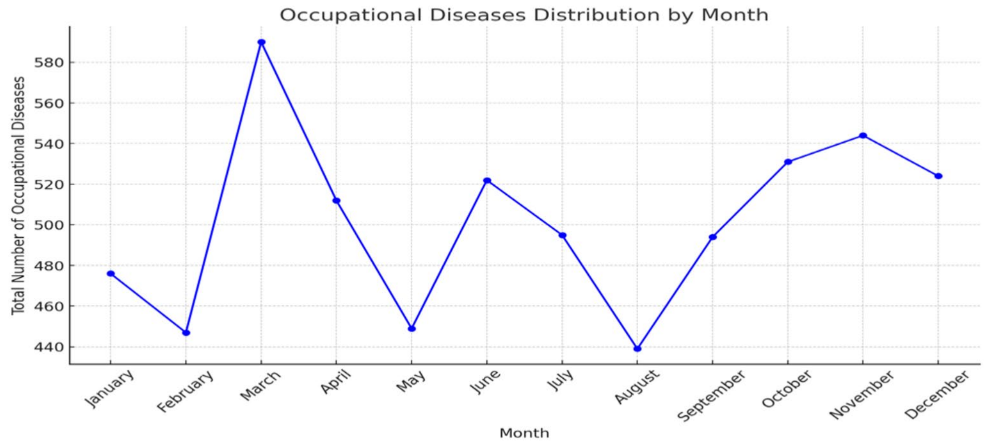


Fig. 3 Monthly distribution of occupational diseases (2010–2022)

Table 6 Performance of ML models for predicting work accidents and occupational diseases

Target Variable	Model	MAE	MSE	R^2
Number of Work Accidents	DT	471.42	2549695.44	0.8605
	RF	353.37	1215936.88	0.9334
	GB	391.02	1014090.74	0.9445
	LR	722.91	2091387.69	0.8856
Number of Occupational Diseases	DT	2.808	153.406	0.0047
	RF	2.226	109.48	0.2897
	GB	2.329	113.43	0.2641
	LR	2.909	113.86	0.2613

4.2 ML-based Prediction of Work Accidents and Occupational Diseases

The predictive performance of the ML models differed substantially between the two target variables. As presented in Table 6, models trained to predict the number of work accidents achieved notably strong results, with ensemble methods outperforming the other approaches across all evaluation metrics. GB and RF exhibited the lowest error values (MAE and MSE) and the highest explanatory power (R^2), indicating their effectiveness in capturing meaningful and complex relationships within the dataset. Although the LR achieved reasonably good predictive performance ($R^2 = 0.8856$), it did not match the capabilities of ensemble-based models. In contrast, the DT showed higher errors and lower predictive capability, reflecting limited generalization.

In comparison, the prediction of occupational diseases resulted in considerably lower performance across all models. RF achieved the highest explanatory strength ($R^2 = 0.2897$), followed by GB ($R^2 = 0.2641$), while DT and LR models demonstrated minimal predictive capability. To assess whether class imbalance contributed to this decline, SMOTE and random undersampling techniques were applied during model development. However, neither method improved predictive outcomes meaningfully, suggesting that the primary limitation originates from a weak predictive signal in the available features rather than class imbalance effects alone.

The feature importance analysis shown in Figs. 4 and 5 demonstrates how the relevance of predictors shifts depending on the target variable. For the prediction of the number of work accidents, the most influential variable was total incapacity days (work accident), indicating a strong association between accident frequency and total incapacity days. Additional contributors, such as year and economic activity, also played a role, although their relative importance was substantially lower. These results reinforce the critical need for preventive safety strategies, as reducing accidents may directly decrease incapacity time and associated economic and human costs.

Figure 5 shows that the most influential predictors for occupational diseases were the number of permanent incapacities and total incapacity days (occupational disease). Variables related to work accidents, such as the number of work accidents and total incapacity days (work accident), also contributed moderately, while the

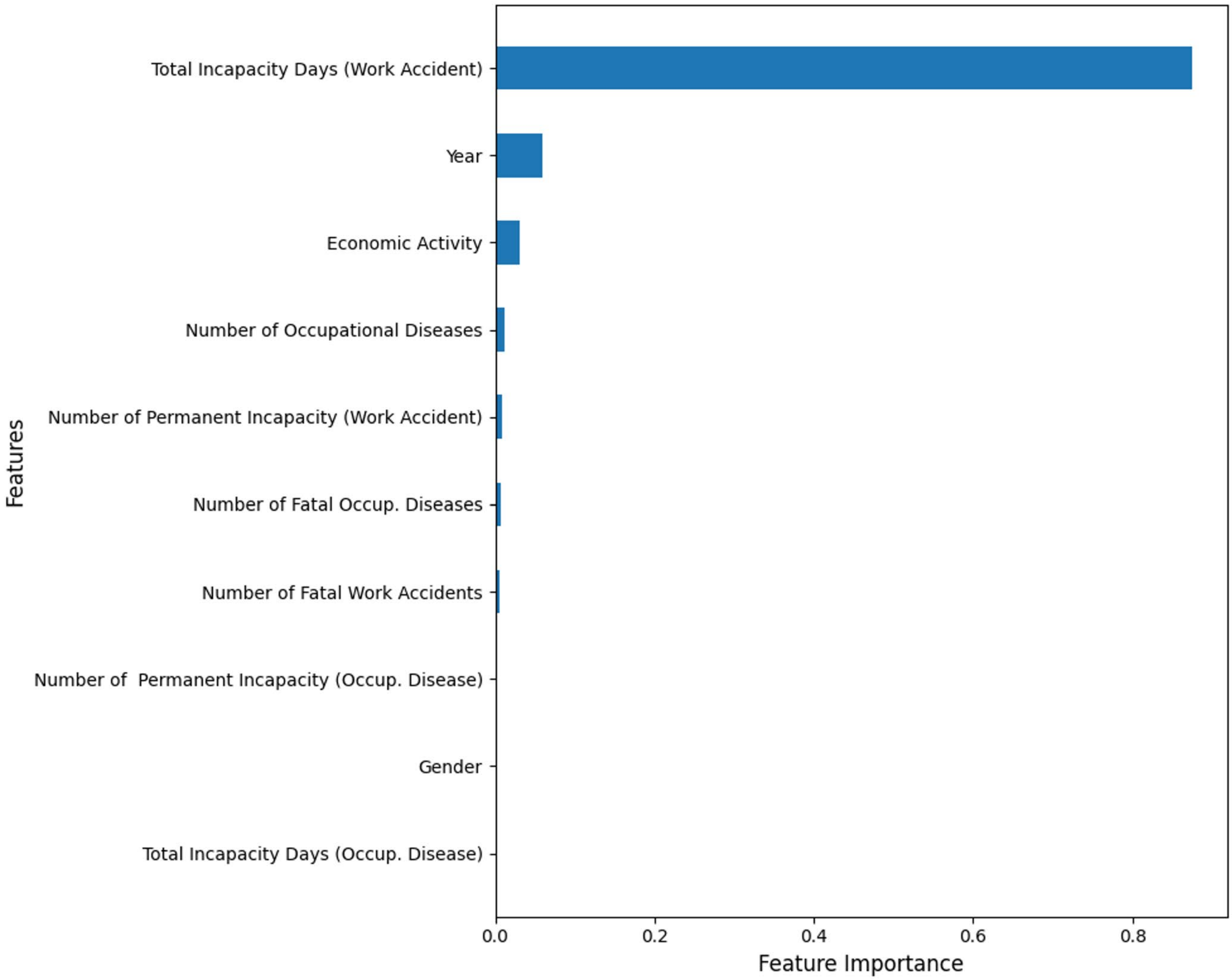


Fig. 4 Feature importance for the number of work accidents

remaining predictors had limited influence. Overall, the feature pattern suggests that occupational disease reporting in the dataset is more strongly associated with recorded incapacity measures than with demographic or temporal factors.

