Ana I. Pereira · Armando Mendes ·
Florbela P. Fernandes · Maria F. Pacheco ·
João P. Coelho · José Lima (Eds.)

Communications in Computer and Information Science 1981

Optimization, Learning Algorithms and Applications

Third International Conference, OL2A 2023
Ponta Delgada, Portugal, September 27–29, 2023
Revised Selected Papers, Part I

Part 1

Springer

OL2A INTERNATIONAL CONFERENCE

Ana I. Pereira · Armando Mendes ·
Florbela P. Fernandes · Maria F. Pacheco ·
João P. Coelho · José Lima
Editors

Optimization, Learning Algorithms and Applications

Third International Conference, OL2A 2023
Ponta Delgada, Portugal, September 27–29, 2023
Revised Selected Papers, Part I

Springer

Editors
Ana I. Pereira
Instituto Politécnico de Bragança
Bragança, Portugal

Florbela P. Fernandes
Instituto Politécnico de Bragança
Bragança, Portugal

João P. Coelho
Instituto Politécnico de Bragança
Bragança, Portugal

Armando Mendes
University of Azores
Ponta Delgada, Portugal

Maria F. Pacheco
Instituto Politécnico de Bragança
Bragança, Portugal

José Lima
Instituto Politécnico de Bragança
Bragança, Portugal

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-53024-1 ISBN 978-3-031-53025-8 (eBook)
https://doi.org/10.1007/978-3-031-53025-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

The volumes CCIS 1981 and 1982 contains the refereed proceedings of the III International Conference on Optimization, Learning Algorithms and Applications (OL2A 2023), a hybrid event held on September 27–29.

OL2A provided a space for the research community in optimization and learning to get together and share the latest developments, trends and techniques as well as develop new paths and collaborations. OL2A had the participation of more than four hundred participants in an online and face-to-face environment throughout three days, discussing topics associated with areas such as optimization and learning and state-of-the-art applications related to multi-objective optimization, optimization for machine learning, robotics, health informatics, data analysis, optimization and learning under uncertainty and 4th industrial revolution.

Six special sessions were organized under the topics Learning Algorithms in Engineering Education, Optimization in the SDG context, Optimization in Control Systems Design, Computer Vision Based on Learning Algorithms, Machine Learning and AI in Robotics and Machine Learning and Data Analysis in Internet of Things. The event had 66 accepted papers. All papers were carefully reviewed and selected from 172 submissions. All the reviews were carefully carried out by a scientific committee of 115 PhD researchers from 23 countries.

The OL2A 2023 volume editors,

September 2023

Ana I. Pereira
Armando Mendes
Florbela P. Fernandes
Maria F. Pacheco
João P. Coelho
José Lima

Organization

General Chairs

Ana I. Pereira	Polytechnic Institute of Bragança, Portugal
Armando Mendes	University of the Azores, Portugal

Program Committee Chairs

Florbela P. Fernandes	Polytechnic Institute of Bragança, Portugal
M. Fátima Pacheco	Polytechnic Institute of Bragança, Portugal
João P. Coelho	Polytechnic Institute of Bragança, Portugal
José Lima	Polytechnic Institute of Bragança, Portugal

Special Session Chairs

João P. Teixeira	Polytechnic Institute of Bragança, Portugal
José Cascalho	University of the Azores, Portugal

Technology Chairs

Paulo Medeiros	University of the Azores, Portugal
Rui Pedro Lopes	Polytechnic Institute of Bragança, Portugal

Program Committee

Ana Isabel Pereira	Polytechnic Institute of Bragança, Portugal
Abeer Alsadoon	Charles Sturt University, Australia
Ala' Khalifeh	German Jordanian University, Jordan
Alberto Nakano	Federal University of Technology – Paraná, Brazil
Alexandre Douplik	Ryerson University, Canada
Ana Maria A. C. Rocha	University of Minho, Portugal
Ana Paula Teixeira	University of Trás-os-Montes and Alto Douro, Portugal
André Pinz Borges	Federal University of Technology – Paraná, Brazil

