Avrupa Bilim ve Teknoloji Dergisi
Sayı 53, S. 84-96, Ocak 2024

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 53, pp. 84-96, January 2024

Research Article

Evaluation of Risk Factors Causing Occupational Accidents in the Textile Sector Using Data Mining Methods

Büşra Tunçman[1*], Tülin Gündüz[2], Duygu Yılmaz Eroğlu[3]

[1*] Bursa Uludag University, Faculty of Engineering, Department of Industrial Engineering, Bursa, Turkey, (ORCID: 0009-0005-8361-1708), 502006001@ogr.uludag.edu.tr

[2] Bursa Uludag University, Faculty of Engineering, Department of Industrial Engineering, Bursa, Turkey, (ORCID: 0000-0002-7134-3997), tg@uludag.edu.tr

[3] Bursa Uludag University, Faculty of Engineering, Department of Industrial Engineering, Bursa, Turkey, (ORCID: 0000-0002-7730-2707), duygueroglu@uludag.edu.tr

(First received 18 July 2023 and in final form 17 December 2023)

(**DOI:** 10.5281/zenodo.10646656)

ATIF/REFERENCE: Tunçman, B., Gündüz, T. & Yılmaz Eroğlu, D. (2024). Evaluation of Risk Factors Causing Occupational Accidents in the Textile Sector Using Data Mining Methods. *Avrupa Bilim ve Teknoloji Dergisi*, (53), 84-96.

Abstract

This study suggests that data mining methods can be helpful in preventing occupational accidents in the textile industry. Within the scope of the study, 89,963 occupational accident data that occurred in the textile sector between the years 2019-2021 were examined and the number of samples was reduced to 11,710 with the data preprocessing study. In estimating accidental injury types, model selection map was taken as reference and SVM, Extra Trees, Random Forest, Gradient Boosting and XGBoost algorithms were chosen. Models were compared using the macro F-score performance metric. The estimation performance of models has increased with data balancing and parameter optimization methods. XGBoost algorithm performed better than other algorithms with 70% prediction success. The SVM (69%) and Extra Trees (68%) have been among the algorithms that correctly interpreted the data set by reaching high macro F-score values. It has been seen that the features that have the most effect on the forecast result are cause of accident, material agent, sub-sector, and company size, respectively.

Keywords: Textile Sector, Occupational Accidents, Work Safety, Data Mining, Data Balancing and Hyperparameter Optimization.

Tekstil Sektöründe İş Kazalarına Neden Olan Risk Faktörlerinin Veri Madenciliği Yöntemleriyle Değerlendirilmesi

Öz

Bu çalışma, veri madenciliği yöntemlerinin tekstil sektöründe iş kazalarının önlenmesinde yardımcı olabileceğini önermektedir. Çalışma kapsamında, 2019-2021 yılları arasında tekstil sektöründe meydana gelen 89,963 iş kazası verisi incelenmiş ve veri ön işleme çalışması ile örneklem sayısı 11,710'a düşürülmüştür. Kaza sonucu oluşan yaralanma türlerinin tahmin edilmesinde model seçme haritası referans alınarak SVM, Ekstra Ağaçlar, Rastgele Orman, Gradient Boosting ve XGBoost algoritmaları seçilmiştir. Modeller makro F-skor performans metriği kullanılarak karşılaştırılmıştır. Veri dengeleme ve parametre optimizasyonu yöntemleri ile modellerin tahmin performansı artış göstermiştir. XGBoost algoritması %70 tahmin başarısı ile diğer algoritmalara göre daha iyi performans göstermiştir. SVM algoritması (%69) ve Ekstra Ağaçlar (%68) algoritması, yüksek makro F-skor değerlerine ulaşarak veri setini doğru yorumlayan modeller arasında yer almıştır. Tahmin sonucuna en çok etki eden özelliklerin sırasıyla kaza sebebi, kaza anında kullanılan araç/metaryel, alt sektör bilgisi ve firma büyüklüğü olduğu görülmüştür.

Anahtar Kelimeler: Tekstil Sektörü, İş Kazaları, İş Güvenliği, Veri Madenciliği, Veri Dengeleme ve Hiperparametre Optimizasyonu.

[*] Corresponding Author: 502006001@ogr.uludag.edu.tr

1. Introduction

A safe and healthy workplace environment is a fundamental principle and right. The most important aspect to consider in working life is to create a safe working environment that does not jeopardize the health of the employees. In this context, occupational accidents and diseases that result from a dangerous situation or behavior are the biggest problems. Governments, employers, and employees are responsible for knowing and applying their rights and responsibilities defined for a safe and healthy working environment. The basic aim of the OHS (Occupational Health and Safety) rules, which are globally significant, is to take the most appropriate preventive and protective measures to prevent occupational accidents and diseases. The International Labour Organization (ILO), which promotes a human-centered approach to the future of work, defines occupational accidents as incidents resulting in fatal or non-fatal injuries caused by work-related activities or occurrences. Occupational accidents and illnesses have destructive effects on employees, employers, and national economies.

According to recent statistics, approximately 270 million occupational accidents occur globally per year, with 360,000 of these resulting in fatalities. Analyses shared by the ILO show that occupational accidents and illnesses result in a loss of four percent of the global GDP (Santos, 2021). According to the records of the Social Security Institution (SSI), the recorded number of occupational accidents was 74,871 in 2012, reaching 423,551 in 2019. Industry-specific research indicates that in emergency department admissions due to work-related injuries, the metal and machinery industries rank first with a rate of 30.1%, while the textile sector ranks second with a rate of 28.7%. In another study that examined the causes of occupational accidents in a wider scope, the rate of dangerous behavior was calculated as 68.1% and the rate of dangerous conditions as 11.6%. The fact that the rate of dangerous behavior is significantly higher than that of dangerous conditions indicates that accidents can be largely prevented with small improvements (Güllüoğlu and Taçgın, 2018).

The increasing industrialization and technological developments both globally and in our country have led to some problems related to the working environment and conditions. During this process, certain rules and regulations have been established to ensure a safe working environment and conditions (Çakmak, 2019). The first major improvement initiative in Turkey was the creation of a separate regulation on occupational health and safety in 2012 (Recal ve Demirel, 2021). In addition to legal initiatives to prevent increasing occupational accidents, more comprehensive research and analysis are needed in the field of OHS. To make the right decisions to prevent occupational accidents, it is necessary to identify the factors that lead to occupational accidents and present them to decision-makers. One of the most important sources that can be used to determine these factors is work accident data. However, the increasing size of the occupational accident data, which includes various parameters defining accidents, has rendered traditional statistical methods insufficient. Therefore, data mining technique in big data analysis, which combines statistical methods with artificial intelligence techniques to extract valuable information from complex datasets, has become a powerful alternative to traditional statistical methods (Recal, 2022).

The analysis of occupational accident data with large data sizes has been facilitated by the data mining technique. This situation has led to an increasing share of research focusing on the analysis of occupational accidents in the literature. Studies analyzing occupational accidents using data mining techniques have mostly focused on national or international occupational accident data. In sector-specific studies, work accidents in heavy industries such as mining, petrochemical, and metal sectors have been the focus of research. As an example of sector-specific studies, Cheng et al. (2022) analyzed work accidents in the petrochemical industry in Taiwan for the analysis of work accidents. Classification and regression tree (CART) methods from data mining techniques were used to examine the distribution and rules of accident factors. In some studies in the literature, occupational accidents occurring in different sectors are analyzed at the same time. Gul et al. (2016) analyzed occupational accidents occurring in different sectors in Turkey for two purposes. In the first stage, the characteristics of the employees who had work accidents and accident information were analyzed with clustering algorithms, and the management style that would maximize the efficiency of the investments made to the employees was investigated. In the second stage, hypotheses assuming whether there was any relationship between the characteristics of the employees and occupational accidents were evaluated in binary confusion matrices. WEKA (data mining program) was used in the study.

