Impact of Organizational Factors on Accident Prediction in the Retail Sector

Inês Sena[1,3][0000−0003−4995−4799], João Mendes[1,3][0000−0003−0979−8314], Florbela P. Fernandes[1,2][0000−0001−9542−4460], Maria F. Pacheco[1,2][0000−0001−7915−0391], Clara B. Vaz[1,2][0000−0001−9862−6068], José Lima[1,2][0000−0001−7902−1207], Ana Cristina Braga[3][0000−0002−1991−9418], Paulo Novais[3][0000−0002−3549−0754], and Ana I. Pereira[1,2,3][0000−0003−3803−2043]

[1] Research Center in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
[2] Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC),Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
{ines.sena, joao.cmendes,fflor, pacheco, clvaz, jllima, apereira}@ipb.pt
[3] ALGORITMI Research Centre, LASI, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
acb@dps.uminho.pt, pjon@di.uminho.pt

Abstract. Although different actions to prevent accidents at work have been implemented in companies, the number of accidents at work continues to be a problem for companies and society. In this way, companies have explored alternative solutions that have improved other business factors, such as predictive analysis, an approach that is relatively new when applied to occupational safety. Nevertheless, most reviewed studies focus on the accident dataset, i.e., the casualty's characteristics, the accidents' details, and the resulting consequences. This study aims to predict the occurrence of accidents in the following month through different classification algorithms of Machine Learning, namely, Decision Tree, Random Forest, Gradient Boost Model, K-nearest Neighbor, and Naive Bayes, using only organizational information, such as demographic data, absenteeism rates, action plans, and preventive safety actions. Several forecasting models were developed to achieve the best performance and accuracy of the models, based on algorithms with and without the original datasets, balanced for the minority class and balanced considering the majority class. It was concluded that only with some organizational information about the company can it predict the occurrence of accidents in the month ahead.

Keywords: Predictive analytics · Occupational accidents · Preprocessing techniques · Machine Learning algorithms.

1 Introduction

Over the years, companies have invested and implemented several preventive measures to improve security and working conditions, such as safety training, updating tools and machines, safety equipment, strengthening of Safety and Health at Work (OSH) teams, awareness actions, a more significant number of audits, collection of non-conformities, among others [6].

These actions have achieved good results in reducing accidents at the workplace: for example, in 2020, Portugal recorded the lowest number of workplace accidents compared to the previous ten years [1]. Despite the growing efforts of organizations, there was an increase in the number of fatal accidents compared to 2018, accounting for 131 fatalities from work accidents in 2020 [1]. Considering the associated costs, these events continue to be a critical issue for companies, affecting their productivity, competitiveness, and capital. Such occurrences also generate a sense of resentment and anger in society, affecting personal lives and brand image and even causing loss of lives [7, 10].

Therefore, to reduce and prevent the occurrence of accidents, companies began to look for new solutions such as Artificial Intelligence (AI) methods and data analysis since they have achieved good results when applied to other business fields: increase in productivity, forecasting sales, identification of buying behavior, among others [16].

It was found in the literature that predictive analytics exists in several domains, from clinical analysis to forecasting stock markets; however, it is relatively new when applied in predicting the outcome of occupational safety incidents [13]. There are already several areas, such as industry [14], construction [15, 24], and agribusiness [13], that explore these methods and achieve good results in predicting accidents. However, no studies have been found so far regarding the retail sector, where there is a perception that employees are generally at low risk of accidents at work. However, retail workers are involved in various demanding work activities, thus being exposed to multiple risks and hazards [4].

In Portugal, the wholesale and retail trade activity, vehicle and motorcycle repair is the second economic activity with the highest percentage of workplace-related accidents, 14.62% [1]. Due to this fact and the gap in the bibliographic review, this study aims to identify the occurrence of workplace accidents in a retail company. Specifically, this paper seeks to analyze and process data referring to the company's demographic information, absenteeism rates, preventive safety actions, action plan, and accident history, followed by the application of Machine Learning (ML) algorithms (Decision Tree, Gradient Boost Model, K-Nearest Neighbor, Naive Bayes, and Random Forest) to classify accident or non-accident behavior.

The article is organized into five sections. Section 2 presents a bibliographical review regarding the application of Machine Learning in classifying and predicting work-related accidents. Section 3 introduces the discussion of the methodology, in particular, datasets, pre-processing, theoretical concepts of the used forecasting algorithms, the developed predictive models, and the performance evaluation of the models. Section 4 aims at comparing the forecast results achieved. Finally, Section 5 concludes the study by enumerating possible directions for future research.

2 Related Work

This section aims to present state of art related to applying Machine Learning (ML) approaches in predicting occupational accidents.

Although the use of predictive analysis for the minimization, prevention, and prediction of accidents at work is recent, there are already several studies that claim that with Machine Learning algorithms, it is possible to identify and predict with high precision the occurrence of injuries and accidents at work in various business sectors [22].

During the bibliographical review, the construction industry was the business sector that presented a more significant number of studies on applying ML algorithms for predicting workplace accidents, which is justifiable due to the high frequency of accidents at work in this industry [15]. There are several approaches, one of which is through the application of time series data using the coupling of the Discrete Wavelet Transform (DWT) with different methods of Machine Learning, Support Vector Machine (SVM), Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Network (ANNs) to predict the number of daily work accidents for periods of 1 day (short term), 7 days (medium term) and 30 days (long time), with the wavelet-ANN pair achieving the best performance with high accuracy rates in the short and medium term, 0.9275 and 0.7678 respectively, based on the Nash-Sutcliffe efficiency index [24].