4.3 Workplace Risk Classification

Based on the weighted risk scoring framework described in Sect. 3.2, the predefined thresholds and weights resulted in a clear separation of observations into distinct risk groups. The resulting risk distribution was imbalanced: 90 group-level observations were classified as High-Risk, while 222 were classified as Low-Risk. These risk categories were subsequently used as the target labels for ML-based workplace risk classification.

Table 7 summarizes the model performance for workplace risk classification. Among all tested models, LightGBM achieved the best performance, with an accuracy of 98.4%, followed closely by CatBoost and RF. These results demonstrate the effectiveness of ensemble methods in handling the complexities of the dataset.

When traditional statistical methods are considered, LogReg and LDA achieved relatively high accuracy scores (95.19% and 93.6%, respectively); however, both were outperformed by the advanced ML models. This

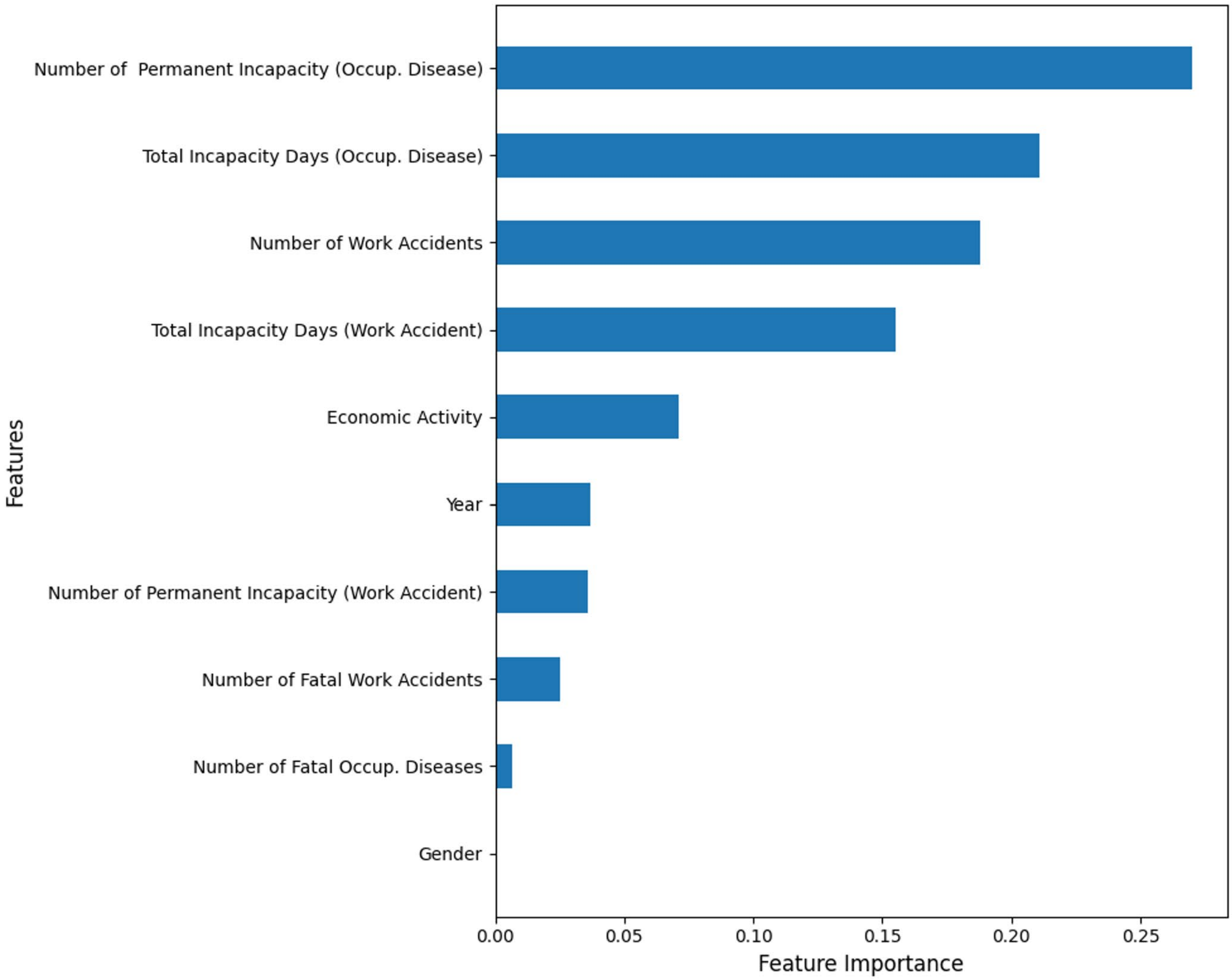


Fig. 5 Feature importance for the number of occupational diseases

Table 7 Performance of classification models for workplace risk classification

Model	Accuracy	Recall	Precision	F1-score
SVM	0.9423	0.9423	0.9492	0.9424
KNN	0.8527	0.8527	0.8773	0.8342
CART	0.9679	0.9679	0.9697	0.9683
RF	0.9711	0.9711	0.9732	0.9710
XGBoost	0.9646	0.9646	0.9686	0.9653
LightGBM	0.9840	0.9840	0.9854	0.9842
CatBoost	0.9807	0.9807	0.9822	0.9809
LogReg	0.9519	0.9519	0.9550	0.9521
LDA	0.9360	0.9360	0.9420	0.9365

Table 8 Performance of forecasting models for predicting work accidents and occupational diseases

Target Variable	Model	MAE	MSE
Number of Work Accidents	MA (3)	50,219	3,398,272,934
	ARIMA (0,1,1)	43,714	3,283,588,372
Number of Occupational Diseases	MA (3)	116.18	22,657
	ARIMA (0,1,1)	210.43	82,481

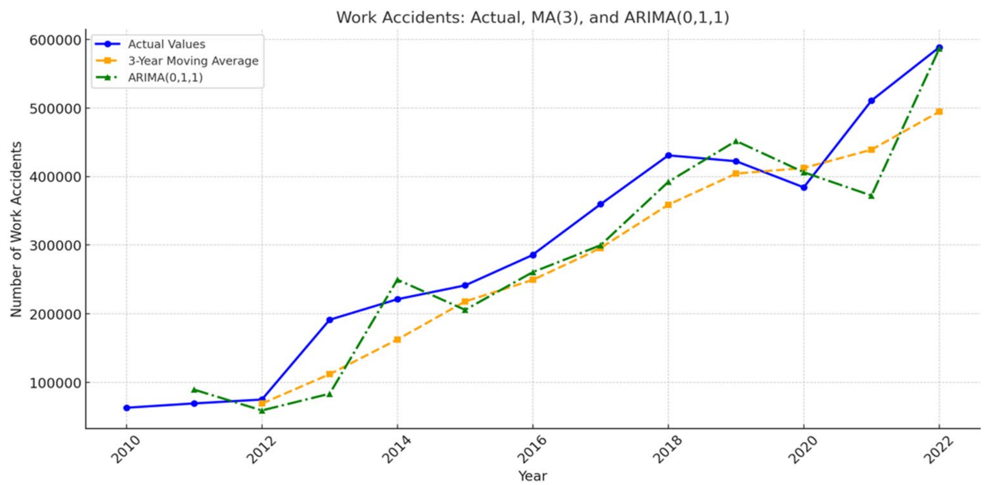


Fig. 6 Comparison of actual values of work accidents with MA (3) and ARIMA (0,1,1) prediction

comparison demonstrates that while LogReg and LDA offer ease of use, their classification performance is relatively limited when dealing with high-dimensional and non-linear workplace safety data.

4.4 Time Series-based Forecasting of Work Accidents and Occupational Diseases

Table 8 presents the forecasting results obtained from the time series models. For the number of work accidents, ARIMA (0,1,1) achieved slightly lower MAE and MSE values compared with the MA (3) model, indicating a marginally better fit to the temporal structure of the data. For occupational diseases, the MA (3) model outperformed ARIMA (0,1,1), yielding lower error values. The weaker performance of ARIMA (0,1,1) suggests that the occupational disease time series may exhibit irregular patterns and limited autocorrelation, which reduces the benefit of more complex time series structures. Therefore, simpler smoothing-based approaches appear to be more appropriate for forecasting this variable.