André Rodrigues da Cruz	Federal Center for Technological Education of Minas Gerais, Brazil
Andrej Košir	University of Ljubljana, Slovenia
António José Sánchez-Salmerón	Universitat Politècnica de València, Spain
António Valente	University of Trás-os-Montes and Alto Douro, Portugal
Armando Mendes	University of the Azores, Portugal
Arnaldo Cândido Júnior	Federal Technological University – Paraná, Brazil
B. Rajesh Kanna	Vellore Institute of Technology, India
Bilal Ahmad	University of Warwick, UK
Bruno Bispo	Federal University of Santa Catarina, Brazil
C. Sweetlin Hemalatha	Vellore Institute of Technology, India
Carlos Henrique Alves	CEFET - Rio de Janeiro, Brazil
Carmen Galé	University of Zaragoza, Spain
Carolina Gil Marcelino	Federal University of Rio de Janeiro, Brazil
Christopher Expósito Izquierdo	University of Laguna, Spain
Clara Vaz	Polytechnic Institute of Bragança, Portugal
Damir Vrančić	Jožef Stefan Institute, Slovenia
Dhiah Abou-Tair	German Jordanian University, Jordan
Diamantino Silva Freitas	University of Porto, Portugal
Diego Brandão	CEFET - Rio de Janeiro, Brazil
Dimitris Glotsos	University of West Attica, Greece
Eduardo Vinicius Kuhn	Federal Technological University – Paraná, Brazil
Elaine Mosconi	Université de Sherbrooke, Canada
Eligius M. T. Hendrix	Malaga University, Spain
Elizabeth Fialho Wanner	Federal Center for Technological Education of Minas Gerais, Brazil
Felipe Nascimento Martins	Hanze University of Applied Sciences, The Netherlands
Florbela P. Fernandes	Polytechnic Institute of Bragança, Portugal
Florentino Fernández Riverola	University of Vigo, Spain
Francisco Sedano	University of León, Spain
Fredrik Danielsson	University West, Sweden
Gaukhar Muratova	Dulaty University, Kazakhstan
Gediminas Daukšys	Kauno Technikos Kolegija, Lithuania
Gianluigi Ferrari	University of Parma, Italy
Glaucia Maria Bressan	Federal University of Technology – Paraná, Brazil
Glotsos Dimitris	University of West Attica, Greece
Humberto Rocha	University of Coimbra, Portugal
João Paulo Carmo	University of São Paulo, Brazil
João Paulo Coelho	Polytechnic Institute of Bragança, Portugal
João Paulo Teixeira	Polytechnic Institute of Bragança, Portugal

Jorge Igual	Universitat Politécnica de Valencia, Spain
Jorge Ribeiro	Polytechnic Institute of Viana do Castelo, Portugal
José Boaventura-Cunha	University of Trás-os-Montes and Alto Douro, Portugal
José Cascalho	University of the Azores, Portugal
José Lima	Polytechnic Institute of Bragança, Portugal
José Ramos	Nova University Lisbon, Portugal
Joseane Pontes	Federal University of Technology – Ponta Grossa, Brazil
Josip Musić	University of Split, Croatia
Juan A. Méndez Pérez	University of Laguna, Spain
Juan Alberto García Esteban	University de Salamanca, Spain
Júlio Cesar Nievola	Pontifícia Universidade Católica do Paraná, Brazil
Kristina Sutiene	Kaunas University of Technology, Lithuania
Laura Belli	University of Parma, Italy
Lidia Sánchez	University of León, Spain
Lino Costa	University of Minho, Portugal
Luca Davoli	University of Parma, Italy
Luca Oneto	University of Genoa, Italy
Luca Spalazzi	Marche Polytechnical University, Italy
Luis Antonio De Santa-Eulalia	Université de Sherbrooke, Canada
Luís Coelho	Polytechnic Institute of Porto, Portugal
M. Fátima Pacheco	Polytechnic Institute of Bragança, Portugal
Mahmood Reza Khabbazi	University West, Sweden
Manuel Castejón Limas	University of León, Spain
Marc Jungers	Université de Lorraine, France
Marco Aurélio Wehrmeister	Federal University of Technology – Paraná, Brazil
Marek Nowakowski	Military Institute of Armoured and Automotive Technology in Sulejowek, Poland
Maria do Rosário de Pinho	University of Porto, Portugal
Martin Hering-Bertram	Hochschule Bremen, Germany
Matthias Funk	University of the Azores, Portugal
Mattias Bennulf	University West, Sweden
Michał Podpora	Opole University of Technology, Poland
Miguel Ángel Prada	University of León, Spain
Mikulas Huba	Slovak University of Technology in Bratislava, Slovakia
Milena Pinto	Federal Center of Technological Education Celso Suckow da Fonseca, Brazil
Miroslav Kulich	Czech Technical University Prague, Czech Republic
Nicolae Cleju	Technical University of Iasi, Romania

Paulo Alves	Polytechnic Institute of Bragança, Portugal
Paulo Leitão	Polytechnic Institute of Bragança, Portugal
Paulo Lopes dos Santos	University of Porto, Portugal
Paulo Medeiros	University of the Azores, Portugal
Paulo Moura Oliveira	University of Trás-os-Montes and Alto Douro, Portugal
Pavel Pakshin	Nizhny Novgorod State Tech University, Russia
Pedro Luiz de Paula Filho	Federal Technological University – Paraná, Brazil
Pedro Miguel Rodrigues	Catholic University of Portugal, Portugal
Pedro Morais	Polytechnic Institute of Cávado e Ave, Portugal
Pedro Pinto	Polytechnic Institute of Viana do Castelo, Portugal
Roberto Molina de Souza	Federal University of Technology – Paraná, Brazil
Rui Pedro Lopes	Polytechnic Institute of Bragança, Portugal
Sabrina Šuman	Polytechnic of Rijeka, Croatia
Sancho Salcedo Sanz	Alcalá University, Spain
Sandro Dias	Federal Center for Technological Education of Minas Gerais, Brazil
Sani Rutz da Silva	Federal Technological University – Paraná, Brazil
Santiago Torres Álvarez	University of Laguna, Spain
Sara Paiva	Polytechnic Institute of Viana do Castelo, Portugal
Shridhar Devamane	Global Academy of Technology, India
Sławomir Stępień	Poznań University of Technology, Poland
Sofia Rodrigues	Polytechnic Institute of Viana do Castelo, Portugal
Sudha Ramasamy	University West, Sweden
Teresa Paula Perdicoulis	University of Trás-os-Montes and Alto Douro, Portugal
Toma Rancevic	University of Split, Croatia
Uta Bohnebeck	Hochschule Bremen, Germany
Virginia Castillo	University of León, Spain
Vítor Duarte dos Santos	Nova University Lisbon, Portugal
Vitor Pinto	University of Porto, Portugal
Vivian Cremer Kalempa	State University of Santa Catarina, Brazil
Wojciech Giernacki	Poznań University of Technology, Poland
Wojciech Paszke	University of Zielona Gora, Poland
Wynand Alkema	Hanze University of Applied Sciences, The Netherlands
Zahia Guessoum	University of Reims Champagne-Ardenne, France