There are also many accidents at work in the steel industry. In the paper [14], the authors aim to predict the outcome of accidents at work using two types of Decision Trees, Classification and Regression Trees (CART) and Automatic Chi-square Interaction Detection (CHAID). To do so, they collected 12 variables (age, working days of the worker, number of past accidents of the worker, number of past calamities in the company, daily wage, number of workers in the company, gender, education level, construction type, cause material, severity level, date) for 2127 accidents, and created five predictions for each method. CART achieved better accuracy, 81.78%, and the most predictive variables are age, cause of the accident, and level of education. Therefore, the authors claim that these methods can be used to predict the outcome of workplace accidents in the steel industry [14].

In addition to using ML techniques to predict accidents at work, several studies use these techniques to predict injuries and the severity of accidents at work [13, 27, 7]. In the agribusiness industry, the authors tested the performance of ML techniques, such as Support Vector Machine (with linear, quadratic, and RBF kernels), Boosted Trees, and Naïve Bayes, in modeling and predicting the severity of occupational accidents using workers' complaints injured. The authors state that through the injured part of the body, the body group, the nature of the injury, the heart of the group, the cause of the injury, the group of reasons, age, and stability of the injured workers, they can classify the severity of the damage with an accuracy rate of 92–98% [13].

Also, in the construction area, the severity of accidents is worrying, as the involved tasks can easily cause victims and property losses; it is, therefore, essential to predict the severity of construction accidents. The authors of this study used 16 critical accident factors, 39 attributes, and eight ML algorithms (Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, K-Nearest Neighbor, Random Forest, Multi-Layer Perceptron, and AutoML) and achieved the best F_{score} of 78.3% with Naive Bayes and logistic regression [27].

Apart from ML algorithms, there are other approaches to be explored: for example, the fuzzy logic of Artificial Intelligence that helps to map inputs and outputs efficiently to build the inference engine so that various types of accidents can be predicted [6].

In summary, most of the studies analyzed rely only on data about the characteristics of the victims, events, causes, and injuries for predicting accidents at work. In this context, this work intends to study whether, with only some organizational data, it is possible to identify the number of accidents and non-accidents.

3 Methodology

This research uses information from a Portuguese retail company and ML classification and/or regression algorithms to identify the number of accidents and non-accidents during January 2023.

The developed datasets, pre-processing techniques, ML algorithms, prediction models, and metrics are applied for the best results and performance.

3.1 Dataset Characterization

For this case study, the intention is to use information that all companies generally store without using it to predict accidents to find out if, with this information, it is possible to identify the occurrence of accidents in the month ahead. For this purpose, two sets of data called ”internal information” and ”safety actions” were prepared, which will be associated individually and simultaneously with the history of accidents and subsequently implemented in different forecasting algorithms.

The set of internal information consists of a combination of demographic data and absenteeism rates for 2022. Demographic data is information about the different characteristics of a population; in this case, it includes information on the number of working hours, average age, percentage the average length of service, percentage of female employees, number of employees, number of full-time and part-time employees, and levels of education per organizational unit among others. Absenteeism at work represents the absence of an employee or more during the working period, whether due to delays, for a few hours, or even missing for several days. This study includes total absenteeism rates for sick leave, covid-19, unjustified absences, parenting, accidents, and other causes. Thus, the set of internal information is composed of 25 input variables.

The set of security actions includes the record of preventive security actions and the action plans established to resolve the identified situations. The descriptions of preventive safety actions are nonconformities observed by those responsible for the unit and Safety and Health at Work (OSH) elements when they visit the field. For each nonconformity, action plans are developed for the intervention of others to repair and improve conditions to prevent workplace incidents. This dataset was created by dividing the information by the status of the action (in progress or concluded) and by its sum per month and the organizational unit, adding up to 50 variables, with information only from August since the company only started collecting these records from that month onwards.

In this way, the data from the internal information set was understood for the same period so that the connection between the groups and the comparison of results was coherent. Therefore, this study will be applied to the internal information dataset, the security actions dataset, and the combined dataset, in which the standard fields, organizational unit, and month will make their integrated.

In addition to these two sets, the accident history was used, including the date of the occurrence and the organizational unit to which the victim belonged, to classify the information from the remaining datasets into non-accident (0) or accident (1), accounting for a total of 22138 data representative of the non-accident class, and 369 referring to accident situations.

3.2 Data Preprocessing

Data preprocessing is a fundamental step when intending to use Machine Learning algorithms for data classification/regression since the quality of the data and the valuable information that can be derived from them directly affect the model's learning capacity.

In this case, when analyzing the datasets in detail, some problems that can influence the learning of the intended model were identified, such as duplicate data, invalid data, and imbalance between the output variables. Therefore, preprocessing techniques were applied to remove duplicates and null values and balance the data before inserting it into the developed model.

Removing duplicate values is critical because such values can distort the analysis and cause incorrect predictions [15]. Likewise, eliminating null values is also essential. Null values can indicate missing information or errors in the dataset. These values can significantly affect the accuracy of predictions since ML algorithms may have difficulty dealing with missing data or may misinterpret it [15].

