Integrated feature selection and classification algorithm in the prediction of work-related accidents in the retail sector: A comparative study

Inês Sena[1,2,3][0000−0003−4995−4799], Laires A. Lima[1,2][0000−0002−3094−3582], Felipe G. Silva[1,2][0000−0002−3612−9645], Ana Cristina Braga[3][0000−0002−1991−9418] Paulo Novais[3][0000−0002−3549−0754], Florbela P. Fernandes[1,2][0000−0001−9542−4460], Maria F. Pacheco[1,2][0000−0001−7915−0391], Clara Vaz[1,2][0000−0001−9862−6068], José Lima[1,2][0000−0001−7902−1207], and Ana I. Pereira[1,2][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 para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
{ines.sena, laires.lima,gimenez,fflor, pacheco, clvaz, jllima, apereira}@ipb.pt,
[3] ALGORITMI Centre, University of Minho, 4710-057 Braga, Portugal
acb@dps.uminho.pt, pjon@di.uminho.pt

Abstract. Assessing the different factors that contribute to accidents in the workplace is essential to ensure the safety and well-being of employees. Given the importance of risk identification in hazard prediction, this work proposes a comparative study between different feature selection techniques (χ^2 test and Forward Feature Selection) combined with learning algorithms (Support Vector Machine, Random Forest, and Naive Bayes), both applied to a database of a leading company in the retail sector, in Portugal. The goal is to conclude which factors of each database have the most significant impact on the occurrence of accidents. Initial databases include accident records, ergonomic workplace analysis, hazard intervention and risk assessment, climate databases, and holiday records. Each method was evaluated based on its accuracy in the forecast of the occurrence of the accident. The results showed that the Forward Feature Selection-Random Forest pair performed better among the assessed combinations, considering the case study database. In addition, data from accident records and ergonomic workplace analysis have the largest number of features with the most significant predictive impact on accident prediction. Future studies will be carried out to evaluate factors from other databases that may have meaningful information for predicting accidents.

Keywords: Feature selection · Classification algorithms · Accident prediction.

1 Introduction

In 2019, according to Eurostat, in the EU, there were 3.1 million non-fatal accidents, with a higher incidence in the manufacturing industry (18.7% of the total in the EU in 2019), wholesale and retail trade (12.3%), construction (11.8%) and human health and social assistance activities (11.0%). Spain, France, and Portugal were the EU Member States with the highest number of non-fatal accidents per 100 000 employed persons [9].

According to the statistics of accidents at work listed in the Pordata database, in Portugal, the tertiary sector is the primary economic activity sector that contributes to the high number of accidents in the workplace, being the sector of "wholesale and retail, repair of motor vehicles and motorcycles" the main contributor to the number of accidents at work [18].

Occupational accidents and diseases impact operations and costs for companies, workers, and society, decisively affecting the quality of life and the economy. Thus, it is evident the loss of productive capacity, in terms of lost working days, as well as the compensation and pensions to be paid to a large extent by companies. Each non-fatal accident in 2019 in the EU resulted in at least four days of absence from work, which implies a cost between 2.6% and 3.8% of the Gross Domestic Product (GDP) [6].

Then, the health and safety issues in the workplace should be a priority in social and economic policies since security-related concerns are empirically integrated into work performance [1]. The most common indicator of lack of safety in the workplace is the occurrence of accidents [2]. As such, work-related accidents have been the subject of several studies over the years, with the development of theories that try to explain, prevent and reduce them. The most used actions to combat occupational accidents in other sectors are investigating accidents and implementing preventive measures, for example, reducing the excessive workload [5] and providing information and training on Occupational Safety and Health (OSH) to workers [6], among others.

To study hazards, risks, and accidents at work, the other sectors use descriptive statistic tools or tools based on conventional statistical techniques [16], which consist of a historical summary based on detailed information collected about the circumstances of an accident at work. However, there are already different methods of analysis and identification of hazards and risks that help prevent workplace accidents, for example, the ergonomic assessment in the workplace and the hazards identification and risk assessment.

Some studies indicate the ergonomic conditions of a company's workplace as influential factors in the quality of work, stress and physical exhaustion of the employee, and the occurrence of accidents [10, 13, 14]. Ergonomics, therefore, plays an important role in identifying risks, developing and applying reliable assessment tools, methodologies, and techniques, and defining preventive recommendations to help reduce exposure risks, improve work organization, and design workplaces. adequate work [3]. This evaluation is performed by an analyst and checks the postures and movements adopted by an employee performing his/her duties, by calculating different methods, from RULA, REBA, among others.