Contents – Part I

Machine Learning

A YOLO-Based Insect Detection: Potential Use of Small Multirotor Unmanned Aerial Vehicles (UAVs) Monitoring 3
Guido S. Berger, João Mendes, Arezki Abderrahim Chellal, Luciano Bonzatto Junior, Yago M. R. da Silva, Matheus Zorawski, Ana I. Pereira, Milena F. Pinto, João Castro, António Valente, and José Lima

A Comparison of Fiducial Markers Pose Estimation for UAVs Indoor Precision Landing 18
Luciano Bonzatto Junior, Guido S. Berger, Alexandre O. Júnior, João Braun, Marco A. Wehrmeister, Milena F. Pinto, and José Lima

Effect of Weather Conditions and Transactions Records on Work Accidents in the Retail Sector – A Case Study 34
Lucas D. Borges, Inês Sena, Vitor Marcelino, Felipe G. Silva, Florbela P. Fernandes, Maria F. Pacheco, Clara B. Vaz, José Lima, and Ana I. Pereira

Exploring Features to Classify Occupational Accidents in the Retail Sector 49
Inês Sena, Ana Cristina Braga, Paulo Novais, Florbela P. Fernandes, Maria F. Pacheco, Clara B. Vaz, José Lima, and Ana I. Pereira

Resource Dispatch Optimization for Firefighting Using a Differential Evolution Algorithm 63
Marina A. Matos, Rui Gonçalves, Ana Maria A. C. Rocha, Lino A. Costa, and Filipe Alvelos

A Pattern Mining Heuristic for the Extension of Multi-trip Vehicle Routing 78
Leila Karimi, Connor Little, and Salimur Choudhury

Time-Dependency of Guided Local Search to Solve the Capacitated Vehicle Routing Problem with Time Windows 93
Adriano S. Silva, José Lima, Adrián M. T. Silva, Helder T. Gomes, and Ana I. Pereira

Federated Learning for Credit Scoring Model Using Blockchain 109
Daniel Djolev, Milena Lazarova, and Ognyan Nakov

Exploring Features to Classify Occupational Accidents in the Retail Sector

Inês Sena[1,2,3(✉)], Ana Cristina Braga[3], Paulo Novais[3], Florbela P. Fernandes[1,2], Maria F. Pacheco[1,2], Clara B. Vaz[1,2], José Lima[1,2], and Ana I. Pereira[1,2]

[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,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. The Machine Learning approach is used in several application domains, and its exploitation in predicting accidents in occupational safety is relatively recent. The present study aims to apply different Machine Learning algorithms for classifying the occurrence or non-occurrence of accidents at work in the retail sector. The approach consists of obtaining an impact score for each store and work unit, considering two databases of a retail company, the preventive safety actions, and the action plans. Subsequently, each score is associated with the occurrence or non-occurrence of accidents during January and May 2023. Of the five classification algorithms applied, the Support Vector Machine was the one that obtained the best accuracy and precision values for the preventive safety actions. As for the set of actions plan, the Logistic Regression reached the best results in all calculated metrics. With this study, estimating the impact score of the study variables makes it possible to identify the occurrence of accidents at work in the retail sector with high precision and accuracy.

Keywords: Workplace Accidents Classification · Machine Learning algorithms · Score Impact

1 Introduction

Over the years, accidents at work have been the subject of numerous studies to understand, prevent, and reduce them. Among the most adopted strategies in various sectors to fight workplace accidents, investigating these incidents and implementing preventive safety measures stand out [4].

A. I. Pereira et al. (Eds.): OL2A 2023, CCIS 1981, pp. 49–62, 2024.
https://doi.org/10.1007/978-3-031-53025-8_4

These events can occur due to several factors. Several theories in the literature explain the causes of accidents, such as the accident proneness theory, domino theory, injury epidemiology, and macro-ergonomic theory, among others. However, if the causes are known, accidents can be predicted through predictive models that can identify patterns and trends that help to understand the leading causes of accidents at work and to develop effective prevention strategies [17].

Although there is still a need for more information regarding predicting accidents at work, some studies already demonstrate the successful application of Machine Learning techniques in predicting accidents in different business sectors. Ajavi et al. (2020) conducted a study focused on predicting accidents in energy infrastructures, exploring the methods of Particle Swarm Optimization (PSO), Decision Tree, Random Tree, and Gradient Boosting Machine (GBM) [1]. They built four predictive analysis models for the occurrence of accidents and the frequency index, in which the GBM-PSO was the model that presented the best predictive capacity [1].