Furthermore, it is essential to balance the dataset so that each class has the same number of samples and therefore has the same weight in the analysis, avoiding vices [9]. In this case, data on the non-occurrence of accidents are much higher than data on the occurrence of accidents, which may lead to machine learning algorithms favoring this class and leading to high prediction values, but inaccurate predictions and distortions in the analysis. For this reason, it is essential to use the confusion matrices to validate the prediction results.

For this study, we intend to apply two methods that solve the problem of data imbalance, namely the Synthetic Minority Oversampling Technique (SMOTE) and Random Undersampling, and compare the results of classification and/or prediction.

SMOTE's main objective is to increase the number of minority samples by inserting n synthetic minority samples among the k samples closest to a given model with a smaller dimension [27]. Random Undersampling is the contrast, a form of subsampling of the majority class, balancing the data up to the size of the minority class, reducing the sample size and the pressure on storage which will improve the execution time; however, the removal of data can lead to the loss of helpful information [9].

Finally, the categorical variables that specify the unit of work were converted into integer variables since ML algorithms produce better results with data of numerical typology [22].

3.3 Methods

Accident occurrence classification is a typical recent classification problem [13]. This study will evaluate the intended result by comparing five supervised classification algorithms.

Decision Tree (DT) is a supervised learning algorithm that can be used for both classification and regression, which relates decisions and possible consequences [16]. This recursive partitioning technique creates a decision tree with several nodes obtained and divided through division criteria [3]. The tree-building procedure stops when the learning dataset is fitted with predictions. Classification trees are developed for categorical and continuous dependent variables that can assume a finite number of unordered values [20].

Gradient Boost Model (GBM) is an ensemble model that combines several weak predictive models that relate predictors to an outcome. This tree construction method is used to reduce the errors of previous trees, which makes the current model focus on data that previous models failed to predict or classify when fitting an underlying model [3]. It is a reinforcement method in which the dataset is resampled repeatedly, with the results produced as a weighted average of the resampled datasets [16].

K-Nearest Neighbor (KNN) is a non-parametric method used in classification and regression problems, which assumes that similar objects are close to each other [16]. This method examines the k most immediate observations and categorizes a statement. Using the closest point of a group of previously classified matters is used as a basis for categorizing a new topic using the nearest neighbor decision rule [22]. The necessary parameters for the algorithm are the value of k and the distance function; the correct value of k is the value, which after several runs with different values of k, reduces the number of errors found. The distance function is calculated using the Euclidean distance, understood as the physical separation between two-dimensional points [22].

Naive Bayes (NB) is a technique to build classifiers with high bias and low variance that can make a solid model even with a small dataset [8]. It is based on Bayes' theory which uses conditional probabilities to classify a categorical target variable based on the input variables [13]. This algorithm predicts the likelihood of different categories through different attributes [8]. When the response variable has two classes, as in injury severity with non-severe and severe types, NB models typically exhibit excellent accuracy [13].

Random Forest (RF) is a popular Machine Learning technique that uses several independent decision trees created from randomly selected variables [3]. RF models are composed of multiple decision trees, each trained using a portion of the original training data and looking only at a randomly chosen subset of the input variables to find a split. Each tree casts a single vote to define the final categorization further, and the classifier's output is finally chosen by the majority of tree votes [20].

Machine learning algorithms undoubtedly guarantee good results in most of their applications, however, it is necessary to ensure some basic conditions for obtaining results. Within these conditions, it is possible to highlight the dataset, which is the basis of all learning. Getting good results with these methods will be difficult without a sufficiently large and concise dataset. Another condition essential to ensure when applying this type of method is a good combination of hyperparameters; these can guarantee a boost in the results obtained, noting that they must be optimized for each model and each dataset used, with no fixed optimal combination. To carry out this optimization, there are different ways, starting with manual fitting, which consists of trial and error, applying different values, and observing the behavior of the results with them. Evaluating the number of hyperparameters present in most artificial intelligence algorithms is easily concluded that it was necessary to evolve these optimization techniques, with several others emerging, such as grid search [2], random search [5], and later gradient-based algorithms like Bayesian Optimization [26], SMAC [18] as well as metaheuristic algorithms like Genetic Algorithm [21] Particle Swarm Optimization [25].

In this specific case, the random search method was chosen mainly because of its effectiveness combined with a less expensive computational cost when compared to some of the other hyperparameter optimization methods.

Another reason that led us to choose the random search was its ability to cover a broader range of values compared to the grid search since it is a method that does not require prior knowledge of the hyperparameter space, which can be quite complex for some models. It randomly samples hyperparameters from a distribution, which becomes useful when working with more than one algorithm, which is the case when comparing five different algorithms; it is notoriously difficult to know all their hyperparameters.

3.4 Model Development

To develop the model, combining all the previously collected information is necessary. Focusing on identifying the number of accidents and non-accidents in January 2023, it is necessary to train the different models with data from the year 2022. The proposed methodology will be applied to the internal information dataset to ensure the security of the action and both of them together.