Figures 6 and 7 illustrate the comparison between the actual values and the predictions obtained by the MA (3) and ARIMA (0,1,1) for work accidents and occupational diseases, respectively. These visualizations reinforce the quantitative results by highlighting the relatively better alignment of ARIMA (0,1,1) with the temporal structure of work accident data, while MA (3) provides a more stable fit for the less structured occupational disease trends.

These findings highlight a critical distinction between the two target variables: while work accidents follow more stable and predictable temporal patterns, occupational diseases demonstrate more erratic behavior, potentially due to their long latency periods and complex etiology. Thus, the integration of both ML and time series models provides a comprehensive understanding of the predictability landscape, where ML models benefit from external features and ARIMA-based models offer insights into intrinsic temporal dynamics.

4.5 Clustering Similar Economic Activities Based on Accident Characteristics

To cluster similar economic activities based on work accident and occupational disease characteristics and better understand risk profiles, we employed k-means clustering algorithms. Determining the optimal number of clusters (k) is crucial in k-means clustering, requiring the user to specify the number of clusters to generate. While

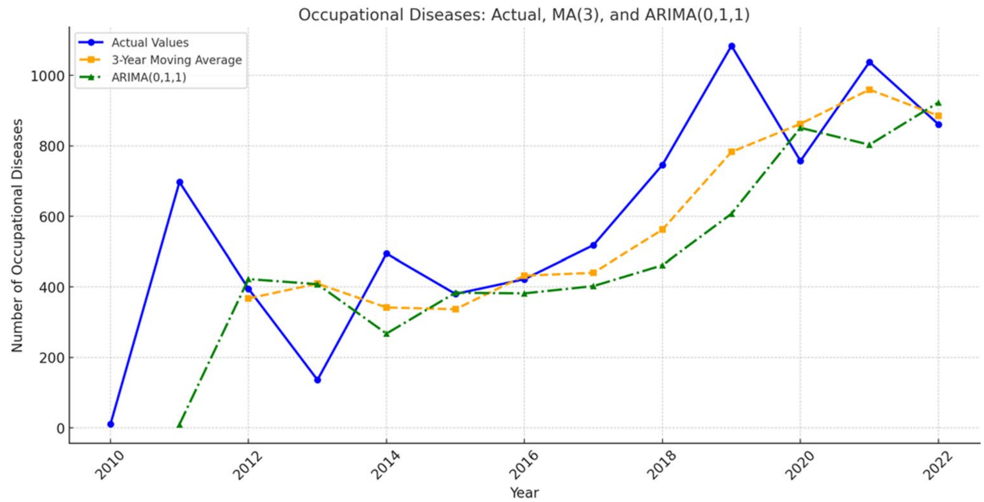


Fig. 7 Comparison of actual values of occupational diseases with MA (3) and ARIMA (0,1,1) prediction

the optimal number is somewhat subjective, it depends on the dataset's distribution, scale, and the user's desired resolution. Common methods for determining k include data-driven techniques and leveraging domain knowledge [50]. In our analysis, we applied the elbow method, adjusted inertia, silhouette analysis, and gap statistics as data-driven techniques to identify the optimal k.

Figure 8 shows an example elbow method derived from the k-means algorithm. The chart shows how the distortion score (within-cluster sum of squares), represented by the blue line, decreases as k increases, signifying improved clustering. The "elbow point" at $k = 5$, marked by the black dashed line, indicates where the rate of decrease slows significantly. The green dashed line represents the computation time for each k, which remains minimal across all k values.

Analyses suggested that two, three, four, or five clusters could be suitable options for grouping the data. Table 9 compares the outcomes of the k-means clustering algorithm for these cluster numbers, evaluated using three performance metrics: SC, DB, and CH. Higher values indicate better performance for SC and CH, whereas lower values indicate better performance for DB. Among the four scenarios, the highest SC (0.6736) is achieved with $k = 3$, indicating that this configuration provides the most cohesive and well-separated clusters. The lowest DB value (0.7689) occurs at k=3, suggesting more compact and well-separated clusters. CH indicates that k=2 is the best option, as it has the highest CH value among the four scenarios.

Scenarios with two and three clusters yielded the highest performance outcomes based on data-driven methods. To refine the results, experts with field experience reviewed the clusters generated in these scenarios. Their evaluation concluded that the three-cluster scenario produced more meaningful and interpretable groupings. Based on this analysis, economic activities were categorized into three distinct clusters, which are depicted in Fig. 9.

- Cluster 1 Low-Risk Sectors: This cluster includes industries such as agriculture, forestry, and fishing, as well as other sectors that exhibit low-risk characteristics. These sectors are notable for their minimal accident rates, very few occurrences of occupational diseases, and an almost negligible number of fatal incidents. As a result, they are considered the least hazardous group within the overall classification.
- Cluster 2 Medium-Risk Sectors: Industries in this cluster include manufacturing, transportation, and certain resource extraction activities. These economic activities are characterized by moderate accident rates and a noticeable frequency of permanent incapacities among workers. The risks in these industries often stem from the use of machinery and logistical operations, which necessitate careful monitoring and the implementation of effective hazard reduction measures to ensure worker safety.
- Cluster 3 High-Risk Sectors: This cluster is primarily composed of high-risk industries such as coal mining and heavy industrial processes. These economic activities stand out due to their high accident rates, frequent fatalities, and significant occurrences of occupational diseases. Given the inherently dangerous working conditions