In his study examining occupational accidents on commercial ships, Çakır (2019) used data mining methods to identify factors affecting injury severity. In the study, the association rule method was used to examine the relationships between accident features and to create accident patterns. The association rule method has been frequently observed in recent studies on occupational accident analysis. Another example of such studies is developed by Çakmak (2019). In the study, accident data recorded by the Social Security Institution (SGK) for the years 2016 and 2017 were used. The study has shown that the association rules method commonly used in market research can also be used in research on the analysis of occupational accidents. Ayanoğlu and Kurt (2019) analyzed accident data in the metal industry using data mining methods (artificial neural network) and developed an accident prediction model. The output label was selected as accident severity (minor injury, injury, limb loss, and fatal accident). WEKA 3.8.1 data mining program was used for analyzing. In studies analyzing occupational accidents with data mining methods, different output labels expressing the accident outcome were preferred.

In some studies in the literature, accident severity has been evaluated using the duration of lost workdays. As an example of this Recal (2022) developed a two-stage data mining method: determining the factors affecting the severity of accidents using the machine learning method and revealing the accident chains formed by the determined factors using the association rule.

Choi et al. (2020) conducted an analysis of industrial accident data in the construction sector using logistic regression, decision trees, random forest, and AdaBoost techniques. They developed a prediction model for categorizing accident risk categories. All model performances were compared and random forest analysis yielded the most successful result in predicting imbalanced binary classifications (91.98%). Similarly, Koç et al. (2021) used ensemble machine learning methods to forecast the condition of

construction workers after an accident. The XGBoost method based on genetic algorithm showed the best performance. Kakhki et al. (2019) used SVM, Boosted Trees, and Naive Bayes algorithms to classify injury severity data resulting from agricultural accidents. They used the F-score metric for model evaluation and found that support vector machines outperformed all other models. The F-score metric, being the harmonic mean of precision and recall, is one of the useful criteria for evaluating the efficacy of classifier algorithms (Mathews, 2016). Khairuddin et al. (2022) found that, the Random Forest method was more successful than other methods in achieving a higher F-score value.

Studies on occupational accidents are generally based on predicting the outcomes of accidents. Parameter optimization in classification algorithms is crucial for the accuracy of prediction models. When building a machine learning model, it is necessary to adjust the hyperparameters for each specific problem. Hyperparameters are configuration variables that are evaluated during the training stage to obtain optimized average values after various trial processes. Optimizing hyperparameters for machine learning directly affects the performance of models (Yang and Shami, 2020). Sarkar et al. (2019) utilized machine learning-based predictive models with optimized parameters to predict accident outcomes using accident data from an integrated steel plant in India between 2010-2013. Wu et al. (2019) proposed the use of automated search algorithms to overcome the disadvantages of manual parameter selection.

Grid search involves training a machine learning model with all conceivable hyperparameter configurations from the training set, and assessing performance based on a predefined metric in the cross-validation set. Shekar and Dagnew (2019) applied a grid search method for parameter optimization along with a cross-validation method in which samples were divided into k randomly selected folds. The sample sizes of classes in a dataset may not always be balanced. In multi-class classification problems, the examples of a class can significantly outnumber other classes. These datasets are defined as imbalanced datasets. In some cases, the minority class may be crucial. However, most classifiers are trained assuming that the numbers of samples in the majority and minority classes are the same (Sarkar et al., 2019).

Imbalanced datasets negatively affect the performance of classifiers (Bulut, 2016). Studies and statistics show that occupational accident datasets classified based on accident outcomes often have an imbalanced distribution. The high frequency of non-fatal occupational accidents compared to fatal ones causes the imbalance of the workplace accident dataset. Consequently, it becomes challenging for a machine learning model to predict fatal occupational accidents. Using resampling techniques in these types of datasets provides significant improvements in predicting accident outcomes (Koc and Gurgun, 2022). The SMOTE algorithm creates a high-speed sampling technique to address the imbalance in the original training set. Rather than creating a direct replica of the minority class samples, it generates synthetic samples (Fernández et al., 2018).

When examining the studies conducted on occupational accidents analyzed by data mining methods, it is evident that no similar research has been conducted to predict the types of injuries resulting from occupational accidents in the textile sector. This article aims to investigate and analyze the risk factors for controlling work accidents in the textile sector using data mining methods. For this purpose, 89,963 work accident data reported to the Social Security Institution (SSI) by companies in the textile sector between 2019 and 2021 were analyzed using data mining methods. The types of injuries resulting from occupational accidents were predicted by XGBoost (eXtreme Gradient Boosting), SVC (Support Vector Classification), Extra Trees, Gradient Boosting, and RF (Random Forest) methods. All methods were compared in terms of their prediction performances and the most successful method was selected. The remaining part of our study consists of three main sections. The second section includes material and method, the third section includes result and discussion and the last section includes the conclusions and recommendations.

2. Material and Method

2.1. Material

In this section, data mining analysis methods used in the study are explained technically.

2.1.1. Machine Learning Algorithms

Support Vector Machines (SVM)

SVM, initially introduced by Vladimir Vapnik and Alexey Chervonenkis in 1963, is a machine learning method with theoretical roots in statistical learning and aims at structural risk minimization (Li et al., 2012; Sánchez et al., 2011). Initially developed for regression tasks, the method was later used as a powerful classifier (Sarkar et al., 2019). SVM is one of the strongest supervised machine learning algorithms due to its ability to avoid the overfitting problem in classification tasks (Yang, 2015; Baby et al., 2021).

Random Forest

Random Forest was developed by Leo Breiman in 2001. It is a popular supervised machine learning algorithm suitable for both classification and regression tasks. It operates on the principle of ensemble learning, where multiple classifiers are combined to tackle complex problems and enhance model performance.

Extra Trees Classifier

Similar to Random Forest, The Extra Trees algorithm is an ensemble learning method that trains numerous decision trees and aggregates their results to obtain a prediction output. The concept of averaging is used to increase accuracy and control overfitting (Abhishek, 2020). Extra Trees allows for the reduction of model bias by using the entire dataset to train decision trees.

Gradient Boosting Classification

Gradient Boosting Classifier, developed by Friedman and his colleagues (2021), is an ensemble learning algorithm used for regression and classification tasks. Gradient boosting aims to minimize a loss function that represents the predictive performance of a model for a particular set of parameters (Bahad and Saaxena, 2020). The name "gradient boosting" comes from the combination of the gradient descent algorithm and boosting technique.

XGBoost (eXtreme Gradient Boosting)

The XGBoost algorithm, created by Tianqi Chen and Carlos Guestrin in 2016, is an advanced version of the gradient boosting algorithm (Parsa et al., 2020). Each tree in the ensemble learns from the previous trees and influences the subsequent trees, collectively enhancing the overall model performance (Friedman, 2001). The XGBoost method is designed for speed, ease of use, and performance on large datasets. Therefore, it offers high performance compared to Gradient Boosting (Dhaliwal et al., 2018). The development of the XGBoost method over the years is shown in Figure 1 (Samur, 2020).