Bearing this in mind, it is essential to examine the ergonomic conditions in the retail sector due to the high proportion of manual work in handling heavy goods and physiologically unfavorable postures that can trigger future musculoskeletal injuries to employees, which start losses in the sector in terms of money, time and productivity.

Another method for preventing and reducing accidents is the identification of hazards and risk assessment, which investigates situations in the existing activities of an organization that can be harmful to the health and safety of employees.

Risk assessment is the basis for preventing accidents at work and occupational health problems, as it aims to implement the required measures (such as adequate information and training for workers) to improve the health and safety of workers, as well as, in general, contributing to improving the performance of companies [8, 22].

However, with the evolution of Artificial Intelligence (AI) techniques, companies must invest in new technologies to improve the safety of employees, and consequently, their productivity and safety. A predictive analytics solution based on AI has already been implemented in several areas, such as the construction industry [19], manufacturing [11], among others, for these purposes, and consequently, in the reduction of injuries in the workplace.

An example of use, is the application of various data-mining techniques to model accident and incident data obtained from interviews carried out shortly after the occurrence of an incident and/or accidents to identify the most important causes for the occurrence of accidents in two companies in the mining and construction sectors, and consequently, develop forecast models that predict them [19]. Another example is the development of a model, through the Random Forest algorithm, to classify and predict the types of work accidents at construction sites using the importance of characteristics and analyzing the relationship between these factors and the types of occupational accidents (OA) [11].

Thus, predictive analytics can analyze different data about the causes of OA and find connecting factors. However, there are few studies aimed at determining the range of factors that give rise to an accident, their interactions, and their influence on the type of accident. Hence, it is challenging to design the kind of well-founded accident prevention policy that is the fundamental objective of safety systems—security management [16].

Taking into account the factors enumerated above, the main objective of the present work is to acquire the features that have the most significant impact on the occurrence of accidents, using different and large databases from a retail company under study. The secondary aim of this research is to compare feature selection methods and classification algorithms for the prediction of workplace accidents. In order to achieve the listed goals, five pre-processed databases (accident records, ergonomic workplace analysis, hazard identification and risk assessment, climate database, and holiday records) will be combined. Thus, two selection methods – χ^2 test and Forward Feature Selection - will be compared and combined with the three selection algorithms prediction - Support

Vector Machine, Random Forest, and Naive Bays – in order to get the relevant features and the selection method-algorithm prediction pair that offer the best accuracy.

The rest of the paper is organised as follows. The methods used during the preparation are reported in Section 2. Section 3 presents the database as well as the pre-processing techniques that were used. The methodology that was followed to achieve the listed objectives as set out in Section 4. In Section 5, the results obtained are presented and discussed. Finally, the study is concluded, and future work is presented in Section 6.

2 Methods

To achieve the specified objectives and find the features that obtain the best precision in accident prediction, two methods were selected, and three Machine Learning classification algorithms were established for comparison. The feature selection methods and the prediction algorithms were used and will be described below.

2.1 Feature Selection

A common problem in Machine Learning is determining the relevant features to predict a specific output while maintaining a good model performance. Feature selection is essential because not all variables are equally informative. Selecting an appropriate number of features can directly lead to a better fit, a more efficient and straightforward predictive model, and better knowledge concerning how they interact, which are fundamental factors in the correct interpretation of the problem [4]. Although the use of feature selection techniques implies an additional complexity added to the modeling process, it has the benefits of decreasing processing time and predicting associated responses [4]. In the case of classification problems where the input and output variables are categorical and numerical, it is advisable to use a suitable statistical test to handle the heterogeneity of the data to determine the level of relationship between the variables. This step also simplifies the database, considering the high number of features present in the global database. The most significant features in accident prediction were selected by the χ^2 test and Forward Feature Selection (FFS). These methods were chosen because they have different approaches and for comparison purposes. The output variable adopted was the cause of the accident, divided into 12 categories:

- Physical aggression.
- Hit by object, equipment, or particles.
- Collision (between two or more vehicles).
- Contact with sharp or sharp material agent.
- Object/Environment contact - hot or cold.
- Contact with hazardous substances (gases, vapors, liquids).