Another relevant study is Kakhki et al. (2020), who developed a predictive model using Random Forest, Decision Tree, and Naive Bayes methods to predict accidents in agricultural installations, more specifically with grain elevators, achieving an accuracy between 80% and 95% [8].

In addition, other business areas already have studies with the application of Machine Learning methods for the prediction of accidents at work, such as the steel industry [9], construction [10,19], agribusiness [7], among others. It was found, after an extensive bibliographic review, that a business sector that has not carried out studies on this subject is the retail trade sector, which, although it seems to have a low risk of injuries or deaths compared to sectors such as agriculture and construction, is a sector that is involved in a variety of challenging work activities and exposed to various hazards [3,16]. It occupied the second place of economic activity in Portugal during the year 2020, with high records of accidents at work, compared to other sectors [18].

That way, it is essential to emphasize the importance of research in this area since the safety of employees is a priority in any economic activity. However, one of the main problems is the need for more knowledge about which data, variables, or parameters drive the occurrence of an accident. In addition, companies in the retail sector have a large amount of information that can be used to implement models for predicting accidents at work, from accident histories, information about the work environment, and employee demographic data, among others. However, it is necessary to study and analyze the amount and type of data inserted in Machine Learning models since the learning capacity of the model depends mainly on the dataset used.

Thus, this study aims to apply different Machine Learning algorithms in a new database approach to classify the occurrence or non-occurrence of accidents at work in the retail sector. This new approach comprises using only impact scores per database. Two databases were used in this case: preventive safety actions and action plans. Being a large company, it is distributed throughout the country in

different stores, each with varying work units. Thus, the impact score calculation will be based on the number of records for each store and work unit.

To demonstrate the effectiveness of this new approach, the occurrence (1) or non-occurrence (0) of accidents will be associated with the score of each store and unit, taking into account the period from January to May 2023. Subsequently, it intended to apply each set of data to several Machine Learning methods and observe whether it is possible to classify the occurrence of accidents at work in the retail sector through the two designed impact scores.

The paper is organized into four sections. The Sect. 2 presents a discussion of the methodology, in particular, datasets, pre-processing, theoretical concepts of classification algorithms, and the applied performance evaluation metrics. The Sect. 3 aims to compare the achieved results for each data set. Finally, Sect. 4 concludes the study and indicates possible directions for future research.

2 Methodology

This section presents the collected databases and the process of designing the datasets to be used. The pre-processing techniques applied to improve the data quality will also be demonstrated, like the approach to achieve the best results for the listed objective.

2.1 Characterization of Datasets

For the present study, it is intended to use three databases made available by the Portuguese flap company:

- **Accident history**, which contains information about the general characteristics of the injured workers (age, length of service, etc.), the conditions of the accident (place, time, sector, function served at the time of the accident, etc.), the damage caused (severity, type of injury, etc.) and the cause of the accident.
- **Preventive safety actions** are records of risk situations or unsafe conditions observed by members of the Occupational Safety and Health (OSH) team when they visit the stores.
- **Action plans** are drawn up after the intervention of third parties or employees to correct and improve the working conditions observed during an audit of your workplace.

The obligation to create action plans to solve the problems that persist in the employee's well-being and working conditions has existed in the company since 2008, a period to be considered in this study. However, the observation and registration by the company of preventive safety actions to improve employees' conditions and quality of work is relatively recent, being practiced only from August 2022, counting 8681 registrations of action plans and 7757 preventive

security actions. Each record will be associated with the occurrence of accidents through the company's accident history recorded between January 2023 and May 2023.

Considering the number of records in each database, the impact score will be calculated for each store and 16 work units. After obtaining the score, it will be related to the occurrence or not of accidents. It should be noted that, due to the records period, it is only possible to obtain the score for 78 stores, considering the set of preventive security actions, and for 316 stores according to the group of action plans.

For each dataset, the average of records per work unit ($\overline{r}_{uw}$) was considered considering the records per work unit of each store (r_{uws}) and the number of work units (n_{uw}), as shown in Eq. (1):

$$\overline{r}_{uw} = \frac{\sum r_{uws}}{n_{uw}} \tag{1}$$

Subsequently, the score (X) was calculated for each data set, in which the number of records was counted for each store and work unit, and the Eq. (2) was applied:

$$X = \frac{n_{uw} - \overline{r}_{uw}}{n_{uw}} \tag{2}$$

The values obtained were arranged using min-max character scaling, a normalization approach that scales the character to the fixed range of [0,1], as shown in Eq. (3).

$$X_{score} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

Then, each store and work unit's X_{score} was linked to the history of accidents, associating the occurrence or non-occurrence of an accident, the store, and the corresponding work unit. Since not all the stores had accidents during this period, it was only possible to obtain the impact of observation on the occurrence of accidents in 71 stores for preventive safety actions and 283 stores for action plans.

Finally, the datasets used in this study will have the information of store, work unit, and impact score and an output parameter (accident (1) or not an accident (0)).