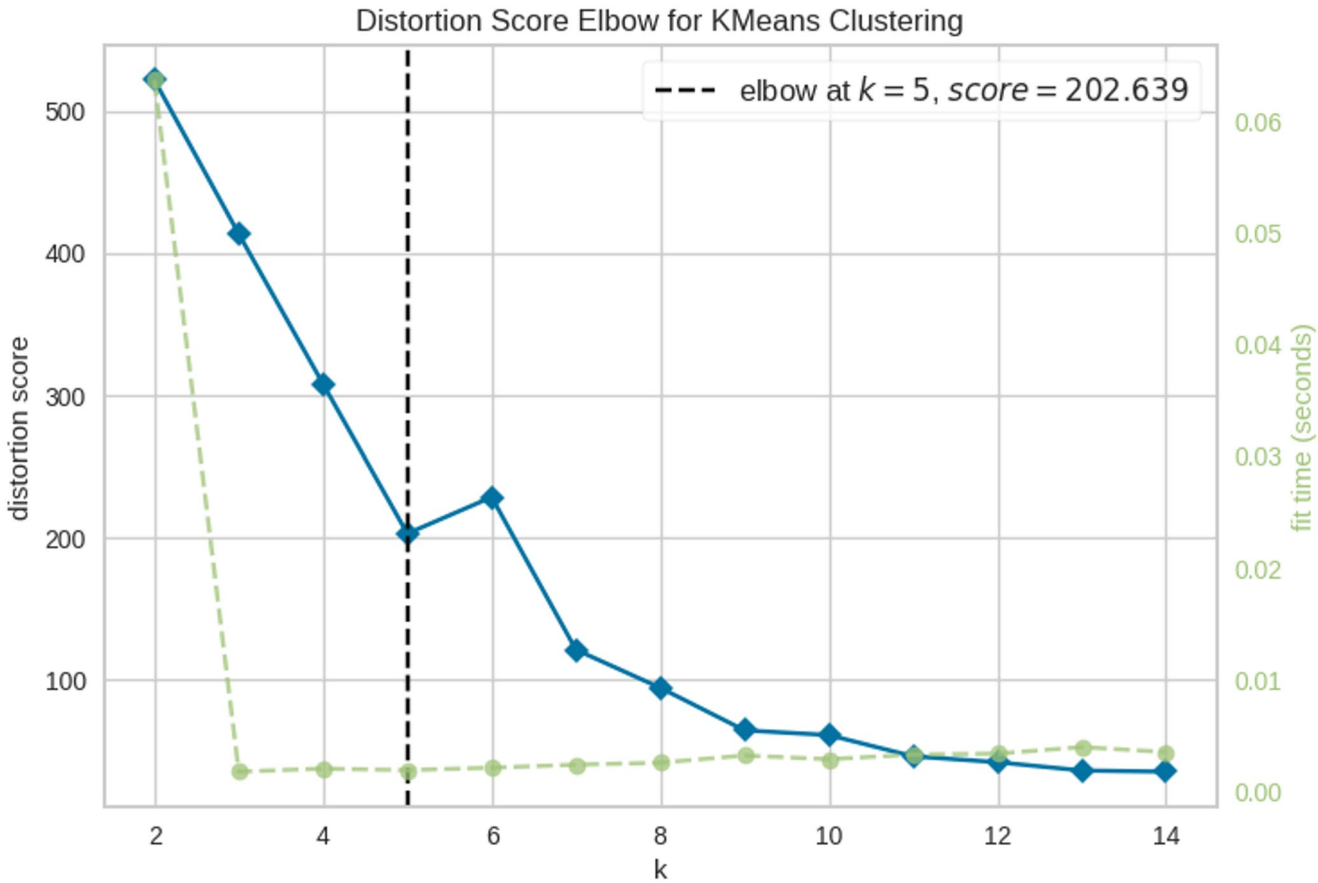


Fig. 8 Distortion score elbow for k-means clustering

Table 9 A comparison of the performance of k-means for k = 2, k = 3, k = 4, and k = 5

	k-means			
	k = 2	k = 3	k = 4	k = 5
SC	0.6701	0.6736	0.6482	0.5447
DB	1.0660	0.7689	1.0545	1.0693
CH	59.367	47.945	53.243	50.207

in these industries, rigorous safety measures and comprehensive risk management strategies are essential to protect workers and reduce the likelihood of adverse outcomes.

4.6 PCA for Identifying Regional Safety Factors

The PCA results reveal clear regional variation in workplace safety indicators, allowing provinces to be grouped based on shared patterns rather than individual metrics. The first two principal components, which account for 93.06% of the total variance, were used to visualize the provinces in a two-dimensional space.

Figure 10 presents the PCA results, which visually reveal relationships between variables as well as similarities and differences among provinces. Principal Component 1 (PC1) explains the majority of the variance (82.27%) and is positively correlated with almost all variables, particularly: the number of work accidents, total incapacity days (work accident), number of permanent incapacity (work accident), and number of permanent incapacity (occupational disease). These variables have high positive loadings, suggesting that PC1 primarily represents the

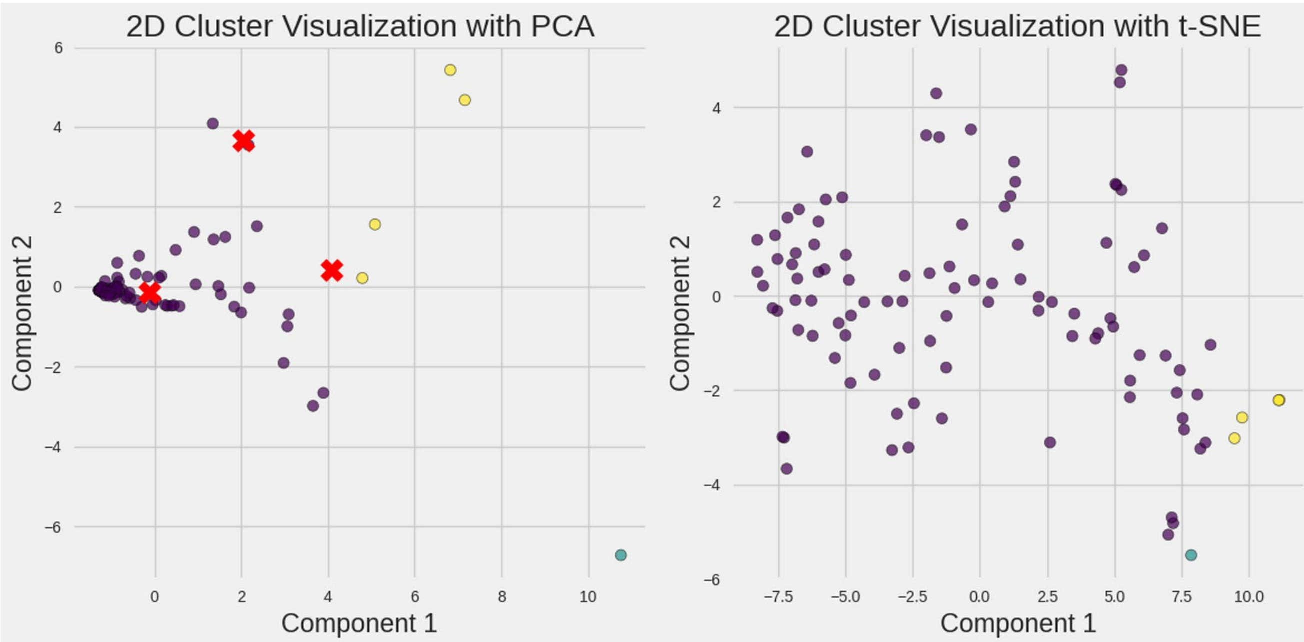


Fig. 9 Representation of economic activities in 2D space based on the three clusters. (Cluster 1: purple, Cluster 2: yellow, Cluster 3: green)