- Welding (machines, equipment).
- Blow/Shock against equipment/fixed structure.
- Sting (insects/fish).
- Fall at the same level.
- Fall from height.
- Over-effort.

The χ^2 test enables to test the independence between input and output variables, whether numerical or categorical, based on observed and expected values. The null hypothesis (H_0) of the χ^2 test states that the two variables i and j, in a contingency table, are independent while the alternative hypothesis (H_1) states that these variables are not independent. The χ^2 statistic is calculated according to:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

where O_{ij} is termed as the observed frequency, and E_{ij} is termed as the expected frequency for each pair of variables i an j classified in r and c categories, respectively [21]. If H_0 is true, Q follows the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom; and when two variables i and j are independent, the actual and expected values are similar, implying a low Q value and indicating that the hypothesis of independence (H_0) between them is not rejected [21]. For the prediction model, the input variables must be strongly related to the output variable.

Forward Feature Selection (FFS) is an alternative in selecting relevant features, a wrapper method through its implementation of Forward Feature Selection. FFS corresponds to an iterative method that starts with an empty set. In each iteration, the feature that improves the accuracy is added until the addition of a new variable does not improve the performance of the model. In other words, the method is based on the search and definition of subgroups of features that achieve the best accuracy, as illustrated in Figure 1.

2.2 Predictive Algorithms

Support Vector Machine (SVM) is a supervised Machine Learning algorithm that can be used for both regression and classification challenges, aimed at train and classifying a database. It does not limit data distribution and is mainly used for small samples [21]. SVM aims to find a hyperplane in an n-dimensional space (with n being the number of features) that distinctly classifies the data points, as shown in Figure 2.

There are infinite possibilities for the hyperplane that separates two groups of data, so the algorithm seeks to find a plane with a maximum margin, which is described by the maximum distance between the points of both groups. The database points can be written as $(\vec{x_i}, y_i), ..., (\vec{x_n}, y_n)$ [17], where the vector $\vec{x_i}$ can be represented by -1 or 1. On the other hand, the hyperplane is described by the Equation (2) [17]:

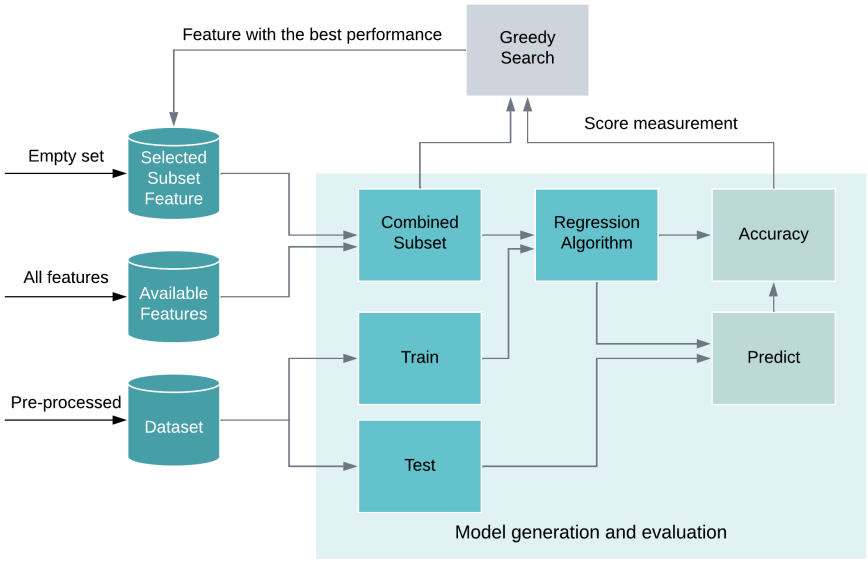

Fig. 1. Functioning of Forward Feature Selection technique. Adapted from [20].

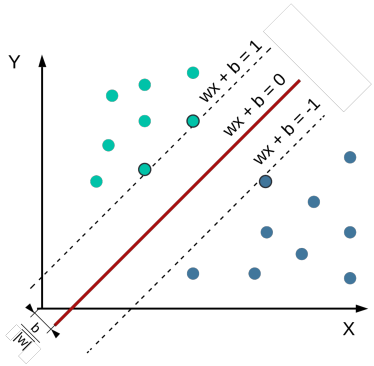

Fig. 2. Support Vector Machine classifier. Adapted from [17].

$$\vec{w}\vec{x} - b = 0 \tag{2}$$

The normal vector is represented by $\vec{w}$ while the offset of hyper place along $\vec{w}$ is by $\frac{b}{||\vec{w}||}$, more details can be seen in [17]. Maximizing the margin distance allows future data points to be sorted more confidently. Support vectors are data

points closest to the hyperplane and influence the position and orientation of the hyperplane. The classifier margin can be maximized using these vectors, and the optimal plane can be drawn.