Starting with the division of the dataset by months, four months were used for training - 15078 events, approximately 80% of the total data - and the remainder for testing (3799 events) corresponds to approximately 20%. Remembering that to predict the next month is needed to ensure that the model is trained with data from the previous month, using the information of the months from August to November as predictors and the accidents or non-accidents information from September to December as outputs to train the five models. Likewise, for testing the models, the same strategy was used. All the information from December was used as predictors and as output, the predicted data will be the accident or non-accident information from January.

To identify cases of accidents and non-accidents, the value of 0 was associated with each existing information of each dataset when there was no accident, and the value of 1 when there were accidents, accounting for a total of 22138 non-accidents and 369 accident occurrence data (September to January). This connection was made taking into account the previous month, that is, an accident that occurred in September will be connected to August information from the remaining datasets.

Observing the accident and non-accident values, it is notorious that the data is not balanced. Due to this data imbalance, pre-processing techniques were applied to balance the data as mentioned in topic 3.2. The proposed methodology was applied to datasets without balancing, with data balancing through the SMOTE method, and with data balancing from Random Undersampling to compare results. A similar process was used to optimize the algorithms, using or without it to compare results.

To simplify the methodology and all the referred information, a flowchart represented in Fig. 1 was created.

In summary, five classification algorithms will be used for three datasets, internal information, security actions, and a total set, which joins the two previous pieces of information. Each dataset will be applied, with the original data, with the training data balanced for the majority class, and with the data balanced considering the minority class, to the algorithms with and without optimization of the hyperparameters. Noting that the test set did not undergo any type of balancing, maintaining the original results with 62 occurrences of accidents and 3737 occurrences without accidents. This is to conclude better the results of using this information to predict the occurrence of accidents in the following month.

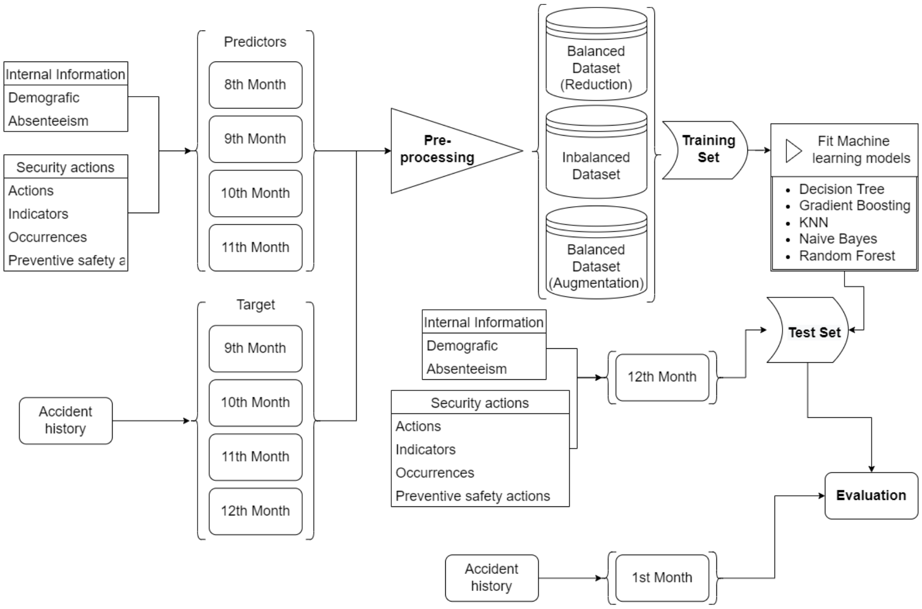


Fig. 1. Flowchart representative of the methodology used in this study.

3.5 Performance Evaluation

To evaluate the performance of the classification model under study, metrics by class were used, such as $Accuracy$, $Precision$, $Recall$, and F_{score}. The metrics are based on the confusion matrix generated for each algorithm [20], which indicates:

- True Positives (TP) – data that were not accidents and were predicted correctly.
- True Negatives (TN) – data points to the model correctly projected as an accident.
- False Positives (FP) – data that the model projected as non-accidents and actually represented the occurrence of accidents.
- False Negatives (FN) – results that were non-accidents and the model identified as an accident.

$Accuracy$ is the most common metric to be used in problems of this magnitude and represents the number of correct predictions as a function of all predictions made, Equation 1 [20]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

The $Precision$, defined in Equation 2, is calculated to evaluate the total correct predictions for a specific class [20]:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The *Recall* measures the number of true positives that were classified correctly through Equation 3 [20]:

$$Recall = \frac{TP}{TP + TN} \tag{3}$$

The F_{score}, defined in Equation 4, is the harmonic mean of *Precision* and *Recall*, which reaches its best value at one and its worst at zero [20]:

$$F_{score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

The confusion matrix for the specific case study is based on the occurrence of accidents, as shown in Table 1.

Table 1. Confusion matrix example based [20].

	Predict Label	
True label	Not an accident (0)	A ccident (1)
Not an accident (0)	TP	FN
Accident (1)	FP	TN

4 Results

In this section, the obtained results are presented and analyzed for each data set, internal information, security actions, and combining the two.

To obtain the best results by the ML algorithms, the hyperparameters of the five algorithms used were optimized with the random search method, as demonstrated in Section 3.3, with equal parameters in terms of evaluation, using 5-fold cross-validation, the random state was used, and a number of 100 iterations for each of the algorithms, and for each of the datasets used. The hyperparameters studied for each algorithm and the chosen values are presented in Table 2. Here, Randint represents uniform random integer values, and Logspace means log-spaced floating point values, with 100 samples evenly spaced on a logarithmic scale.