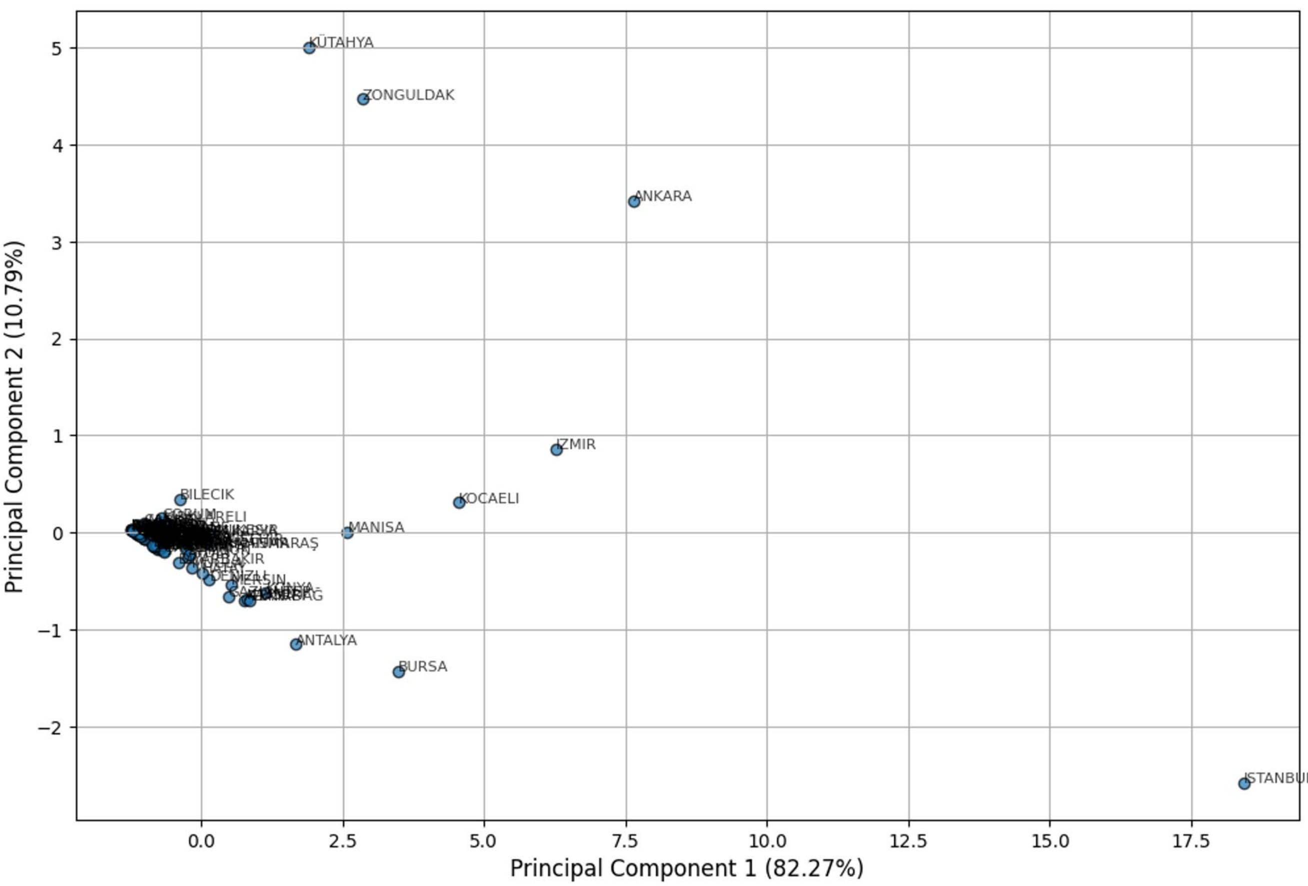


Fig. 10 PCA of provinces by work accidents and diseases

overall severity of work-related accidents and their consequences in terms of incapacity. Principal Component 2 (PC2) captures additional variance (10.79%) and is mainly influenced by: the number of occupational diseases and total incapacity days (occupational disease). These variables have strong positive loadings on PC2, while most other variables have negative or weak contributions.

Provinces like Bilecik, Burdur, Erzurum, and Çankırı cluster near the origin. This indicates that these provinces have similar and relatively low levels of work accidents, occupational diseases, and incapacity days. Their profiles are comparable in terms of overall workplace safety. Positioned far to the right along PC1, İstanbul has significantly higher values across most variables, particularly in work accidents and related incapacity days. Its large population and industrial activity likely contribute to these elevated values. Provinces like Manisa and Bursa are positioned between the origin and İstanbul on PC1, indicating moderately elevated levels of work-related incidents compared to the majority of provinces, but not as extreme as İstanbul. Located further along PC2, Ankara shows a higher relative incidence of occupational diseases and related incapacity days, distinguishing it from other provinces. Izmir and Kocaeli are also moderately separated from the main cluster, reflecting higher values for work accidents and incapacity days, albeit to a lesser extent than İstanbul. Kütahya and Zonguldak are positioned distinctly high on the PC2, reflecting a notable prevalence of occupational diseases and incapacity days associated with these diseases. Both provinces are known for industrial and mining activities, which inherently carry higher risks for such outcomes. Their similar profiles, influenced by the nature of their dominant industries, explain their proximity in PCA space, yet distinct separation from the main cluster of provinces.

5 Discussion

5.1 Interpretation of Key Findings

This study provides a multi-perspective analysis of Turkey's OHS landscape using nationwide records from 2010 to 2022. The descriptive results show strong heterogeneity across sectors, demographics, and regions. The findings highlight significant demographic disparities, with younger workers disproportionately affected by workplace accidents and older workers more vulnerable to occupational diseases, while male workers face substantially higher incident rates overall. Work accidents are concentrated in construction- and manufacturing-related activities, whereas occupational diseases are more prominent in mining and other heavy-industry contexts. This contrast aligns with the differing nature of acute incidents versus long-latency health outcomes and highlights why sector-specific prevention strategies are necessary rather than uniform policies.

A key finding is the clear performance gap between predicting work accidents and predicting occupational diseases. Ensemble-based ML models achieved strong predictive performance for the number of work accidents (with GB reaching the highest explanatory power, $R^2 = 0.9445$), indicating that accident counts exhibit stable relationships with the available predictors. However, the relatively lower predictive power for occupational diseases (with RF reaching the highest explanatory power, $R^2 = 0.2897$) suggests that these incidents are influenced by more complex and long-term factors.

Beyond prediction, the unsupervised analyses support a structured understanding of risk profiles. The k-means clustering results indicate that economic activities can be grouped into three interpretable clusters representing low-, medium-, and high-risk sector profiles based on shared patterns in incident frequency, severity, and health consequences. These groupings provide a practical basis for tailoring inspections, enforcement intensity, and sector-focused training and prevention programs.

The PCA results further emphasize regional differentiation. Provinces separate along a dominant component associated with overall accident burden and its consequences (e.g., accidents and incapacity days), with Istanbul standing out as an extreme case consistent with its scale of industrial and workforce activity. A second component is driven more by occupational disease indicators, distinguishing provinces such as Zonguldak and Kütahya, which is consistent with mining and heavy-industry exposure patterns. Together, these findings indicate that regional OHS profiles are multi-dimensional and that interventions should consider whether a province's risk profile is driven primarily by accident burden, occupational disease burden, or both.

Temporal analyses indicate clear monthly and seasonal variation in incident rates, with workplace accidents peaking during summer months, whereas occupational diseases show no clear seasonal pattern and fluctuate across the year. Complementary to these findings, time series models (ARIMA and MA) applied to yearly data suggest that workplace accidents follow a more predictable temporal trend, while occupational diseases demonstrate more irregular yearly variation. These results underscore the existence of both monthly and annual dynamics in occupational risk patterns, with implications for the timing of inspections, training, and monitoring efforts.

5.2 Practical and Policy Implications

The study's statistical and ML findings have produced concrete evidence to guide decision-makers in critical areas such as resource allocation, risk classification, audit prioritization, and training planning in OHS management. One of the key takeaways is the importance of sector-specific safety strategies. Since the study identified significant variations in workplace accident rates across different industries, policymakers should avoid a one-size-fits-all approach to OHS regulations. High-risk industries, such as coal mining and heavy manufacturing, require stricter enforcement of safety protocols, increased inspections, and mandatory adoption of advanced protective equipment. Meanwhile, industries with lower accident rates, such as agriculture and fisheries, could benefit more from awareness campaigns and targeted training programs rather than overly strict regulatory measures that may not align with their unique work conditions.

Another key implication relates to the role of ML in proactive risk management. The high predictive accuracy of GB and RF models for workplace accidents demonstrates that organizations can use AI-driven insights to anticipate and prevent future accidents. Managers can integrate these predictive models into their decision-making processes, allowing them to allocate safety resources more efficiently. For example, if the ML model indicates an elevated expected accident burden in a specific region or industry, regulatory bodies can conduct more frequent safety audits or mandate additional safety training in those locations. This kind of data-driven decision-making would not only improve workplace safety but also reduce the financial burden of workplace injuries on businesses and the healthcare system.

Monthly and seasonal risk cycles identified in temporal analyses enable public authorities and business managers to schedule inspections, occupational safety training, and protective equipment procurement more effectively. These temporal patterns in workplace incidents also indicate that safety policies should be time-sensitive—for example, accident prevention efforts and inspections can be intensified during summer months when incidents peak, while occupational disease monitoring may require greater attention throughout the year.

From a regional perspective, the study's findings indicate that provinces such as Istanbul, Kocaeli, and Izmir have higher risks of workplace accidents due to their industrial density. This suggests that regional authorities and local governments should take a more active role in occupational safety policies. Investing in regional safety education programs, developing industry-specific safety guidelines, and offering financial incentives to companies that implement effective risk-reduction measures could significantly lower accident rates. Moreover, collaboration between government agencies, industry associations, and academic institutions can create a more structured and standardized approach to workplace safety management, ensuring that preventive measures are continuously updated based on emerging trends.

Demographic differences—namely, the higher accident risk among younger workers and the greater prevalence of occupational diseases among older employees—underscore the need for age-sensitive safety policies. Accordingly, organizations and human resources departments should offer targeted training for younger staff and implement long-term health monitoring and preventive programs for aging employees.