Figure 1. The development of decision tree-based XGBoost algorithm over the years (Şekil 1. Karar ağacı tabanlı XGBoost algoritmasının yıllar içindeki gelişimi)

2.1.2. Classification Metrics

Classification metrics are used to evaluate the performance of machine learning models. Different evaluation criteria exist that reflect the performance of a classification model. Each machine learning model has different constraints and working procedures. Therefore, choosing an appropriate evaluation metric for the problem is crucial for model setup and optimization. The confusion matrix provides information about the prediction performance of the machine learning model (Çelik et al., 2022). The actual and predicted values used in the computations of performance metrics for classification problems are based on the confusion matrix. The expressions corresponding to the fields in the confusion matrix are shown in Table 1.

Table 1. Confusion Matrix (Tablo 1. Karmaşıklık matrisi)

Actual	Negative	False Positive	True Negative
	Positive	True Positive	False Negative
		Positive	*Negative*
		Predicted	

TN (True Negative) indicates that the result is correctly predicted by predicting a negative class label for the values whose true class label is negative. TP (True Positive) indicates that the result is correctly predicted by predicting a positive class label for the values whose true class label is positive. FP (False Positive) indicates that the result is incorrectly predicted by predicting a positive class label for the values whose true class label is negative. FN (False Negative) indicates that the result is incorrectly predicted by predicting a negative class label for the values whose true class label is positive.

Accuracy indicates the ratio of correct predictions made by the classifier method (Equation 1). Precision shows how many of the samples predicted as positive in the test section are correctly classified (Equation2). The recall is a metric that shows how many of the positive examples we should have predicted as positive (Equation 3). A low recall rate confirms that there are many false positives and few true positives. F-Score provides the harmonic mean of precision and recall values (Equation 4). The formulas for the metrics are given in Table 2. All metrics take a value between 0 and 1. The best value is 1 and the worst value is 0.

Table 2. Performance metrics (Tablo 2. Performans metrikleri)

Accuracy	*Precision*	*Recall*	*F- Score*
$\frac{TP+TN}{TP+TN+FP+FN}$ (1)	$\frac{TP}{TP+FP}$ (2)	$\frac{TP}{TP+FN}$ (3)	$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision}+\text{Recall}}$ (4)

2.1.3. Problem of Imbalanced Datasets and Solution Methodologies

The purpose of classification algorithms is to maximize the accuracy rate of predictions. However, in imbalanced datasets where examples are concentrated in some classes, the prediction performance of classification algorithms tends to decrease. Classification algorithms usually assume that the examples in the training set are evenly distributed among classes. Minority classes, which contain rarely occurring events are generally the most important data that needs to be learned (Weiss, 2004; Fernández et al., 2013). The main goal of algorithms focused on imbalanced datasets is to increase the accuracy rate and reduce error rates (Pir, 2022). Four different methods are generally suggested to deal with the issue of class imbalance in both standard and ensemble learning algorithms:

NearMiss, Random-Undersampling (RUS), Random-Oversampling (ROS), and Synthetic Minority Over-Sampling Technique (SMOTE).

NearMiss is a method that aims to prevent information loss. It is based on the KNN algorithm. The distance between majority class examples and minority class examples is calculated and examples with short distances are preserved based on the specified k value. RUS is a method that involves removing random examples from the majority class so that the count of majority class examples is the same as the number of minority class examples. This method reduces the time required for classification and improves classification accuracy. However, the biggest disadvantage of RUS is the loss of potentially useful data for the induction process. ROS is an approach designed against class imbalance by randomly replicating minority class examples. The disadvantage of this method is that it can lead to overfitting and increase classi fication time due to the creation of exact copies of existing examples (Fernández et al., 2013; Pir, 2022). SMOTE is an oversampling method that generates synthetic minority examples. It is one of the most commonly used methods for addressing the problem of imbalanced datasets in data mining projects. The main idea of this method is to select random neighbors from the nearest k neighbors of the minority class examples and create synthetic samples along the line segments that connect any one/all of the selected neighbors depending on the required amount of oversampling. This method prevents the problem of overfitting and provides good classification performance.

2.1.4. Hyperparameter Tuning Based On Gridsearch

Hyperparameter optimization is the process of finding the optimal combination of hyperparameters for a machine learning method based on a selected performance metric. Each machine learning algorithm has multiple hyperparameter expressions and different value options on a hyperparameter basis. It can be challenging to find the best combination of hyperparameters by manually testing each one to achieve a successful model. Therefore, various methods have been developed for hyperparameter optimization. GridSearch is one of the most commonly used methods for hyperparameter optimization. The goal of the Grid Search method is to select the hyperparameter combination that maximizes the prediction model's performance. Accuracy values obtained from different test datasets vary. To solve this problem, grid search is often used in conjunction with the k-fold cross-validation method. In this method, the data is partitioned into k groups, where k-1 groups are used for training and the remaining group is used for testing. The algorithm is trained and tested k times. The performance of the model is recorded at each iteration and the average of all performances is calculated at the end (Ranjan et al., 2019).

2.2. Method

In this study, the roadmap shown in Figure 2 was followed. All the steps followed in the data mining process were detailed in this roadmap. In the first stage, occupational accident data that occurred in the textile industry in the last 3 years was requested from the Social Security Institution (SSI) within the scope of the necessary features required for the study.