Noting that all the algorithms presented in this work were tested, trained, and implemented on a computer equipped with an Intel(R) Core(TM) i7-10875H processor, with a RAM DDR4 32GB memory and Python version 3.8 as well as the libraries scikit-learn in version 1.2.1 [23], Pandas in version 1.3.4 [19], numpy in version 1.19.2 [11], imblearn in version 0.10.1 [17] and finally matplotlib in version 3.5.0 [12].

In this way, several forecast models were developed for each dataset to compare the results taking into account whether the training set is balanced with the SMOTE technique or with Random Undersampling and if the hyperparameters of the algorithms are optimized or not, originating six approaches for the comparison of results:

Table 2. Hyperparameter values obtained for each Machine Learning algorithm used in this study.

	Methods
Hyperparameter	**Decision Tree**
max_depth	Randint(2,10)
splitter	Either ”best” or ”random”
min_samples_split	Randint (2,20)
min_samples_leaf	Randint (1,10)
max_features	Either ”auto”, ”sqrt”, ”log2”, or None
max_leaf_nodes	Randint (1, 10)
criterion	Either ”gini”, ”entropy”, or ”log_loss”
	Random Forest
max_depth	Randint (2,10)
min_samples_split	Randint (2,20)
min_samples_leaf	Randint (1,10)
max_features	Either ”auto”, ”sqrt”, ”log2” or None
criterion	Either ”gini” or ”entropy”
n_estimators/neighbors	Randint (1,200)
	K-Nearest Neighboor
n_estimators/neighbors	Randint (1, 200)
leaf_size	Randint (2, 50)
p	Randint (1, 20)
weights	Either ”uniform”, ”distance” or None
algorithm	Either ”auto”, ”ball_tree”, ”kd_tree” or ”brute”
	Naive Bayes
var_smoothing	Logspace (0, -9, num=100)
	Gradient Boost Model
min_samples_split	Randint (2, 20)
min_samples_leaf	Randint (1, 10)
criterion	Either ”friedman_mse” or ”squared_error”
n_estimators/neighbors	Randint (1, 200)
loss	Either ”exponential”, ”deviance” or ”log_loss”
learning_rate	Logspace(0,-7, num = 100)

– Default - In this approach, the original training set is used in the algorithms without optimizing the hyperparameters.

– Optimized - In this case, the original training set is applied in the algorithms to optimize the hyperparameters.

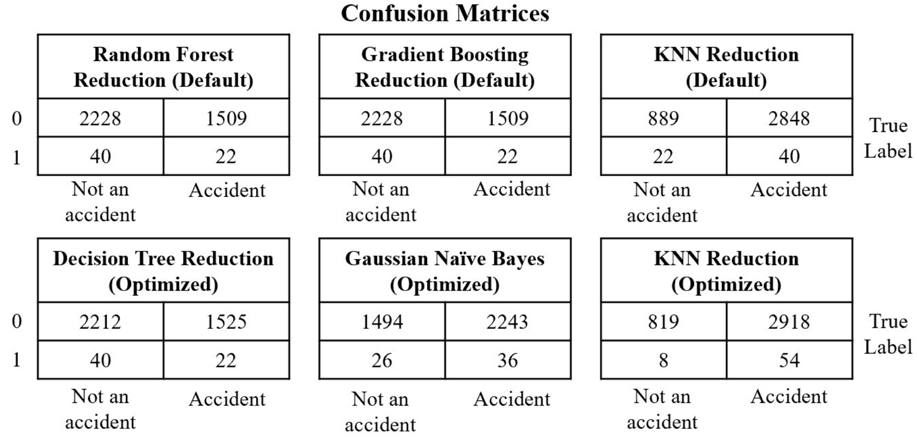
– Reduction Default (RD) - In this strategy, the training set is balanced according to the minority class and implemented in the different algorithms without optimizing the hyperparameters.

– Reduction Optimized (RO) - In this approach, the training set is balanced according to the minority class and implemented in the different algorithms with hyperparameter optimization.

- Augmentation Default (AD) - In this case, the training dataset is balanced with the datasets technique, and in turn, it is applied to the five algorithms without optimizing the hyperparameters.
- Augmentation Optimized (AO) - In this last comparison strategy, the training dataset is balanced with SMOTE technique, and in turn, it is applied to the algorithms taking into account the optimization of the hyperparameters.

4.1 Internal Information Dataset

After obtaining all the results, it is possible to make some observations about using this dataset for accident prediction. Fig. 2 shows the confusion matrices that presented the best results.



Confusion Matrices

Random Forest Reduction (Default)

True Label	Not an accident	Accident
0	2228	1509
1	40	22

Gradient Boosting Reduction (Default)

True Label	Not an accident	Accident
0	2228	1509
1	40	22

KNN Reduction (Default)

True Label	Not an accident	Accident
0	889	2848
1	22	40

Decision Tree Reduction (Optimized)

True Label	Not an accident	Accident
0	2212	1525
1	40	22

Gaussian Naïve Bayes (Optimized)

True Label	Not an accident	Accident
0	1494	2243
1	26	36

KNN Reduction (Optimized)

True Label	Not an accident	Accident
0	819	2918
1	8	54

Fig. 2. Confusion matrices of the approaches and method that presented better results for the internal information set.