A major takeaway for human resource and safety managers is the importance of workplace risk classification. In this study, the best-performing ML-based classification models (LightGBM and CatBoost) achieved accuracy levels above 98% in distinguishing high-risk and low-risk groups derived from aggregated workplace data. This means that organizations can leverage AI-driven tools to identify worker groups that may need additional training, protective measures, or job modifications to ensure their safety. However, it is crucial that such classifications are not used in a discriminatory manner but rather as a tool to enhance safety practices. Employers should be transparent about how these models are used and ensure that workplace interventions remain fair and ethical.

Finally, the study underscores the need for continuous improvement in occupational safety policies. While regulations are essential, companies must foster a culture of safety that goes beyond compliance. Organizations should invest in technology-driven safety solutions, such as IoT-based hazard monitoring systems, wearable safety devices, and AI-driven risk assessment tools. Moreover, safety training should be made more engaging and accessible, using interactive digital platforms and virtual reality simulations to prepare workers for real-life scenarios.

5.3 Strengths and Limitations

One of the main strengths of this study is the use of long-term nationwide OHS data covering Turkey for the period 2010–2022, analyzed through the combined application of statistical analysis and ML techniques (SVM, KNN, CART, RF, XGBoost, LightGBM, CatBoost, LogReg, LDA, DT, GB). Moreover, the analyses incorporate time series–based forecasting to examine trends in workplace incidents, k-means clustering to group similar economic activities, and PCA to identify regional safety-related factors. This integrated approach enables the examination of workplace accidents and occupational diseases not only from a descriptive perspective but also through a predictive lens. In addition, the simultaneous consideration of sectoral, regional, temporal, and demographic dimensions broadens the scope of the study and provides multidimensional and actionable insights for policymakers and practitioners.

Much of the existing OHS research in Turkey has examined specific sectors, most notably construction, with ML methods employed to explore targeted research questions. For example, Koç et al. [25] examined the vulnerability of specific body parts among construction workers using ensemble-based ML models, while Koç et al. [12] focused on predicting fatality risk in construction accidents through RF-based approaches. Similarly, Koç and Gurgun [26] proposed a scenario-based preprocessing framework to predict accident severity, and Koç et al. [27] employed time series and wavelet-based ML methods to forecast accident frequencies within the construction sector. In contrast, the present study adopts a broader, nationwide perspective by analyzing long-term OHS data across multiple economic sectors and regions in Turkey. Rather than concentrating on a single industry or outcome, this study integrates descriptive statistics, time series analysis, clustering techniques, and multiple ML models to simultaneously examine accident occurrences, occupational diseases, and risk patterns across demographic, sectoral, regional, and temporal dimensions. As such, the contribution of this study lies not in direct methodological or performance-based comparison with sector-specific research, but in providing a comprehensive, multi-dimensional framework that complements and extends existing OHS studies in Turkey.

Despite these strengths and the comprehensive nature of the analyses, this study also has several limitations. First, the study is based on secondary data obtained from official records; therefore, potential underreporting—especially for workplace accidents and occupational diseases—cannot be ruled out. In addition, the long-latency nature of occupational diseases may result in cases that have already occurred but have not yet been recorded or reflected in the available data. Moreover, although the ML models employed in this study exhibit strong predictive performance, they primarily identify patterns rather than explaining causal relationships. The analyses are conducted mainly at the sectoral and regional levels and do not include detailed information on firm- or workplace-level safety practices.

5.4 Future Studies

Future studies should focus on integrating additional contextual factors, such as workplace culture, safety training effectiveness, and regulatory enforcement, to further refine predictive models and improve OHS outcomes.

 https://doi.org/10.1007/s44196-026-01351-7

The findings provide a strong foundation for data-driven decision-making, contributing to the long-term goal of reducing workplace hazards and promoting a safer working environment across all industries. Future research on OHS should expand datasets to include workplace safety policies, employee training programs, and compliance levels to refine predictive models. In addition, future studies should investigate the extent to which workplace accidents and occupational diseases are influenced by contextual variables such as working hours, the frequency and quality of safety training, and the use of personal protective equipment (PPE). Integrating these operational and behavioral factors into risk assessment frameworks could significantly enhance predictive accuracy and support the development of more targeted and effective intervention strategies. The integration of real-time monitoring through IoT and wearable devices could enhance accident prevention by providing early hazard detection. More advanced clustering techniques and anomaly detection algorithms could further refine industry risk classifications. Geographical differences in accident rates should be analyzed alongside socioeconomic and policy factors to identify regional trends and best practices. Long-term studies on post-accident recovery, economic impacts, and rehabilitation programs could improve worker support policies. Additionally, ensuring ethical AI implementation in workplace safety is crucial, requiring research on explainable AI (XAI) to make predictive models more transparent and fairer.

6 Conclusion

This study presents a comprehensive data-driven analysis of OHS in Turkey, integrating statistical and ML techniques to assess workplace risks, predict accident occurrences, and identify key determinants of workplace safety. Using nationwide OHS data from 2010 to 2022 and analyzing them, the research provides substantial disparities in the distribution of workplace accidents and occupational diseases across demographic groups, economic sectors, and geographic regions. Overall, this study provides strong evidence that integrating statistical analysis with ML can transform workplace safety management. The high accuracy of the ML models enables the classification of aggregated worker groups into high- and low-risk categories, thereby allowing employers to design proactive risk management practices, targeted training programs, and preventive intervention strategies. In addition, regional and sectoral clustering results offer practical value for policymakers and government agencies by facilitating risk-based audits and targeted regulatory actions. Accordingly, this study not only contributes to the academic literature but also provides a robust, data-driven framework to support policy design, operational planning, and managerial decision-making in the field of OHS. The practical implications suggest that decision-makers must shift from reactive to proactive safety strategies, using data-driven insights to anticipate and mitigate risks before accidents occur. By leveraging these insights effectively, businesses and policymakers can take meaningful steps toward creating safer, more sustainable workplaces in Turkey and beyond.