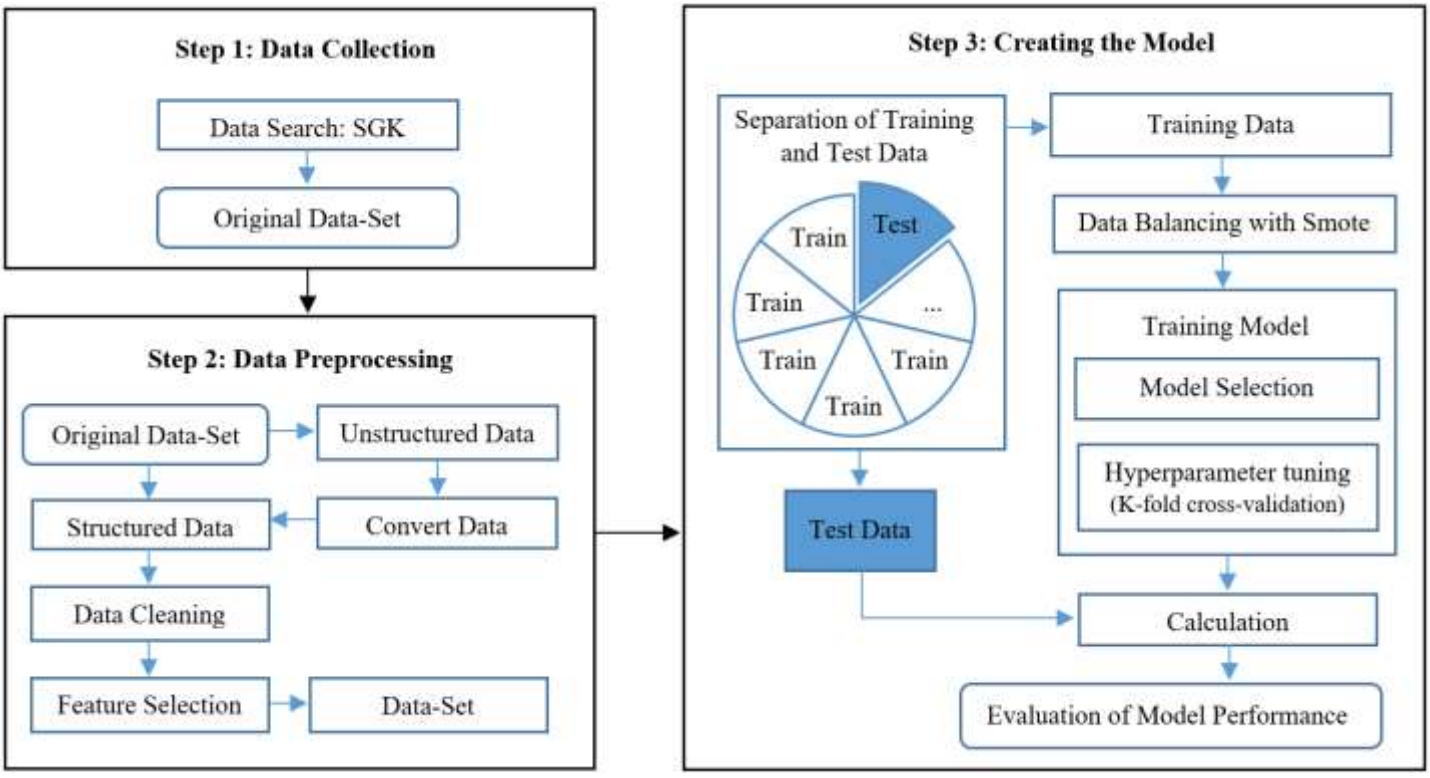


Figure 2. Data mining model selection map (Şekil 2. Veri madenciliği model seçim haritası)

In the second stage, the information in the data set sent by the SGK was examined, and new features that were thought to be meaningful for the study were derived by transforming some of the data. The data set obtained by adding new features was subjected to data preprocessing steps. All the features in the data set were statistically analyzed. Outliers in numerical features were identified

and removed. In categorical features, expressions with insufficient sampling have been cleaned. After the data cleaning process, a feature selection method based on statistics was used. The relationship between each input feature and the output label was evaluated and the input features with the strongest relationship with the output label were selected.

After completing the data preprocessing stage, the model creation stage was initiated. In the model creation stage, the dataset was partitioned into two parts: training and testing data. Different data balancing methods were tried to optimize the prediction performance and reduce the error rate of the model despite the imbalanced distribution of the dataset. The most successful result was achieved with the Smote method. Data balancing techniques should only be applied to the training set. If applied to the test set, the evaluation result will be misleading. Suitable models were identified for the balanced training dataset. The most appropriate hyperparameter combinations were created using the grid search function according to the classification metric for the selected machine learning models. At this stage, 5-fold cross-validation process was utilized to prevent the possibility of different accuracy values obtained by evaluating different test datasets. The machine learning models with optimized parameters using the Grid Search method were trained with the training dataset. After the training process was completed, the models were tested using the test dataset. Based on the test results; accuracy, precision, recall, and F-score values were calculated and model performances were compared.

2.2.1. Data-Set

In this study, the dataset used consists of work accident data reported to the Social Security Institution (SSI) by companies in the textile sector. The dataset sent by the SGK includes 20 different features. The feature description, feature type, and quantity of categorical features are shown in Table 3. Within the context of the study, 89,963 occupational accident data in the textile sector between 2019-2021 were analyzed. The type of Injury, affected parts of the body, and the result of accident are data that can be recorded after the accident. Therefore, these output data are label options. When the studies have been examined, it can be observed that the imbalanced dataset problem significantly negatively affects the model's success when the result of accident data is evaluated as the label. Therefore, the type of injury data has been selected as the label.

Table 3. Informations of dataset features (Tablo 3. Veri seti değişken bilgileri)

Feature	**Description**	**Feature Type**	**Quantity of Category**
Identification Number	Identification number of the employee who had an occupational accident (closed expression).	categorical	89,963
Gender	Gender of the employee who had an occupational accident	categorical	2
Age	Age of the employee who had an occupational accident	numerical	-
Marital Status	Marital status of the employee who had an occupational accident	categorical	7
Sub-Sector	The sub-sector in which the employee who had an occupational accident worked.	categorical	20
City	The city where the employee who had an occupational accident lived.	categorical	77
Work Experience	Work experience(day) in the company where the employee had a work accident.	numerical	-
Business Registration Number	Business registration number of the employee who had an occupational accident (closed expression).	categorical	89,963
Educational Status	Educational status of the employee who had an occupational accident.	categorical	11
Job Title	Job title of the employee who had a work accident.	categorical	1242
Accident Time	The hour at which the accident occurred.	categorical	24
Accident Year	The accident's occurrence year.	categorical	3
Accident Month	The accident's occurrence month.	categorical	12
Arrival Time at Work	The employee's arrival time on the day of the accident.	categorical	24
Company Size	The number of employees at the workplace where the accident occurred.	numerical	-
Cause of Accident	The situation/activity that caused the employee to have an occupational accident.	categorical	9
Material Agent	The tool, object, or instrument being used by the victim when the accident happened, just before the accident.	categorical	22
Affected Parts of The Body	The parts where the employee's body is affected as a result of an occupational accident.	categorical	9
Type of Injury	The type of injury as a result of an occupational accident.	categorical	14
Result of The Accident	The status of the employee as a result of an occupational accident.	categorical	5

2.2.2. Preprocessing

Upon examination of the dataset provided by the Social Security Institution (SSI), it was observed that the Identification Number and Business Registration Number features contain specific information unique to each occupational accident (All personal information has been provided to us by the SSI in closed expressions by the Personal Data Protection Law). The first step in the data preprocessing phase was to clean these expressions. In addition, a new feature has been derived defined as pre-pandemic/post-pandemic, was derived using the data of the month and year of the accident. Before March 2020 is defined as pre-pandemic and after March 2020 is defined as post-pandemic. The effect of this feature on the accident outcome was tested during the modeling phase. After adding and removing features in the dataset, data preprocessing steps were carried out. All features in the dataset were examined statistically. In the categorical features, expressions with a percentage of sample size equal to zero were deleted to clean the expressions without sufficient sampling.

The type of injury label contains a total of 14 different expressions. In Table 4, the quantity and rate of each label in the dataset are expressed. 'Other specified injuries' and 'Unknown or unspecified type of wound' were excluded from the analysis as they do not have a clear definition. To avoid the rare types of injuries resulting from occupational accidents from affecting the prediction performance of the machine learning model, injury type labels with a sample size of less than one per thousand were not included in the analysis. These labels are 'multiple injuries', 'concussion and internal injuries', 'loss of part of the body', 'effects of extreme heat, light, and radiation', 'effects of sound, vibration and pressure', 'drowning and shortness of breath', and 'brain concussion and internal injuries'. Through the research and analysis conducted, it has been observed that the injury type labels defined as ‘dislocations, sprains and strains’ and ‘bone fractures’ are the result of similar incidents. Based on this information, examples belonging to two labels are combined and a new label is defined under the title of 'dislocations, sprains, strains, and bone fractures'. The numerical equivalence of the type of injury label with label encoding is as follows: burns and frostbite are represented as 0, wounds and superficial injuries as 1, poisoning and infections as 2, dislocations, sprains, strains, and bone fractures as 3, and shock as 4.