Random Forest (RF) is composed of Decision Trees formed by a data sample extracted from a training set with replacement, called a bootstrap sample. Of this training sample, one-third of it (out-of-bag sample) is reserved as test data. Once more, a group of random data is injected through feature clustering, adding more diversity to the database and reducing the correlation between the decision trees. The determination of the forecast is based on majority voting, that is, the selection of the most frequent variable. Finally, an out-of-bag sample is used for cross-validation [12]. Figure 3 shows the typical structure of Random Forests.

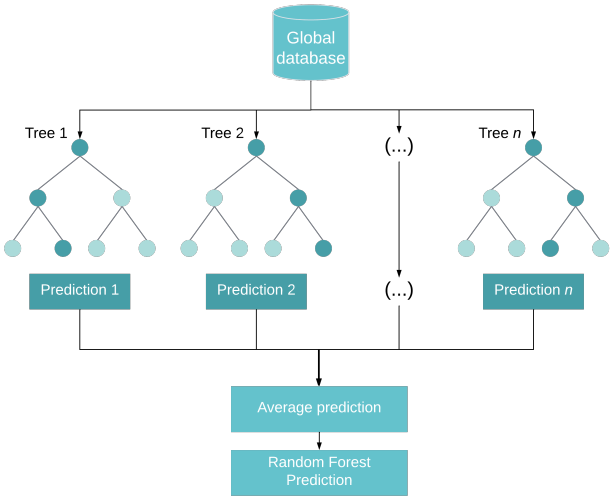

Fig. 3. Random Forest classifier. Adapted from [12].

Random Forest algorithms start with three primary hyperparameters, namely the node size, the number of trees, and the number of features sampled, defined before the training step. From there, the random forest classifier can be used to solve regression or classification problems [12].

Naive Bayes (NB) is a simple probabilistic classifier based on the application of Bayes' theorem [17]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

where A and B are the events (or features), $P(A|B)$ is the probability of A given B is true, $P(B|A)$ is the probability of B given A is true, and $P(A)$ and $P(B)$ are the independent probabilities of A and B.

In basic terms, a Naive Bayes classifier assumes that each feature independently contributes to the probability of the output variable [17]. The conditional probability was calculated considering that the global database has numerical and categorical data. Although Naive Bayes does not account for resource interactions, it is efficient in predicting classification problems. The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables needed for classification. As the independent variables are not specified, only the variances of the variables for each label need to be determined and not the entire covariance matrix [7].

3 Case Study

This section examines and describes how the database was developed and its characteristics. In addition, the pre-processing techniques used to adapt the database to the algorithms' attributes to improve their performance will also be described.

3.1 Database

The database used comprises information from five different databases:

- Accident records – contain information on the general characteristics of the injured employees (age, seniority, etc.), the conditions of the accident (place, time, sector, a task performed at the time of the accident, etc.), the damage caused (severity, type of injury, etc.) and the cause of the accident.
- Ergonomic Workplace Analysis (EWA) – includes the values calculated in an analysis of the postures and movements adopted by employees performing their duties by calculating different methods from RULA, REBA, among others.
- Hazard Identification and Risk Assessment (HIRA) – is composed of risk levels associated with the detail of the micro tasks in each work section.
- Climate database – includes weather data from across the country, since 2019 and 2021 automatically extracted from external sources, more information at [20].
- Holiday records – comprises registration of all national and municipal holidays that occurred between 2019 and 2021 in Portugal, automatically extracted from external sources, see in [15].

Data from accident records, EWA, and HIRA, were collected between 2019 and 2021 from different supermarkets of a company in the retail sector. Thus, the global database was developed through the integration of the five databases mentioned above, resulting in a total of 128 instances or records, totaling 122 input factors and 12 output factors after removing record lines with missing information or out of context. In Figure 4, the method of association of the five databases can be observed, since they are integrated by the common factors which are the microtask (accident records, EWA, HIRA) and the event date (accident records, climate database, and holiday records).