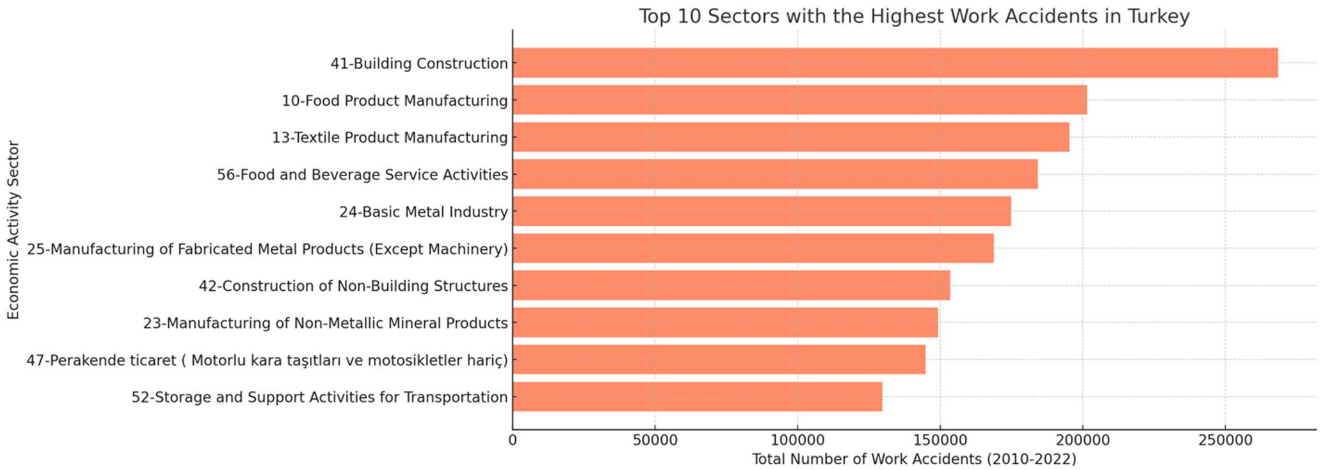


Fig. 11 Top 10 sectors with the highest work accidents in Turkey

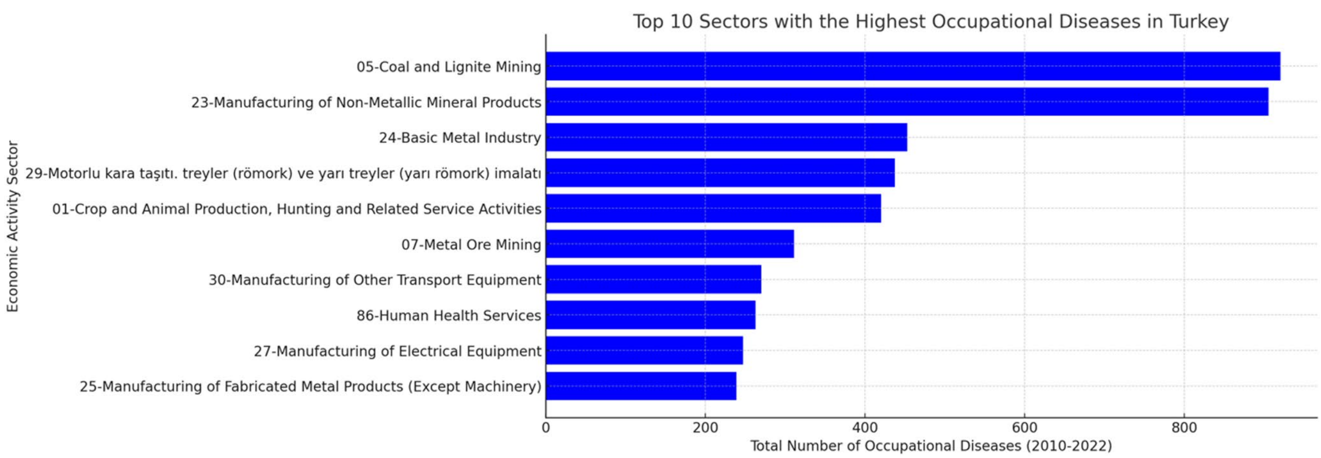
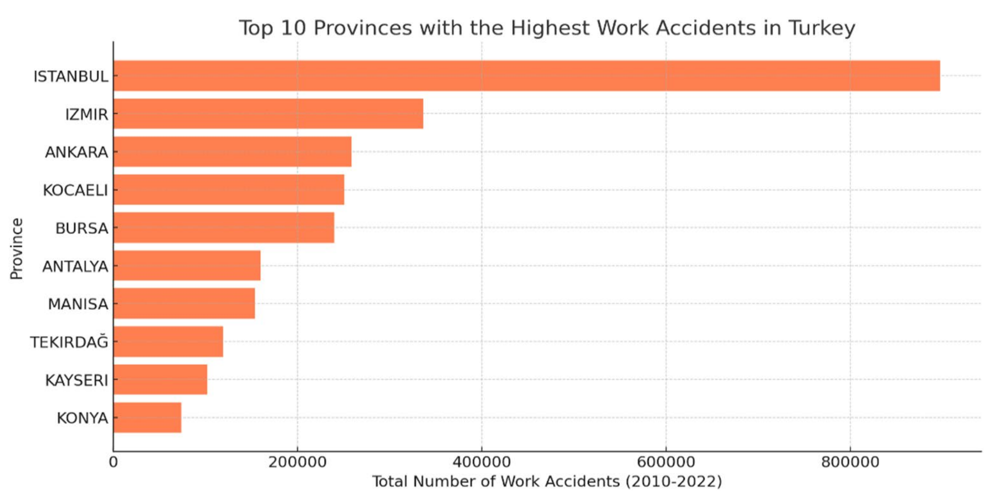


Fig. 12 Top 10 sectors with the highest occupational diseases in Turkey

Fig. 13 Top 10 Provinces with the highest work accidents in Turkey



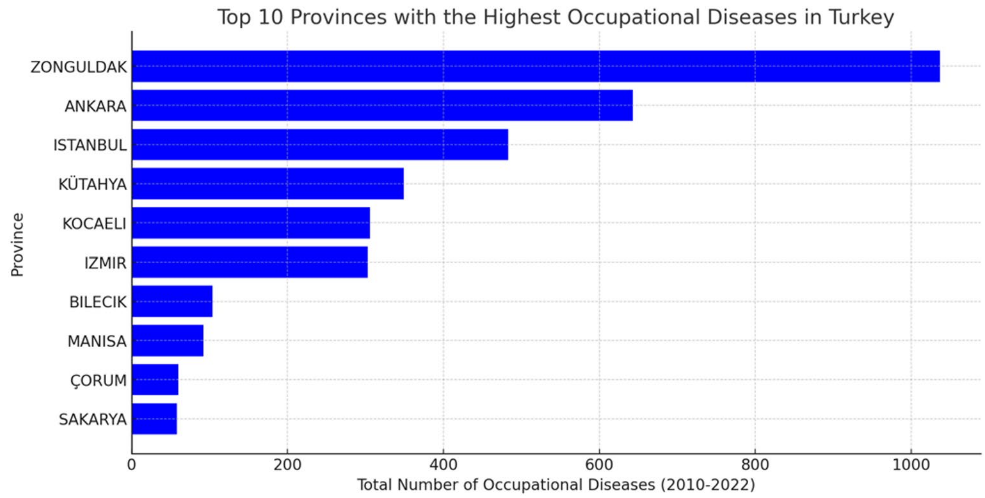


Fig. 14 Top 10 Provinces with the highest occupational diseases in Turkey

Appendix A. Sectoral and Regional Distributions

This appendix provides supplementary descriptive visualizations that support the sectoral and regional analyses presented in Sect. 4.1 (Descriptive Analysis). Figures 11 and 12 display the top 10 economic sectors with the highest numbers of work accidents and occupational diseases, respectively, illustrating the concentration of occupational risks across industries. Figures 13 and 14 present the top 10 provinces with the highest numbers of work accidents and occupational diseases, highlighting regional disparities associated with industrial activity and sectoral composition.

Author Contributions M.A., A.A., and M.D. designed the study and developed the methodology. M.A. performed the data preprocessing and conducted the statistical and machine learning analyses. A.A. led the literature review. M.A. and A.A. drafted the manuscript, while M.D. and I.Y. contributed to result interpretation and manuscript refinement. I.Y. also supported data analysis. M.D. provided conceptual guidance throughout the study. All authors reviewed and approved the final version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received for conducting this study.

Data Availability The data analyzed in this study originate from official national OHS statistics released by public institutions in Turkey. Although the underlying aggregate statistics are publicly available, they are dispersed across multiple sources and formats. For this study, the data were collected, merged, cleaned, and reorganized into a structured dataset suitable for analysis. Researchers interested in accessing the processed version may contact the authors directly.

Declarations

Competing Interests The authors declare no competing interests.