Table 4. Type of injury labels (Tablo 4. Yara türü etiketleri)

Label	Quantity	%
Wounds and superficial injuries	44,707	49,7%
Other specified injuries	19,126	21,3%
Dislocations, sprains, and strains	13,503	15,0%
Unknown or unspecified type of wound	5,649	6,3%
Bone fractures	3,129	3,5%
Burns and frostbite	1,696	1,9%
Poisoning and infections	1,430	1,6%
Shock	178	0,2%
Multiple injuries	130	0,1%
Concussion and internal injuries	125	0,1%
Loss of part of the body	118	0,1%
Effects of extreme heat, light, and radiation	53	0,1%
Effects of sound, vibration, and pressure	39	0,0%
Drowning and shortness of breath	12	0,0%

After the data cleaning process, the relationship between each input feature and the output label has been evaluated using the Filter-Based Feature Selection Method. It has been observed that the input features with the strongest relationship with the output label are cause of the accident, material agent, sub-sector, and company size, respectively.

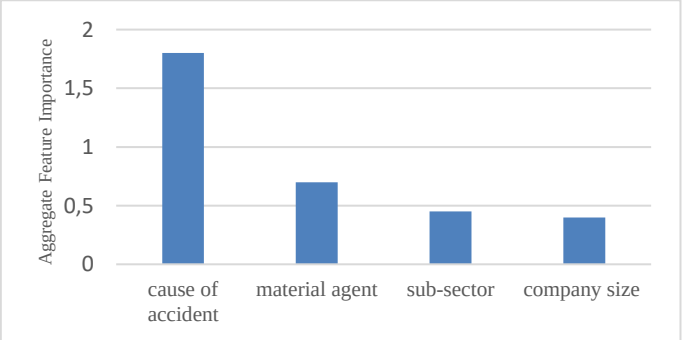


Figure 3. Feature importance graph (Şekil 3. Özellik önem grafiği)

The Covid-19 pandemic that started in 2020, resulted in employees working from home as much as possible and being away from their workplaces. As a result, there was a decrease in the number of days and hours worked in the workplace. The impact of the pandemic on occupational accidents was examined using Phi_K correlation analysis, which works consistently among categorical

features. The correlation matrix was examined separately by eliminating the occupational accidents that occurred during the pandemic period in the occupational accident dataset. Figure 4 shows the analysis results of occupational accident data for the three-year period (2019-2021) that included the pandemic period, while Figure 5 shows only the analysis results of occupational accident data before the pandemic. The binary evaluations in the correlation matrix, the correlation value of 0 indicates that there is no connection between the two features. It has been observed that the correlation values in the correlation matrix including the pandemic process are much lower compared to the correlation values in the correlation matrix without the pandemic process. This shows that the relationship of the label to many features (cause of accident, material agent, company size and gender) are stronger in the occupational accident dataset, in which the pandemic process is not included [Figure 5]. To avoid misleading the analysis of the pandemic process, which has largely lost its effect today, only occupational accidents that occurred before the pandemic were analyzed.

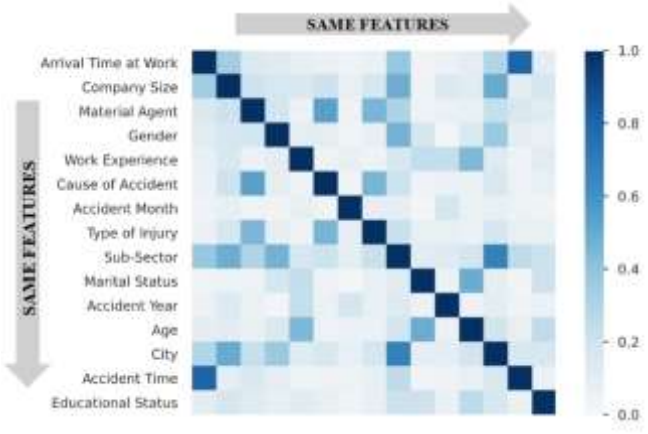


Figure 4. Correlation matrix (2019-2021)
(Şekil 4. Korelasyon matrisi (2019-2021))

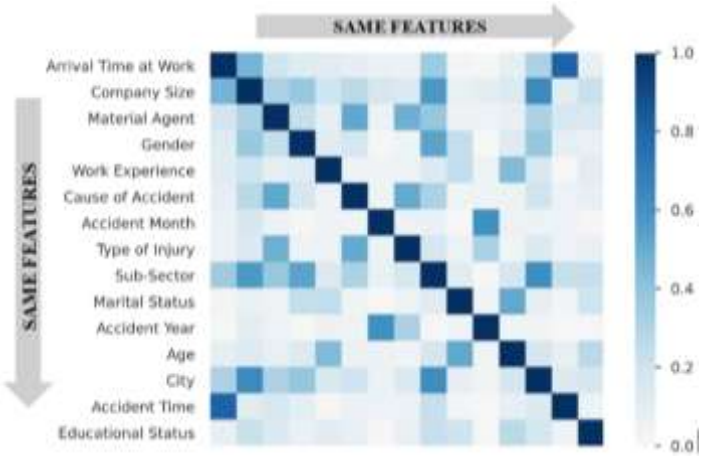


Figure 5. Correlation matrix (2019-2020 March)
(Şekil 5. Korelasyon matrisi (2019-2020 mart))

3. Result and Discussion

In this section, the methods described in the second section were applied step by step to the occupational accident dataset and model performance values were recorded and compared at each step. Data preprocessing steps (feature selection, converting categorical data to numerical data with one-hot encoding, normalization), model creation (data balancing, hyperparameter optimization), and all tests performed to calculate model performance were carried out in Python 3.

During the data cleaning process, the features were initially inspected. In the categorical features, expressions with a percentage of sample size equal to zero were deleted to clean the expressions without sufficient sampling. Next, the output labels were scrutinized. Two output labels that unclear definitions were eliminated from the dataset. Furthermore, injury type labels with a sample size of less than one per thousand were excluded from the analysis to prevent rare injury types from affecting the performance of the machine learning model.

Lastly, occupational accidents that occurred during the pandemic process, whose misleading effect was proven by correlation analysis, were excluded from the dataset. As a result of the analysis and evaluations conducted in the data preprocessing step, the number of samples in the dataset has been decreased to 11,710 to optimize the prediction performance of machine learning algorithms. The category quantities of the features in the dataset after the data preprocessing step is shown in Table 5.

Table 5. Information of features after preprocessing (Tablo 5. Ön işleme sonrası özellik bilgileri)

Feature	**Category**	**Feature**	**Category**
Gender	2	*Accident Time*	24
Age	-	*Accident Year*	2
Marital Status	6	*Accident Month*	12
Sub-Sector	16	*Arrival Time at Work*	24
City	71	*Company Size*	-
Work Experience	-	*Cause of Accident*	8
Educational status	11	*Material Agent*	19
Job Title	567	*Type of Injury*	14

3.1. Application of Algorithms

After completing the data preprocessing step, the dataset was separated randomly into 80% for training and 20% for testing. Next, the MaxAbs Scaler one of the feature scaling methods, was used to normalize the input features of the dataset. In this way, all features were brought to a common scale using their maximum value. Different prediction models are used for different data types and different problems. The model selection map in Figure 6 has been a guide in the selection of the predictive models used in the study.