2.2 Pre-processing Data

Data pre-processing is a significant initial step when using Machine Learning (ML) algorithms to classify events or information. This is because the data quality directly impacts the model's learnability.

Some issues that may affect learning the desired model can be identified by closely examining the datasets. These problems include an imbalance between output variables and data typology. Therefore, it is necessary to apply pre-processing techniques to deal with these issues before using the data in the developed model.

The balance of the data set is crucial to improve the performance of ML algorithms since it is essential that each class has the same number of samples and, thus, has equal relevance in the analysis, avoiding any bias [23]. In this research, the data related to the non-occurrence of accidents is significantly higher than in the other classes. This can lead the algorithms to favor this class and generate predictions with high values but inaccurate and distorted. Therefore, it is essential to use confusion matrices to validate the prediction results.

In this way, the Synthetic Minority Oversampling Technique (SMOTE) technique will be applied, whose main objective is to increase the number of minority samples by inserting n synthetic minority samples among the k samples that are closest to a given sample with lower dimension [23].

The one-hot encoding technique was applied since Machine Learning algorithms tend to obtain better results when dealing with numerical data [14], and datasets contain categorical typology information. This technique consists of removing the categorical variable and dividing it into n new binary variables, depending on the number of categories in the data.

2.3 Machine Learning Techniques

To find out if an impact score can replace the information from a database, it is intended to calculate the correlation coefficient between the two variables, applying different ML algorithms – to conclude if obtaining a relationship with the accident event is possible.

The strength of the association varies between -1 and $+1$, indicating a strong relationship between variables. If the correlation coefficient is near 0, it represents a weaker relationship between variables [2].

Different types of coefficients can be calculated, such as Pearson, Spearman, Kendall Tau, and others. The selection of the coefficient to be calculated depends on the data type. In this case, as the data sets have a non-Gaussian distribution, the Kendall-Tau Correlation was chosen, which is a non-parametric method that measures the association between two variables X, Y based on ratings of sampled observations from X and Y [21]. In addition, it should be used when the same classification is repeated many times on a small dataset, as in the present case [2] (Table 1).

Table 1. Interpretation of the correlation coefficient values based [2].

Correlation Coefficients	Relation Interpretation
[-1, -0.9[or]0.9, 1]	Perfect
[-0.9, -0.7[or]0.7, 0.9]	Strong
[-0.7, -0.4[or]0.6, 0.4]	Moderate
[-0.3, -0.1[or]0.3, 0.1]	Weak
[-0.1, 0[or]0.1, 0]	None

To identify the occurrence of accidents, it is intended to apply and compare different classification algorithms to understand whether it is conceivable to predict the occurrence of accidents through the impact score. Thus, the following algorithms were used:

- **Decision Tree (DT)**, is a versatile Supervised Learning algorithm applicable for both classification and regression and for categorical and continuous dependent variables [23]. Its main objective is to develop a tree structure that identifies the values of test samples through training samples [13]. It is easy to understand and interpret and is often used to support decision-making since each branch represents a choice between several alternatives. Each node represents a [11] decision. This algorithm uses a recursive partitioning technique, building a decision tree composed of several nodes created and divided based on specific criteria. This process is interrupted when the training dataset is adjusted to the predictions [1].
- **K-Nearest Neaighbor (KNN)**, classifies an observation by analyzing the k nearest [23] observations. The algorithm uses the nearest neighbor technique to assign a classification to a new sample point based on its proximity to a set of previously classified points [14]. This method involves two main parameters, the value of k and the distance function. The value of k is determined through several executions with different values, selecting the one that minimizes the number of errors found and provides greater forecast accuracy. The distance function used by KNN is the Euclidean distance, which represents the distance physics between two-dimensional points [14].
- **Random Forest (RF)**, is a popular Machine Learning approach that uses multiple independent decision trees that are built from previously selected variables [1]. Each decision tree is trained using a portion of the original training data. It performs the divisions considering only a random subset of the input variables. The final categorization is defined through the classifier's output that obtains the most votes from the trees [1,23].
- **Logistic Regression (LR)**, is an algorithm used for classification capable of estimating discrete values based on a set of dependent variables [23]. It calculates the probability of an event by fitting the data into a logical function. As a result, the algorithm's output is always between 0 and 1 [23].
- **Support Vector Machine (SVM)**, is a widely used Machine Learning algorithm with a solid theoretical basis, which seeks to find a hyperplane that separates the training data into different classes, maximizing the margin between them [13]. Different kernels, such as linear, RBF, and sigmoid, can be applied for this task. However, proper training parameters are essential to ensure satisfactory prediction accuracy. In general, SVM is recognized for its effectiveness in classifying binary sentiments and its ability to deal with classification and regression problems, outperforming many other statistical and ML methods [13].

To implement the datasets in the five referred algorithms, it is necessary to divide them into training (70%) and testing (30%). To evaluate the performance

of each applied algorithm, specific metrics per class were used, including accuracy, precision, recall, and F_{score}. To calculate them, it is necessary to identify the following:

- True Positives (TP) - data that were accidents and the model predicted as an accident.
- True Negatives (TN) - non-accident samples that the model correctly projected as non-accident.
- False Positives (FP) - data representing accidents and the model projected as non-accidents.
- False Negatives (FN) - accidents samples and the model predict as a non-accident.