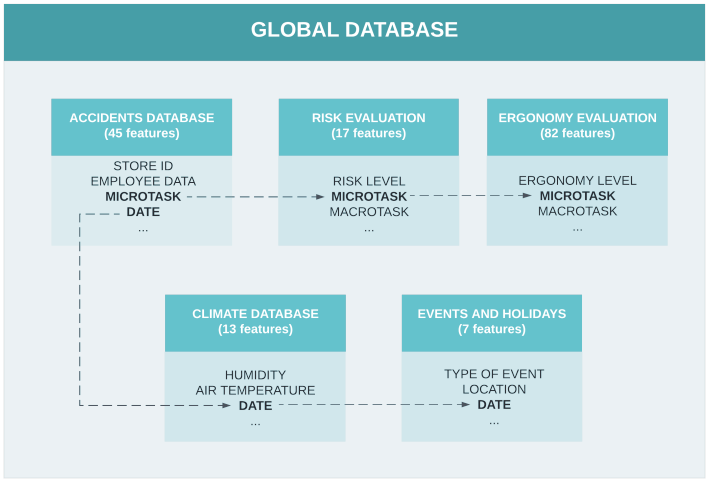

Fig. 4. Methodology adopted for the preliminary integration of simple databases in the formation of the global database.

3.2 Database Pre-processing

The original database was encoded in numerical form since many Machine Learning algorithms do not operate directly on the label data, requiring the input and output variables to be numeric. In other words, categorical data in the global database were converted into a numerical form, more precisely using One Hot Encoding. In this method, each feature is identified with 0 and 1 according to the presence or absence of the feature. This pre-processing step allows the binarization of the categories. Thus, they can be included in the training of the models without assigning weight, as in the case of a simple label encoding.

Then, the global database was normalized in order to increase the learning algorithms' efficiency. Once applied, database normalization ensures that all values are in the range between zero and one.

4 Methodology

The methodology for developing supervised Machine Learning models for accident prediction based on the historical database of a retail company is shown in Figure 5. A detailed description of each step is discussed.

For the execution of the χ^2 test, the steps adopted were the definition of the null and alternative hypotheses, the creation of a contingency table, the calculation of expected values, calculation of the χ^2 statistic, and acceptance or rejection of the null hypothesis. The significance level considered was 0.10, and the features were selected based on the level of correlation with each target.

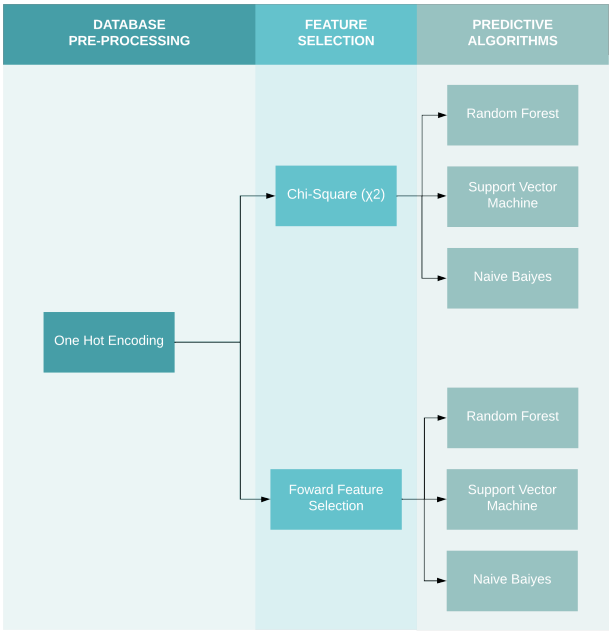

Fig. 5. Methodology for building classification algorithms to predict accidents in the retail sector.

In contrast to the χ^2 test, FFS relies on learning machine methods to select the significant features. Thus, the method becomes more computationally expensive but has better chances of assertiveness. This technique builds feature subgroups with the highest accuracy, which was performed according to the steps in Figure 1. Since some algorithms (like SVM and NB) are not prepared to predict multiple targets and the database used is composed of 12 different accident causes (targets), the procedure described was carried out for each target (that is, when a target was selected one, the rest were set to zero) to increase the certainty in the selection of features and the performance of the learning algorithms.