Considering the dataset and the aim of the study, it was decided to use the SVM, Random Forest, Extra Trees, Gradient Boosting, and XGBoost (eXtreme Gradient Boosting) algorithms.

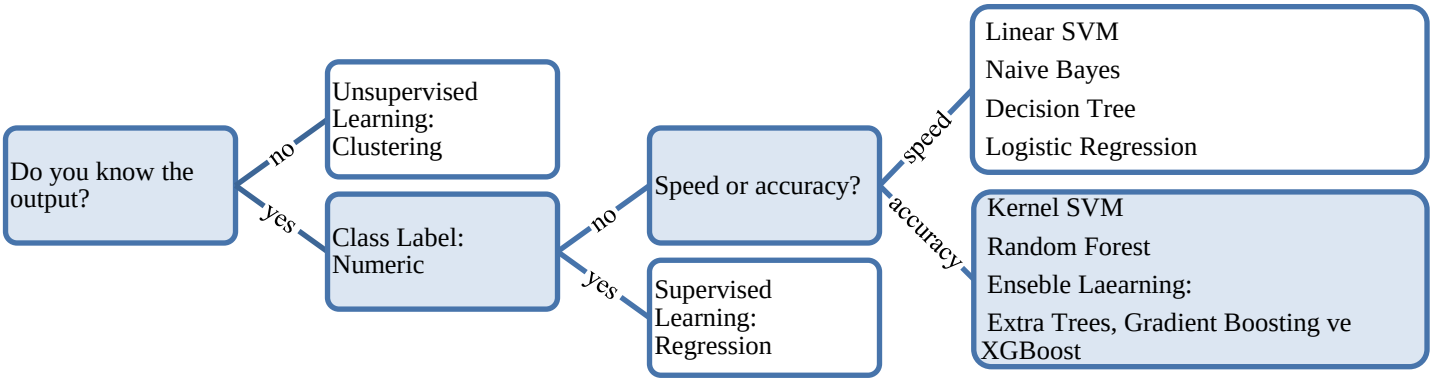


Figure 6. Model selection map (Şekil 6. Model seçim haritası)

3.1.1. Evaluation of Data Balancing Methods

When measuring model performance in imbalanced datasets, low precision and high recall or vice versa success metrics can be obtained. This makes it difficult to compare models. To properly compare the performance of models in this situation, the F-score metric should be used. In the process of building and evaluating the model in this study, the F-score metric has been taken as the basis. However, the high F-score value of the label with a high sample size significantly can raise the weighted F-score value. Therefore, the model performances were compared using the macro F-score, which represents the average of F-score values.

Due to the imbalanced distribution in the dataset, the test sample sizes for some class labels remained very low. In highly imbalanced datasets, accuracy values that favor the majority class can be misleading in terms of the overall success of the model. To overcome this problem, data balancing methods were employed. RUS, ROS, SMOTE and Near Miss methods were applied to the training dataset. The most successful result was achieved with the SMOTE method. Table 6 shows the success values for all scenarios.

Table 6. Evaluation of data balancing methods (Tablo 6. Veri dengeleme yöntemlerinin değerlendirilmesi)

Data Balancing Methods	Metrics	SVM	Extra Trees	Random Forest	Gradient Boosting	XGBoost
Original	accuracy	0.76	0.76	0.77	0.76	0.76
	macro F-score	0.58	0.61	0.62	0.61	0.61
Near Miss	accuracy	0.25	0.34	0.34	0.42	0.37
	macro F-score	0.25	0.24	0.24	0.28	0.31
RUS	accuracy	0.61	0.61	0.61	0.59	0.63
	macro F-score	0.50	0.46	0.45	0.42	0.46
ROS	accuracy	0.76	0.76	0.76	0.75	0.76
	macro F-score	0.57	0.52	0.50	0.53	0.57
SMOTE	accuracy	0.76	0.76	0.76	0.75	0.75
	macro F-score	0.67	0.66	0.64	0.64	0.65

3.1.2. Evaluation of Hyperparameter Tuning Method

Parameters are configuration variables that machine learning algorithms used to predict output labels in the dataset. To achieve successful results from the chosen models, ideal hyperparameter values must be determined. For this purpose, all possible combinations have been tried using the GridSearch method with 5-fold 2-repeat configuration and the parameters that provide the best combination have been applied to the models. The hyperparameter selection modules are designed to optimize the F-score value. Models that completed the data balancing with SMOTE and hyperparameter optimization with GridSearch were tested and confusion matrices that represent the model outputs are shown in Figure 7. Class labels' numerical equivalences are burns and frostbite: 0, wounds and superficial injuries: 1, poisoning and infections: 2, dislocations, sprains, strains, and bone fractures: 3, and shock: 4.