Accuracy is a metric widely applied in problems of this nature. It returns a general value of how much the model is correctly predicting the class concerning the entire data set (as defined in Eq. (4)) where:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

The *Precision*, defined in Eq. (5), refers to the model's reliability when correctly predicting a specific class [5].

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

The *Recall* measures the number of true positives that were classified correctly, using Eq. (6), [5].

$$Recall = \frac{TP}{TP + TN} \tag{6}$$

The F_{score} is the harmonic mean of *Precision* and *Recall*, as can see in Eq. (7), which reaches its best value at one and its worst at zero [5].

$$F_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

The referred metrics are based on the confusion matrix generated for each algorithm. The confusion matrix for the specific case study is based on the occurrence of accidents [5].

All the algorithms presented in this work were tested, trained, and implemented on a computer equipped with an 11th Gen Intel(R) Core(TM) i7-1185G7 processor, with a RAM 16 GB memory and Python version 3.10.6 in Google Colab. For this study, different libraries were used, such as the Numpy library (version 1.22.4) [6] and Pandas (version 1.5.3) [20] for efficient data manipulation and analysis, the Scipy library (version 1.10.1) [22] to calculate the correlation, the Imbalanced-learn library (version 0.10.1) [12] for data imbalance, and finally, the Scikit-learn library (version 1.2.2) [15], also known as sklearn, for creating robust predictive models and evaluating performance through appropriate metrics.

3 Results

In this section, the results obtained from the relationship between the impact score and the occurrence of accidents will be presented, as well as the values of the performance evaluation metrics obtained by each iteration of the classification algorithms applied to the dataset. These results enable an understanding of whether it is possible to predict the occurrence of accidents through the calculated impact score.

3.1 Preventive Security Actions

As previously mentioned, this database is recent in the company, containing a reduced period of recording information. However, the impact score of recording preventive safety actions was calculated between August and December 2022 and connected to accidents from January 2023 to May 2023. Therefore, the dataset has three input variables (store identification, identification of the work unit for each store, and impact score) and an output variable (accident or not an accident), as shown in Table 2.

Table 2. Distribution of data by output variables.

Predict Label	Number of data
Not an accident	536
Accident	84

As mentioned in Sect. 2.3, the Kendall-Tau coefficient calculates the association between impact score and crash occurrence. The value 0.0984 refers to a fragile association between them. However, it reveals a positive association, referring to a level of agreement between the impact score (X_{score}: continuous variable) and the occurrence of accidents (0 and 1).

Although there is a fragile association between the variables, the possibility of predicting the occurrence of accidents through the impact score was studied. In this way, different classification algorithms were applied with five cross-validations, and the performance results can be analyzed in Table 3.

Through Table 3, it is possible to observe high accuracy values. However, it is worth noting the lack of assertiveness in identifying the occurrence of accidents. Thus, the algorithms that achieved the highest accuracy values, RF and SVM, could not classify any accidents. However, the remaining algorithms were able to identify a small number of cases. The run time of the prediction models tested was 0.17 s.

The inability of the algorithms to predict the occurrence of accidents can be justified due to the imbalance of information between the outputs, which may hamper the performance of the models in identifying accidents at work. To this end, the SMOTE method was applied, balancing the data from the minority class according to the majority class [23].

Table 3. Results obtained for preventive security actions by the metrics for each algorithm.

Learning algorithms	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
Logistic Regression (LR)	Not an accident	0.87	0.91	0.89	0.80
Logistic Regression (LR)	Accident	0.23	0.16	0.18	0.80
Decision Tree (DT)	Not an accident	0.86	0.89	0.88	0.78
Decision Tree (DT)	Accident	0.12	0.10	0.10	0.78
Random Forest (RF)	Not an accident	0.86	1.00	0.92	0.86
Random Forest (RF)	Accident	0.00	0.00	0.00	0.86
Support Vector Machine (SVM)	Not an accident	0.86	1.00	0.92	0.86
Support Vector Machine (SVM)	Accident	0.00	0.00	0.00	0.86
K-Nearest Neighbors (KNN)	Not an accident	0.85	0.90	0.88	0.78
K-Nearest Neighbors (KNN)	Accident	0.04	0.03	0.03	0.78

Once the data set was modified, the association between the two variables was calculated again using the Kendall-Tau coefficient, showing an increase in the coefficient value to 0.1422, revealing a weak positive association between the variables. Thus, the prediction tests were repeated for the balanced data set. In Table 4, the results of the metrics calculated to evaluate the performance of each of the tested algorithms with a cross-validation of five are presented. The run time of the prediction models tested was 0.16 s.

Table 4. Results obtained for preventive security actions balanced data by the metrics for each algorithm.