Analyzing Fig. 2, it appears that the approach that presents the best results in the different methods is the Reduction Default, except the Naive Bayes method, which is the Optimized approach, and KNN where the Reduction Optimized (RO) approach also demonstrates promising results. Observing all matrices, these approaches reached higher true negative and lower actual positive values than the others.

The Random Forest, Gradient Boost Model, and Decision Tree methods, on the other hand, present very similar confusion matrices, so it is necessary to resort to evaluation metrics so that it is possible to draw a conclusion on which approach and method best predicts the occurrence of accidents with this dataset. The metrics obtained for each mode can be seen in Table 3.

According to the metrics achieved for each strategy and method, it is possible to mention that with the dataset of internal information of a retail company, it

Table 3. Metric values obtained for the methods and approaches that presented better results in the confusion matrices.

		Metrics			
Method/Strategy	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
RF with RD	Not an accident	0.98	0.70	0.82	0.70
RF with RD	Accident	0.02	0.31	0.03	
DT with RD	Not an accident	0.98	0.59	0.74	0.59
DT with RD	Accident	0.01	0.35	0.03	
GBM with RD	Not an accident	0.98	0.60	0.74	0.59
GBM with RD	Accident	0.01	0.35	0.03	

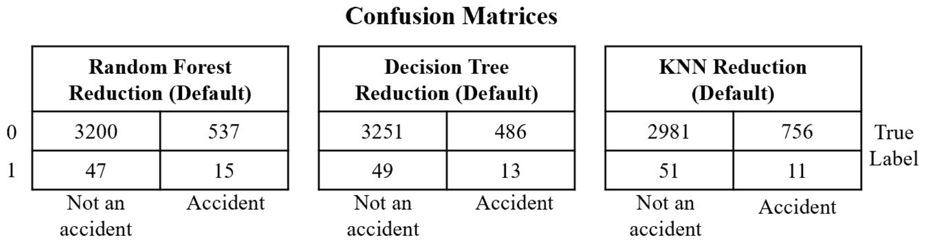
is possible to predict with 70% accuracy the occurrence of accidents in the next month through the Random Forest algorithm in the approach Reduction Default.

4.2 Security Actions Dataset

For this dataset, Naive Bayes obtained the best results in predicting the accident event, ranging from 84 to 100% accuracy, however, it was less assertive than the other algorithms in identifying non-accidents, reaching between 2 to 15% accuracy. Considering the final accuracy, the maximum reached was 16% in the Default approach.

The Decision Tree, Random Forest, and Gradient Boost Model methods presented similar results for the Reduction Default and Reduction Optimized approaches. However, when approaching the optimized algorithm, the value of false negatives increases considerably, although there is an increase in true negatives. Of these three methods, the GBM is the one with the least assertiveness, reaching 10% in Reduction Default and 11% in Reduction Optimized, because it has an acceptable value of true negatives (45 accidents predicted out of 62) but a reduced number of true positives (330 in 3737 no accidents). Also, the K-Nearest Neighbor algorithm performs well in Reduction Optimized.

In Fig. 3, one can visualize the confusion matrices of the approaches that obtained the best results for the set of security actions.



Confusion Matrices

	Random Forest Reduction (Default)		Decision Tree Reduction (Default)		KNN Reduction (Default)		
0	3200	537	3251	486	2981	756	True Label
1	47	15	49	13	51	11	
	Not an accident	Accident	Not an accident	Accident	Not an accident	Accident	

Fig. 3. Confusion matrices of the approaches and method that demonstrated better results for the security actions dataset.

Since the results presented in Fig. 3 are very similar, the values of the metrics that can be seen in Table 4 were used.

Table 4. Metric values obtained for the methods and approaches that presented better results in the confusion matrices.

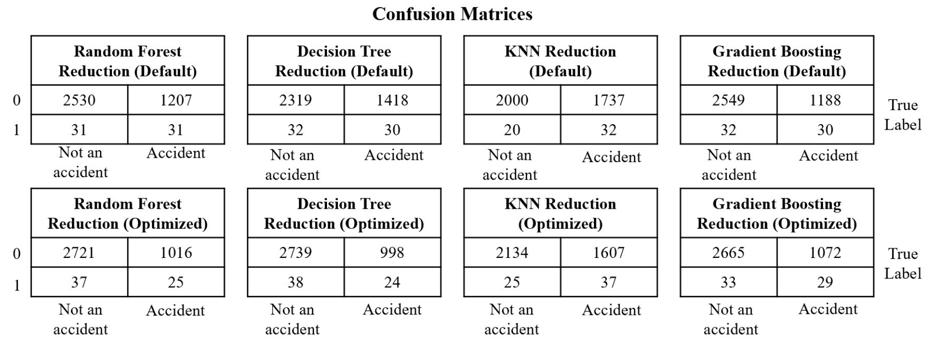
		Metrics			
Method/Strategy	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
RF with RD	Not an accident	0.95	0.86	0.92	0.85
RF with RD	Accident	0.03	0.24	0.05	
DT with RD	Not an accident	0.99	0.87	0.92	0.86
DT with RD	Accident	0.01	0.21	0.05	
KNN with RO	Not an accident	0.98	0.80	0.88	0.79
KNN with RO	Accident	0.01	0.18	0.03	

Analyzing the presented results in detail, the most appropriate model to achieve the intended objective is the Random Forest in the Reduction Default approach.