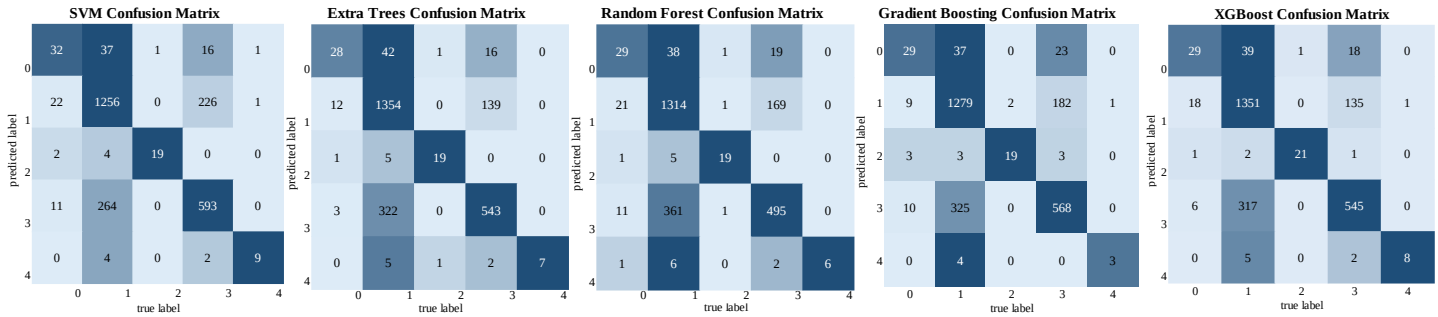


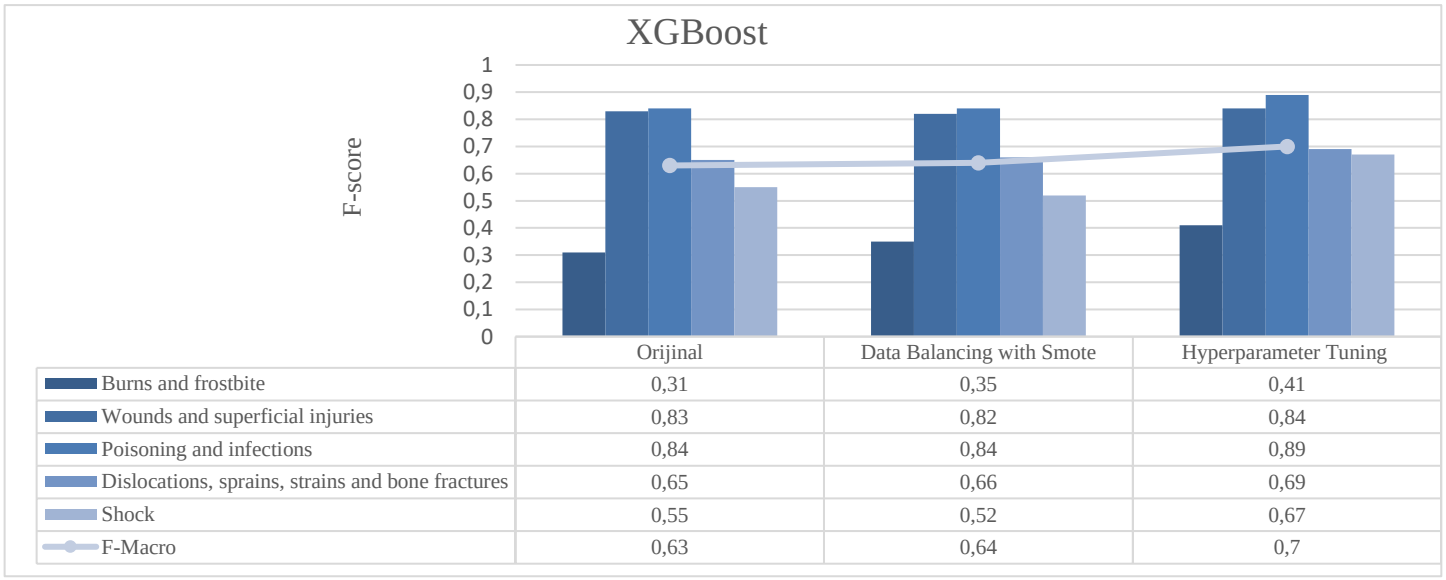
Figure 7. Confusion matrices of models developed with hyperparameter optimization (Şekil 7. Hiperparametre optimizasyonu ile geliştirilen modellerillerin karmaşıklık matrisleri)

The prediction performances of machine learning methods were compared using classification report metrics. The high F-score of the label with a high sample size significantly raised the weighted F-score value. Therefore, the model performances were compared using the macro F-score, represents the average of F-score values. The impact of data balancing and parameter optimization processes on the prediction performance of machine learning methods is shown in Table 7. As a result of the analysis, it was observed that the methods with the highest prediction performance were XGBoost, SVM, Extra Trees, Gradient Boosting, and Random Forest, respectively.

Table 7. Evaluation of Data Balancing and Hyperparameter Optimization Methods (Tablo 7. Hiperparametre ve veri dengeleme yöntemlerinin değerlendirilmesi)

Performance Improvement Steps	Metrics	SVM	Extra Trees	Random Forest	Gradient Boosting	XGBoost
Original	accuracy	0.76	0.76	**0.77**	0.76	0.76
	macro F-score	0.58	0.61	0.62	0.61	0.61
Improvement Step 1: Smote	accuracy	0.76	0.76	0.76	0.75	0.75
	macro F-score	0.67	0.66	0.64	0.64	0.65
Improvement Step 2: Smote + Grid Search	accuracy	0.76	**0.78**	0.75	0.76	**0.78**
	macro F-score	**0.69**	**0.68**	**0.64**	**0.65**	**0.70**

According to the results presented in Table 7, the XGBoost method has the highest prediction performance percentage. Therefore, it is expected that the XGBoost algorithm will provide a more successful basis in accident prediction models. The performance of the XGBoost method, which reached a macro F-score value of 70% through all the performance improvement steps, was also examined in terms of class labels. When data balancing and hyperparameter tuning methods were applied together, improvements were observed in the prediction accuracy of Burns and frostbite label by 10%, wounds and superficial injuries label by 1%, poisoning and infections label by 5%, dislocations, sprains, strains and bone fractures label by 4% and shock label by 12% (Figure 8).



	Orijinal	Data Balancing with Smote	Hyperparameter Tuning
Burns and frostbite	0,31	0,35	0,41
Wounds and superficial injuries	0,83	0,82	0,84
Poisoning and infections	0,84	0,84	0,89
Dislocations, sprains, strains and bone fractures	0,65	0,66	0,69
Shock	0,55	0,52	0,67
F-Macro	0,63	0,64	0,7

Figure 8. Success of the XGBoost model to predict injury types (Şekil 8. XGBoost modelinin yara türlerini tahmin etme başarısı)

4. Conclusions and Recommendations

In this study, risk factors for controlling occupational accidents were analyzed using data mining methods based on accident data from textile sector workplaces in 2019-2021, and the performance of models that would form the basis of accident prediction models was evaluated. In this context, an accident dataset was created from occupational accidents in the textile sector workplaces. The accident dataset was transformed into a more trainable and testable dataset for machine learning by applying data analysis methods. As a result of the analyses, the pandemic process, which was found to have a misleading effect, was excluded from the dataset. Therefore, the scope of the study was limited to occupational accidents that occurred between January 2019 and March 2020, inclusive. Then, machine learning methods that were most suitable for the dataset and the purpose of the study were selected Through data balancing and hyperparameter optimization, machine learning models have been enabled to interpret the dataset with better performance. As a result of the studies, it was seen that the XGBoost algorithm was the best machine learning algorithm to interpret the dataset. Through experiments with 5-fold cross-validation and 2 repetitions, it was revealed that the XGBoost method performs with a success rate of 70%. SVM (69%) and Extra Trees (68%) algorithms also achieved a good level of performance in interpreting the dataset, close to that of the XGBoost algorithm.

The imbalanced distribution of the type of injury labels has caused a low macro F-score value, which was chosen as the performance metric. Both the preprocessing steps in the dataset and the data balancing and hyperparameter tuning methods used in the prediction model generation have improved the macro F-score value to better levels. The SVM algorithm has increased its macro F-score value from 58% to 69%, the Extra Trees algorithm has increased its macro F-score value from 62% to 68%, the Random Forest algorithm has increased its macro F-score value from 62% to 64%, the Gradient Boosting algorithm has increased its macro F-score value from 61% to 65% and the XGBoost algorithm has increased its macro F-score value from 63% to 70%. The prediction performance values of the XGBoost algorithm for output labels are calculated as follows: 89% for poisoning and infections, 84% for wounds and superficial injuries, 69% for dislocations, sprains, strains, and bone fractures, 67% for shock, and 41% for burns, scalding with hot water and frostbite. When the sample sizes and success values were evaluated together, it was observed that the poisoning and infections label had a high prediction success rate, although it was tested with few samples.

According to studies, it has been observed that the XGBoost algorithm performs better compared to other algorithms for datasets with imbalanced distribution. It has also been determined that data balancing and hyperparameter optimization methods are crucial in terms of prediction success for such data sets. As a result, it is understood that an accident prediction model created by data mining steps can be used to identify significant risks that may occur in the textile industry. With the accident prediction model, injury types resulting from accidents in workplaces operating in the textile industry can be predicted based on basic criteria related to employees and work processes. Predicting injury types will contribute to identifying hazardous situations and creating risk maps for textile sector workplaces. Therefore, with the widespread use of accident prediction models, it is expected that occupational safety and health (OSH) personnel in workplaces will take proactive measures against potential accidents and a decrease in the number and severity of occupational accidents will be seen. In future studies, researchers can enhance the findings of this study by expanding the dataset with universal data to develop a more comprehensive accident prediction model. In addition, future studies can focus on the development of real-time monitoring systems that can detect hazardous situations and alert workers and managers in the textile industry. These systems can use machine learning algorithms to analyze data from the workplace and identify potential risks before they lead to accidents.