Learning algorithms	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
Logistic Regression (LR)	Not an accident	0.88	0.88	0.88	0.88
Logistic Regression (LR)	Accident	0.88	0.88	0.88	0.88
Decision Tree (DT)	Not an accident	0.88	0.73	0.80	0.82
Decision Tree (DT)	Accident	0.77	0.90	0.83	0.82
Random Forest (RF)	Not an accident	0.93	0.65	0.77	0.80
Random Forest (RF)	Accident	0.73	0.95	0.83	0.80
Support Vector Machine (SVM)	Not an accident	0.88	0.89	0.88	0.88
Support Vector Machine (SVM)	Accident	0.89	0.88	0.88	0.88
K-Nearest Neighbors (KNN)	Not an accident	0.90	0.65	0.75	0.79
K-Nearest Neighbors (KNN)	Accident	0.73	0.93	0.81	0.79

Observing Table 4, the increase in assertiveness in detecting the occurrence of accidents is notorious, in addition to the increase in accuracy values, in almost all the cases, except Random Forest.

The SVM and LR were the ones that obtained the best results, and the Support Vector Machine was more assertive in identifying the occurrence of accidents, maintaining the accuracy of the LR in detecting non-occurrence. On the other hand, the RF showed the highest precision in the non-occurrence of accidents and the lowest for the classification of accidents, like the KNN.

3.2 Actions Plans

In Table 5, the distribution of the amount of data of the output variables of the dataset that relates the impact score of action plans and the occurrence of accidents is presented.

Table 5. Distribution of data by output variables.

Predict Label	Number of data
Not an accident	1795
Accident	174

The relationship between the impact score and the output variable was also calculated using the Kendall-Tau coefficient for this data set. Thus, a correlation of 0.0131 was obtained, a fragile association between the variables.

However, the occurrence of accidents was also identified based on the impact score of the action plans registered by the company through the implementation of the data set in the different selected classification algorithms. In Table 6, the results obtained by the metrics after cross-validation of five can be seen. The run time of the prediction models tested was 1.58 s.

Table 6. Results obtained for actions plan data set by the metrics for each algorithm executed.

Learning algorithms	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
Logistic Regression (LR)	Not an accident	0.92	0.94	0.93	0.87
Logistic Regression (LR)	Accident	0.19	0.15	0.16	0.87
Decision Tree (DT)	Not an accident	0.91	0.96	0.94	0.88
Decision Tree (DT)	Accident	0.15	0.08	0.10	0.88
Random Forest (RF)	Not an accident	0.91	1.00	0.95	0.91
Random Forest (RF)	Accident	0.28	0.02	0.03	0.91
Support Vector Machine (SVM)	Not an accident	0.91	1.00	0.95	0.91
Support Vector Machine (SVM)	Accident	0.00	0.00	0.00	0.91
K-Nearest Neighbors (KNN)	Not an accident	0.91	0.98	0.94	0.89
K-Nearest Neighbors (KNN)	Accident	0.18	0.04	0.07	0.89

Observing Table 6, the high accuracy values achieved by the tested algorithms are noted. However, the little assertiveness in the classification of "Accident" is noticeable. In this case, the RF and the SVM were the ones that achieved the highest accuracy values. However, the Support Vector Machine could not identify any accidents occurring.

Once again, this lack of ability to predict an accident can be explained by the sharp difference in data between the two possible outputs. In this way, SMOTE was also applied to this set, which allows for increasing the number of data representing the occurrence of accidents depending on the size of the non-occurrence class.

The previous calculations were repeated, starting with the Kendall-Tau correlation coefficient, which increased to 0.026, keeping the fragile association between the two variables. Subsequently, the occurrence of accidents was classified using the same algorithms and five cross-validations. The results obtained by the metrics can be seen in Table 7. The run time of the prediction models tested was 1.24 s.

Table 7. Results obtained for actions plan balanced dataset by the metrics for each algorithm executed.

Learning algorithms	Predict Label	*Precision*	*Recall*	F_{score}	*Accuracy*
Logistic Regression (LR)	Not an accident	0.92	0.93	0.93	0.92
Logistic Regression (LR)	Accident	0.93	0.92	0.92	0.92
Decision Tree (DT)	Not an accident	0.91	0.65	0.76	0.79
Decision Tree (DT)	Accident	0.73	0.93	0.82	0.79
Random Forest (RF)	Not an accident	0.96	0.51	0.66	0.74
Random Forest (RF)	Accident	0.66	0.98	0.79	0.74
Support Vector Machine (SVM)	Not an accident	0.90	0.93	0.92	0.92
Support Vector Machine (SVM)	Accident	0.94	0.90	0.91	0.92
K-Nearest Neighbors (KNN)	Not an accident	0.92	0.73	0.81	0.83
K-Nearest Neighbors (KNN)	Accident	0.78	0.94	0.85	0.83

Observing Table 7, one can indicate a high increase in the accuracy of the algorithms in classifying the occurrence of accidents, maintaining assertiveness in identifying non-occurrence. Random Forest was the algorithm that reached the highest precision values in class "not an accident" prediction and the lowest for class "accident" identification, compared to the other algorithms, reaching the most insufficient precision. SVM and LR were the ones that achieved the best accuracy results, with Logistic Regression standing out in the remaining evaluation metrics.

The attempt to classify the occurrence of accidents by merging the two scores is worth noting since they are complementary information. However, it was impossible due to the information discrepancy in the two databases.

4 Conclusions and Future Works

This study aimed to calculate an impact score for the two databases that identify risk situations and unsafe conditions in the company's work areas and, consequently, to understand whether, through the score obtained, it is possible to predict the occurrence of accidents on work.