References

1. Frick, K.: Worker influence on voluntary OHS management systems: a review of its ends and means. Saf. Sci. **49**(4), 974–987 (2011)
2. Zanko, M., Dawson, P.: Occupational health and safety management in organizations: a review. Int. J. Manag Rev. **14**(3), 328–344 (2012)
3. International Labour Organization.: Nearly 3 million people die of work-related accidents and diseases. ILO News. (2023). Available at: https://www.ilo.org/resource/news/nearly-3-million-people-die-work-related-accidents-and-diseases. Accessed 02 Feb 2026
4. Nastasia, I., Rives, R.: Successful strategies for occupational health and safety in small and medium enterprises: Insights for a sustainable return to work. J. Occup. Rehabil. **35**, 767-782 (2025)
5. Mustard, C.A., Yanar, B.: Estimating the financial benefits of employers' occupational health and safety expenditures. Saf. Sci. **159**, 106008 (2023)
6. Ateeq, A., Al-Refaei, A.A.A., Alzoraiki, M., et al.: Sustaining organizational outcomes in manufacturing firms: the role of HRM and occupational health and safety. Sustainability. **16**(3), 1035 (2024)
7. Bayram, M., Üngan, M.C.: The relationships between OHS prevention costs, OHSMS practices, employee satisfaction, OHS performance, and accident costs. Total Qual. Manag Bus. Excell. **31**(11–12), 1325–1344 (2020)
8. Sinelnikov, S., Inouye, J., Kerper, S.: Using leading indicators to measure occupational health and safety performance. Saf. Sci. **72**, 240–248 (2015)
9. SocialSecurity Institution [dataset]. Annual statistics on occupational accidents and diseases. (2024). Available from: https://www.sgk.gov.tr/Istatistik/Yillik/fcd5e59b-6af9-4d90-a451-ee7500eb1cb4/. Accessed 24 Nov 2024
10. Özkan, E.K., Ulaş, H.B.: Comparison of four machine learning methods for occupational accidents based on national data on metal sector in Turkey. Saf. Sci. **174**, 106468 (2024)
11. Machfudiyanto, R.A., Kim, S., Latief, Y., et al.: Analysis of design-for-safety implementation factors in the Indonesian construction industry: a two-staged SEM-artificial neural network approach. Heliyon **9**(11), e21273 (2023)
12. Koc, K., Ekmekcioğlu, Ö., Gurgun, A.P.: Developing a national data-driven construction safety management framework with interpretable fatal accident prediction. J. Constr. Eng. Manag. **149**(4), 04023010 (2023)
13. Zhou, J., Liu, L., Su, Y., et al.: Statistical investigation on the characteristics of hazardous chemical accidents in special work in China from 2008 to 2021. Process. Saf. Prog. **43**(1), 195–208 (2024)
14. Soykan, O.: Occupational health and safety in the Turkish fisheries and aquaculture; a statistical evaluation on a neglected industry. Saf. Health Work. **14**(3), 295–302 (2023)
15. Costa da Silva, R.D.C., Pires de Novais, M.A., Zucchi, P.: Notifications of work-related injuries and diseases: an observational study on a mining disaster. BMC Health Serv. Res. **23**(1), 936 (2023)
16. Gil, S., Pelon, G.: Development of short-term prediction with regard to a number of accidents at work using the scoring method. Acta Montan Slovaca. **28**(1), 38–46 (2023)
17. Bilim, N., Bilim, A.: Estimation of the risk of work-related accidents for underground hard coal mine workers by logistic regression. Int. J. Occup. Saf. Ergon. **28**(4), 2362–2369 (2022)
18. Eravcı, D.B., Yılmaz, İ.: Statistical analysis and evaluation on the economic cost of permanent disability due to occupational accidents and occupational diseases in Turkey. Anadolu Univ. J. Fac. Econ. Adm. Sci. **25**(4), 51–71 (2024)
19. Małysa, T., Gajdzik, B.: Predictive models of accidents at work in the steel sector as a framework for sustainable safety. Energies. **14**(1), 129 (2020)
20. Tanaya, A.S., Martiana, T., Rahman, Z.F.: Relationship of age, work period, and education level with occupational accidents in the South Borneo coal industry. Indian J. Forensic Med. Toxicol. **14**(4), 992-997 (2020)
21. Jafari, M.J., Barkhordari, A., Eskandari, D., et al.: Relationships between certain individual characteristics and occupational accidents. Int. J. Occup. Saf. Ergon. **25**(1), 61–65 (2019)
22. da Silva, J.M.N., da Silva Braz, M.L., da Silva, J.G., et al.: Evaluation of the performance of neural and non-neural methods to classify the severity of work accidents occurring in the footwear industry complex. Appl. Syst. Innov. **7**(5), 85 (2024)
23. Santurtún, A., Moraes, S.L., Fdez-Arroyabe, P., et al.: Descriptive analysis of occupational accidents in Spain and their relationship with heatwaves. Prev. Med. **175**, 107697 (2023)
24. Mammadov, A., Kazar, G., Koc, K., et al.: Predicting accident outcomes in cross-border pipeline construction projects using machine learning algorithms. Arab. J. Sci. Eng. **48**(10), 13771–13789 (2023)
25. Koc, K., Ekmekcioğlu, Ö., Gurgun, A.P.: Determining susceptible body parts of construction workers due to occupational injuries using inclusive modelling. Saf. Sci. **164**, 106157 (2023)

26. Koc, K., Gurgun, A.P.: Scenario-based automated data preprocessing to predict severity of construction accidents. Autom. Constr. **140**, 104351 (2022)
27. Koc, K., Ekmekcioğlu, Ö., Gurgun, A.P.: Accident prediction in construction using hybrid wavelet-machine learning. Autom. Constr. **133**, 103987 (2022)
28. Goldberg, D.M.: Characterizing accident narratives with word embeddings: improving accuracy, richness, and generalizability. J. Saf. Res. **80**, 441–455 (2022)
29. Koklonis, K., Sarafidis, M., Vastardi, M., et al.: Utilization of machine learning in supporting occupational safety and health decisions in hospital workplace. Eng. Technol. Appl. Sci. Res. **11**(3), 7262–7272 (2021)
30. Sadeghi, H., Mohandes, S.R., Hosseini, M.R., et al.: Developing an ensemble predictive safety risk assessment model: case of Malaysian construction projects. Int. J. Environ. Res. Public. Health. **17**(22), 8395 (2020)
31. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: Classification and regression trees, 1st edn. Chapman and Hall/CRC (1984). https://doi.org/10.1201/9781315139470
32. Breiman, L.: Random forests. Mach. Learn. **45**(5), 32 (2001). https://doi.org/10.1023/A:1010933404324
33. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
34. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. Wiley (2021)
35. Vapnik, V.N.: The nature of statistical learning theory. Springer. (1995)
36. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory. **13**(1), 21–27 (1967)
37. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery. pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
38. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17). Curran Associates Inc. pp. 3149–3157 (2017)
39. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS '18). Curran Associates Inc. pp. 6639–6649 (2018)
40. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, Hoboken, NJ (2013)
41. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
42. Yontar, M., Namli, Ö.H., Yanik, S.: Using machine learning techniques to develop prediction models for detecting unpaid credit card customers. J. Intell. Fuzzy Syst. (2020). https://doi.org/10.3233/JIFS-189080
43. Aksoy, M., Yanik, S., Amasyali, M.F.: A comparative analysis of text representation, classification and clustering methods over real project proposals. Int. J. Intell. Comput. Cybernetics. **16**(3), 595–628 (2023). https://doi.org/10.1108/IJICC-11-2022-0289
44. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control. Wiley, Hoboken, NJ (2008). https://doi.org/10.1002/9781118619193
45. Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J.: K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. **622**, 178–210 (2023)
46. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(1), 53–65 (1987)
47. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (1979). https://doi.org/10.1109/TPAMI.1979.4766909
48. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974)
49. Gewers, F.L., Ferreira, G.R., De Arruda, H.F., Silva, F.N., Comin, C.H., Amancio, D.R., Costa, L.D.F.: Principal component analysis: A natural approach to data exploration. ACM Comput. Surveys. **54**(4), Article70 (2021). https://doi.org/10.1145/3447755
50. Yuan, C., Yang, H.: Research on K-value selection method of K-means clustering algorithm. Multidisciplinary Sci. J. **2**(2), 226–235 (2019). https://doi.org/10.3390/j2020016

Authors and Affiliations

Meltem Aksoy[1] · Aylin Adem[2] · Ibrahim Yilmaz[3] · Metin Dağdeviren[2,4]

✉ Meltem Aksoy
meltem.aksoy@tu-dortmund.de

1 Computer Science, Research Center Trustworthy Data Science and Security of the University Alliance Ruhr, Technical University Dortmund, Dortmund 44227, Germany

2 Department of Industrial Engineering, Gazi University, Ankara 06570, Turkey

3 Department of Industrial Engineering, Ankara Yildirim Beyazit University, Ankara 06010, Turkey

4 Council of Higher Education, Ankara 06539, Turkey