5. Acknowledge

This research did not receive any specific grant from funding agencies in the public, commercial or non-profit sectors.

References

Santos A.J.R. (2021, February 11). Learning from work-related accidents [Conference presentation]. Towards safe, healthy and declared work in Ukraine – ILO, Ukraine. https://www.ilo.org/wcmsp5/groups/public/---europe/---ro-geneva/---sro-budapest/documents/genericdocument/wcms_769666.pdf

Güllüoğlu, E. N. & Taçgın, E. (2018). Türkiye Tekstil Sektöründe İstihdam ve İş Kazalarının Analizi. *Tekstil ve Mühendis*, *25*(112), 344-354. https://doi.org/10.7216/1300759920182511208

Çakmak H. (2019). Analysis of current occupational accidents raw data by data mining process (Publication No. 614088) [Master's dissertation, Gazi University]. Ulusal Tez Merkezi.

Recal, F. & Demirel, T. (2021). Predicting accident severity with machine learning. *Journal of Intelligent & Fuzzy Systems*, *40*, 10981–10998. doi:10.3233/JIFS-202099

Recal F. (2022). Creating a decision support framework from national occupational accidents data with machine learning approach (Publication No. 743913) [Doctoral dissertation, Yıldız Teknik University] Ulusal Tez Merkezi.

Cheng C. W., Yao H. Q. & Wu T. C. (2013). Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, *26*(6), 1269-1278. doi: https://doi.org/10.1016/j.jlp.2013.07.002

Gül M., Guneri A.F., Yilmaz F. & Celebi O. (2016). Analysis of the relation between the characteristics of workers and occupational accidents using data mining. *The Turkish Journal of Occupational / Environmental Medicine and Safety*, *1*(4), 102-118.

Çakır E. (2019). Workplace hazards and occupational risks: A research on occupational accidents aboard merchant ships (Publication No. 564634) [Doctoral dissertation, Dokuz Eylül University]

Ayanoğlu C. & Kurt M. (2019). Proposal of an occupational accident forecasting model with data mining methods in metal sector. *Ergonomi, 2*(2), 78–87. doi: https://doi.org/10.33439/ERGONOMI.481861

Choi J., Gu B., Chin S. & Lee J.S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in Construction, 110*, 102974. doi: https://doi.org/10.1016/j.autcon.2019.102974

Koc K., Ekmekcioğlu Ö. & Gurgun A.P. (2021). Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. *Automation in Construction, 131*, 103896. doi: https://doi.org/10.1016/j.autcon.2021.103896

Kakhki FD., Freeman SA. & Mosher G.A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science, 117*, 257-262. doi: https://doi.org/10.1016/j.ssci.2019.04.026

Mathews D.G. (2016). Data Mining and Machine Learning Algorithms for Workers' Compensation Early Severity Prediction. [Master's dissertation, Middle Tennessee State University]

Khairuddin M.Z.F., Hui P.L., Hasikin K., Razak N.A.A., Lai K.W., Saudi A.S.M. & Ibrahim S.S. (2022). Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. *International Journal of Environmental Research and Public Health, 19*(21), 13962. doi: https://doi.org/10.3390/ijerph192113962

Yang L. & Shami A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295-316. doi: https://doi.org/10.1016/j.neucom.2020.07.061

Sarkar S., Vinay S., Raj R., Maiti J. & Mitra P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research, 106*, 210-224. doi: https://doi.org/10.1016/j.cor.2018.02.021

Wu J., Chen X.Y., Zhang H., Xiong L.D., Lei H. & Deng S.H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology, 17*(1), 26-40. https://doi.org/10.11989/JEST.1674-862X.80904120

Shekar B.H. & Dagnew G. (2019). Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data [Conference presentation]. Second International Conference on Advanced Computational and Communication Paradigms (1-8), Gangtok, Hindistan. doi:10.1109/ICACCP.2019.8882943

Sarkar S., Khatedi N., Pramanik A. & Maiti J. (2019). *An Ensemble Learning-Based Undersampling Technique for Handling Class-Imbalance Problem.* Lecture Notes in Electrical Engineering, Springer.

Bulut F. (2016). Performance Analysis of Ensemble Methods on Imbalanced Datasets. *Bilişim Teknolojileri Dergisi, 9*(2), 153-159. doi: 10.17671/btd.81137

Koc K. & Gurgun A.P. (2022). Scenario-based automated data preprocessing to predict severity of construction accidents. *Automation in Construction, 140*, 104351. doi: https://doi.org/10.1016/J.AUTCON.2022.104351

Fernández A., García S., Herrera F. & Chawla N.V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research, 61*, 863-905. doi: https://doi.org/10.1613/jair.1.11192

Li Z., Liu P., Wang W. & Xu C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention, 45*, 478-486. doi: https://doi.org/10.1016/j.aap.2011.08.016

Sánchez A.S., Fernández P.R., Lasheras F.S., Juez F.J. & Nieto P.G. (2011). Prediction of work-related accidents according to working conditions using support vector machines. *Applied Mathematics and Computation, 218*(7), 3539-3552. doi: https://doi.org/10.1016/j.amc.2011.08.100

Yang Y., Li J., & Yang Y. (2015). The research of the fast SVM classifier method. In 2015 12th international computer conference on wavelet active media technology and information processing (ICCWAMTIP), Chengdu, China, 121-124.

Baby D., Devaraj S. J., & Hemanth J. (2021). Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach. *Turkish Journal of Electrical Engineering and Computer Sciences, 29*(8), 2742-2757. doi: 10.3906/elk-2104-183

Abhishek L. (2020). Optical character recognition using ensemble of SVM, MLP and extra trees classifier. In 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 1-4.

Bahad P. & Saxena P. (2020). Study of adaboost and gradient boosting algorithms for predictive analytics. In International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019, Singapore, 235-244.

Parsa A. B., Movahedi A., Taghipour H., Derrible S. & Mohammadian A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention, 136*, 105405. doi: https://doi.org/10.1016/j.aap.2019.105405

Friedman J.H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232. doi: 10.1214/aos/1013203451

Dhaliwal S.S., Nahid A.A. & Abbas R. (2018). Effective intrusion detection system using XGBoost. *Information, 9*(7), 149. doi: https://doi.org/10.3390/info9070149

Samur O. (2020). Boosting Algorithms. Retrieved from https://www.datasciencearth.com/boosting-algoritmalari/

Çelik H.İ., Dülger L.C., Öztaş B., Kertmen M. & Gültekin E. (2022). A Novel Industrial Application of CNN Approach: Real Time Fabric Inspection and Defect Classification on Circular Knitting Machine. *Textile and Apparel, 32*(4), 344-352. doi: https://doi.org/10.32710/tekstilvekonfeksiyon.1017016

G.M. Weiss. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter, 6* (1), 7-19. doi: https://doi.org/10.1145/1007730.1007734

Fernández A., López V., Galar M., Del Jesus M. J. & Herrera F. (2013). Analyzing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems, 42*, 97-110. doi: https://doi.org/10.1016/j.knosys.2013.01.018

Pir M. Ş. (2022). Dengesiz Veri Setlerinde Sınıflandırma Problemlerinin Çözümünde Melez Yöntem Uygulaması. (Publication No. 720846) [Doctoral dissertation, Bursa Uludağ University]

Ranjan G. S. K., Kumar Verma A. and Radhika S. (2019). K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). Pune, India.