To this end, the impact score was obtained by calculating the average of the records for each dataset per store and work unit. An impact score was obtained through the annual records and then normalized using the min-max feature scaling that establishes a scale of values between 0 and 1 depending on the maximum and minimum values obtained.

By achieving the impact score for each database, a connection was made with the history of accidents between January and May 2023, associating the occurrence or not of accidents at work for each store and work unit with an impact score reached.

With the creation of the two datasets, the Kendall-Tau coefficient was calculated to understand the association between the impact score and the occurrence of accidents. A fragile association was observed between the set of action plans and the event of the accident, and a weak positive association between the impact score of preventive safety actions and the occurrence of accidents.

Afterward, the occurrence of accidents was classified through the achieved impact scores. First, the original datasets were used. However, the results could have been better for their classification due to the data imbalance between the two output variables. In this way, the SMOTE technique was applied, which allows for increasing the information of the minority class as a function of the majority class, equalizing the amount of data for each category.

Thus, five classification algorithms were applied to observe whether it is possible to identify the occurrence of accidents through the impact scores obtained in this study. For the data set of preventive safety actions, the Support Vector Machine was the most assertive in identifying the occurrence of accidents, maintaining high values of accuracy and precision in the classification of the non-occurrence of accidents. As for the action plan dataset, the Logistic Regression algorithm reached the best results in all analyzed metrics.

For both cases, Random Forest was the algorithm that obtained the best precision in predicting class "not an accident" but the lowest values when identifying class "accident".

Thus, it can be concluded that it is possible to classify the occurrence of accidents at work in the retail sector through the impact score obtained by the records of preventive safety actions and action plans carried out by the safety and health team leader and the company's employees.

For future work, it is intended to explore these results further and test this approach in other company databases to find other impact scores that can predict the occurrence of accidents at work in the retail sector.

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. Ajayi, A., et al.: Optimised big data analytics for health and safety hazards prediction in power infrastructure operations. Saf. Sci. **125**, 104656 (2020)
2. Akoglu, H.: User's guide to correlation coefficients. Turkish J. Emerg. Med. **18**(3), 91–93 (2018)
3. 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. Am. J. Ind. Med. **53**(7), 673–685 (2010)
4. Cioni, M., Savioli, M.: Safety at the workplace: accidents and illnesses. Work Employ Soc. **30**(5), 858–875 (2016)
5. Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756 (2020)
6. Harris, C.R., Millman, K.J., et al,: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep2020). https://doi.org/10.1038/s41586-020-2649-2,https://doi.org/10.1038/s41586-020-2649-2
7. Kakhki, F.D., Freeman, S.A., Mosher, G.A.: Evaluating machine learning performance in predicting injury severity in agribusiness industries. Saf. Sci. **117**, 257–262 (2019)
8. Kakhki, F.D., Freeman, S.A., Mosher, G.A.: Applied machine learning in agromanufacturing occupational incidents. Procedia Manufact. **48**, 24–30 (2020)
9. Koc, K., Ekmekcioğlu, Ö., Gurgun, A.P.: Accident prediction in construction using hybrid wavelet-machine learning. Autom. Constr. **133**, 103987 (2022)
10. Koc, K., Gurgun, A.P.: Scenario-based automated data preprocessing to predict severity of construction accidents. Autom. Constr. **140**, 104351 (2022)
11. Kumar, V., Garg, M.: Predictive analytics: a review of trends and techniques. Int. J. Comput. Appl. **182**(1), 31–37 (2018)
12. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18**(17), 1–5 (2017), https://jmlr.org/papers/v18/16-365.html
13. Liu, Y., Bi, J.W., Fan, Z.P.: Multi-class sentiment classification: the experimental comparisons of feature selection and machine learning algorithms. Expert Syst. Appl. **80**, 323–339 (2017)
14. Oyedele, A., et al.: Deep learning and boosted trees for injuries prediction in power infrastructure projects. Appl. Soft Comput. **110**, 107587 (2021)
15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
16. Putz Anderson, V., Schulte, P.A., Novakovich, J., Pfirman, D., Bhattacharya, A.: Wholesale and retail trade sector occupational fatal and nonfatal injuries and illnesses from 2006 to 2016: Implications for intervention. Am. J. Ind. Med. **63**(2), 121–134 (2020)

17. Sanmiquel, L., Rossel, J.M., Vintró, C.: Study of Spanish mining accidents using data mining techniques. Saf. Sci. **75**, 49–55 (2015)
18. dos Santos (FFMS), F.F.M.: Pordata. https://www.pordata.pt/portugal Accessed Jan 6 2023
19. 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. J. Public Health Res.**7**(2), jphr-2018 (2018)
20. pandas development team, T.: pandas-dev/pandas: Pandas (Feb 2020). https://doi.org/10.5281/zenodo.3509134, https://doi.org/10.5281/zenodo.3509134
21. Valencia, D., Lillo, R.E., Romo, J.: A kendall correlation coefficient between functional data. Adv. Data Anal. Classif. **13**, 1083–1103 (2019)
22. Virtanen, P., et al.: SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
23. Zhu, R., Hu, X., Hou, J., Li, X.: Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Saf. Environ. Prot. **145**, 293–302 (2021)