4.3 Total Dataset

After obtaining results individually, the two data sets were merged through common factors, thus obtaining the total set. The same methodology was applied to see if they achieved better precision in predicting accidents together or separately.

Therefore, from the 30 calculated confusion matrices, Fig. 4 will show those that present better results considering the listed objective.



Confusion Matrices

	Random Forest Reduction (Default)	
0	2530	1207
1	31	31
	Not an accident	Accident

	Decision Tree Reduction (Default)	
0	2319	1418
1	32	30
	Not an accident	Accident

	KNN Reduction (Default)	
0	2000	1737
1	20	32
	Not an accident	Accident

	Gradient Boosting Reduction (Default)	
0	2549	1188
1	32	30
	Not an accident	Accident

True Label

	Random Forest Reduction (Optimized)	
0	2721	1016
1	37	25
	Not an accident	Accident

	Decision Tree Reduction (Optimized)	
0	2739	998
1	38	24
	Not an accident	Accident

	KNN Reduction (Optimized)	
0	2134	1607
1	25	37
	Not an accident	Accident

	Gradient Boosting Reduction (Optimized)	
0	2665	1072
1	33	29
	Not an accident	Accident

True Label

Fig. 4. Confusion matrices of the approaches and method that demonstrated better results for the total dataset.

Analyzing Fig. 4, it appears that the Naive Bayes method is not illustrated because comparatively, the confusion matrices presented, the NB shows values of true negatives and false positives lower than the other methods, which indicates high accuracy values but with lower assertiveness in the occurrence of accidents.

It can also be drawn that there is a considerable amount of true negatives and true positives compared to the results observed in the individual datasets. However, it is necessary to use the metrics so that it is possible to draw a more assertive conclusion about the accuracy, approach, and method that makes the dataset reaches the objective. The value of the metrics can be seen in Table 5.

Table 5. Metric values obtained for the methods and approaches that presented better results in the confusion matrices.

		Metrics			
Method/Strategy	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
RF with RD	Not an accident	0.99	0.68	0.80	0.67
RF with RD	Accident	0.03	0.50	0.05	
RF with RO	Not an accident	0.99	0.73	0.84	0.72
RF with RO	Accident	0.02	0.40	0.05	
DT with RD	Not an accident	0.99	0.62	0.76	0.62
DT with RD	Accident	0.02	0.48	0.04	
DT with RO	Not an accident	0.99	0.73	0.84	0.73
DT with RO	Accident	0.02	0.39	0.04	
KNN with RD	Not an accident	0.99	0.54	0.69	0.53
KNN with RD	Accident	0.02	0.52	0.03	
KNN with RO	Not an accident	0.99	0.57	0.72	0.53
KNN with RO	Accident	0.02	0.60	0.04	
GBM with RD	Not an accident	0.99	0.68	0.81	0.68
GBM with RD	Accident	0.02	0.48	0.05	
GBM with RO	Not an accident	0.99	0.71	0.83	0.71
GBM with RO	Accident	0.03	0.47	0.05	

Analyzing Table 5, it can be noted that the methods that obtain a higher accuracy value are the Random Forest, Decision Tree, and the Gradient Boost Model in the Reduction Optimized approach. Observing the results in more detail and taking into account the recall metric, the GBM Reduction Optimized is the one that presents the highest value for the accident class and values for the non-accident class similar to those of the other methods that were mentioned; therefore with the combination of the two datasets it is also possible to predict the occurrence of accidents with an accuracy of 71%.

5 Conclusion and Future Work

The present study aimed to predict the occurrence of accidents in the month ahead of the current one in a retail company, with only organizational data classified as an accident or non-accident event. Taking into account the information provided by the company, two sets of data were developed that were individually applied and combined into five Machine Learning classification algorithms.

Throughout the data analysis, some obstacles were faced, such as the period not being of the same dimension for the different data; therefore, it was necessary to reduce the information to understand the same period. The imbalance of the data was high, with a greater number for the non-accident class, therefore, the results were used and compared with two data balancing approaches. As these mishaps can influence performance and harm the results, hyperparameters optimization was also used for each algorithm to obtain the best possible performance from them.

In this way, six approaches were developed for comparing results and obtaining the best ones considering different techniques to solve the problems encountered. Among the different approaches that were used, the one that achieved the best results in different datasets and methods was the Reduction Default.

It can also be mentioned that the best accuracy achieved was 85% for the safety actions dataset in the RF Reduction Default. However, if we analyze the values of the metrics in detail, the total set reached a higher recall value for the accident class in the GBM method in the Reduction Default approach; however, its accuracy is lower (71%).

In addition, it can also be mentioned that NB is an algorithm that obtains higher values of true negatives; however, it reaches low values of precision due to the values of false positives being higher than those of true positives.

It is possible to conclude that with the information used and without any details of the accidents and characteristics of the victim, it is possible to predict the occurrence of accidents at work in the next month. However, it is necessary to explore this study further to find solutions that increase the value of true negatives and decrease those of false positives.

In this way, and considering the number of workplace accidents and the existing gap in the literature, it is important to deepen the exploration of factors and algorithms that predict the occurrence of accidents at work in any business sector.

As future work, it is intended to explore further the information used, find solutions to the fact that there is a limitation of information regarding the accident event, experiment with other types of organizational information, and add new information to the study.

Acknowledgement

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI (UIDB/05757/2020 and UIDP/05757/2020), ALGORITMI UIDB/00319/2020 and SusTEC (LA/P/0007/2021). This work has been supported by NORTE-01-0247-FEDER-072598 iSafety: Intelligent system for occupational safety and well-being in the retail sector. Inês Sena was supported by FCT PhD grant UI/BD/153348/2022.

References

1. Pordata. https://www.pordata.pt/portugal, accessed: 2023-04-03
2. Abreu, S.: Automated architecture design for deep neural networks. arXiv preprint arXiv:1908.10714 (2019)
3. Ajayi, A., Oyedele, L., Akinade, O., Bilal, M., Owolabi, H., Akanbi, L., Delgado, J.M.D.: Optimised big data analytics for health and safety hazards prediction in power infrastructure operations. Safety science **125**, 104656 (2020)
4. Anderson, V.P., Schulte, P.A., Sestito, J., Linn, H., Nguyen, L.S.: Occupational fatalities, injuries, illnesses, and related economic loss in the wholesale and retail trade sector. American journal of industrial medicine **53**(7), 673–685 (2010)
5. Belete, D.M., Huchaiah, M.D.: Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results. International Journal of Computers and Applications **44**(9), 875–886 (2022)
6. Beriha, G., Patnaik, B., Mahapatra, S., Padhee, S.: Assessment of safety performance in indian industries using fuzzy approach. Expert Systems with Applications **39**(3), 3311–3323 (2012)
7. Carnero, M.C., Pedregal, D.J.: Modelling and forecasting occupational accidents of different severity levels in spain. Reliability Engineering & System Safety **95**(11), 1134–1141 (2010)
8. Chaipanha, W., Kaewwichian, P., et al.: Smote vs. random undersampling for imbalanced data-car ownership demand model. Communications **24**, D105–D115 (2022)
9. Cherian, S.A., Hameed, A.S.: Numerical modelling of concrete filled frp tubes subjected under impact loading (2017)
10. Fernández-Muñiz, B., Montes-Peón, J.M., Vázquez-Ordás, C.J.: Relation between occupational safety management and firm performance. Safety science **47**(7), 980–991 (2009)
11. Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al.: Array programming with numpy. Nature **585**(7825), 357–362 (2020)
12. Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in science & engineering **9**(03), 90–95 (2007)
13. Kakhki, F.D., Freeman, S.A., Mosher, G.A.: Evaluating machine learning performance in predicting injury severity in agribusiness industries. Safety science **117**, 257–262 (2019)
14. Koc, K., Ekmekcioğlu, Ö., Gurgun, A.P.: Accident prediction in construction using hybrid wavelet-machine learning. Automation in Construction **133**, 103987 (2022)
15. Koc, K., Gurgun, A.P.: Scenario-based automated data preprocessing to predict severity of construction accidents. Automation in Construction **140**, 104351 (2022)
16. Kumar, V., Garg, M.: Predictive analytics: a review of trends and techniques. International Journal of Computer Applications **182**(1), 31–37 (2018)
17. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research **18**(1), 559–563 (2017)
18. Li, H., Liang, Q., Chen, M., Dai, Z., Li, H., Zhu, M.: Pruning smac search space based on key hyperparameters. Concurrency and Computation: Practice and Experience **34**(9), e5805 (2022)
19. Wes McKinney: Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman (eds.) Proceedings of the 9th Python in Science Conference. pp. 56 – 61 (2010). https://doi.org/10.25080/Majora-92bf1922-00a

20. Mendes, J., Lima, J., Costa, L., Rodrigues, N., Brandão, D., Leitão, P., Pereira, A.I.: Machine learning to identify olive-tree cultivars. In: Optimization, Learning Algorithms and Applications: Second International Conference, OL2A 2022, Póvoa de Varzim, Portugal, October 24-25, 2022, Proceedings. pp. 820–835. Springer (2023)
21. Nikbakht, S., Anitescu, C., Rabczuk, T.: Optimizing the neural network hyperparameters utilizing genetic algorithm. Journal of Zhejiang University-Science A **22**(6), 407–426 (2021)
22. Oyedele, A., Ajayi, A., Oyedele, L.O., Delgado, J.M.D., Akanbi, L., Akinade, O., Owolabi, H., Bilal, M.: Deep learning and boosted trees for injuries prediction in power infrastructure projects. Applied Soft Computing **110**, 107587 (2021)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
24. Shirali, G.A., Noroozi, M.V., Malehi, A.S.: Predicting the outcome of occupational accidents by cart and chaid methods at a steel factory in iran. Journal of public health research **7**(2), jphr–2018 (2018)
25. Singh, P., Chaudhury, S., Panigrahi, B.K.: Hybrid mpso-cnn: Multi-level particle swarm optimized hyperparameters of convolutional neural network. Swarm and Evolutionary Computation **63**, 100863 (2021)
26. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems **25**, 1–9 (2012)
27. Zhu, R., Hu, X., Hou, J., Li, X.: Application of machine learning techniques for predicting the consequences of construction accidents in china. Process Safety and Environmental Protection **145**, 293–302 (2021)