The FFS method was executed multiple times for each search algorithm. The FFS feature selection process is usually not deterministic, and the search strategy influences the result. Thus, the repetition of the method sought to identify the most commonly selected features to avoid specific scenarios, atypical subgroups, and overfitting. The FFS method chose up to 10 features for each target. The selection of features and quantities per target was decided according to the average performance obtained by the input features.

The predictive model bias and the average accuracy are the selected features considered to perform the quality analysis of the models generated by learning algorithms using the selected features.

The process was performed with unbalanced data since the global database contains only real data. Strategies to simulate data were not used, so 'false' data did not influence the feature selection process

Predictive models can present biased predictions created by unbalanced data and still have high accuracy. To optimize accuracy and reduce the number of input features, the FFS can opt for a predictive model which always uses the most dominant label as the final answer. Due to this factor, the metric (4) more sensitive to unbalanced databases was developed considering the percentage of hits of each label.

$$accuracy = \frac{1}{L} \sum_{i=1}^{L} \frac{C(L_i)}{T(L_i)} \tag{4}$$

where L is the number of labels, C is the number of correct predictions of a label L_i, and T is the total number of occurrences of a label L_i.

The metric was used to guide feature choices. The implemented FFS uses the greedy search algorithm as a search strategy, which determines the best feature according to a previously chosen metric. Thus, the models were optimized regarding their quality and not for the total number of hits, seeking the set of features capable of simultaneously creating models with the lowest prediction bias and better accuracy.

5 Results and Discussion

Taking this procedure into account, a better understanding of the results involved the separation into:

- True Positive (TP) – indicates the number of times the model was assertive in the selected target.
- True Negative (TN) – represents the number of times the learning algorithm detected the non-corresponding causes of the accident, that is, the non-occurrence of an accident.
- False Positive (FP) – shows the number of times the model missed the selected target.
- False Negative (FN) – reveals the number of times the model did not identify the accident.
- Final accuracy – shows the overall score for each learning algorithm.

These metrics were calculated because they differentiate assertiveness from precision value. Since the TP and TN metrics enable the identification of the predictive trend about impact, the FP and FN metrics allow identifying the predictive trend about model errors. Table 1 shows the results obtained for each combination tested.

Observing Table 1, it is notable that the FFS selection method obtained greater absolute precision combined with the different classification algorithms, which indicates that by using this selection method, it is possible to get more

Table 1. Results obtained by the adapted confusion matrix and accuracy tests in terms of TP, TN, FP and FN, and final accuracy. The values consider the average calculated based on 1000 executions of the algorithm.

Parameters		**Results (% correct answers in 1000 tests)**				
Feature Selection	**Classification Algorithm**	**TP**	**TN**	**FP**	**FN**	**Final Accuracy**
χ^2	SVM	0.1233	**0.9967**	**0.0033**	0.0797	0.9239
FFS	SVM	**0.5078**	0.9875	0.0125	**0.0447**	**0.9475**
χ^2	RF	0.2523	**0.9945**	**0.0055**	0.0679	0.9328
FFS	RF	**0.7000**	0.9828	0.0172	**0.0272**	**0.9592**
χ^2	NB	0.5639	0.8407	0.1593	0.0396	0.8176
FFS	NB	**0.7436**	**0.9537**	**0.0463**	**0.0233**	**0.9361**

general hits and maximize the accuracy of less common labels. From all combinations with the FFS, Random Forest had the highest precision value.

Analyzing in more detail this table, it can be noted that the absolute precision of the pairs involving the χ^2 method is also high, however, an imbalance of results between the TP and TN metrics is identified, indicating little predictive assertiveness in the type of target, so the high precision value is not satisfactory for the final objective of the present study.

From these metrics, the FFS-NB pair also seems to be a good solution for accident prediction since it was more assertive in the type of target than the FFS-RF pair and only has a difference of 2.31% of ultimate precision. Through Table 1, it was possible to obtain that the best selection method is the FFS. Table 2 presents the number of features selected from each database in each prediction test with classification algorithms.

Table 2. Number of variables selected from each of the five databases using different combinations of Forward Feature Selection and classification algorithms.

Databases	**Classification algorithms combined with FFS**		
	Support Vector Machine	Random Forest	Naive Bayes
Accidents records	33	39	39
EWA	25	23	22
Climate database	9	11	10
Holiday records	6	5	5
HIRA	3	3	4
Total	**76**	**81**	**80**

Observing Table 2, it can be mentioned that:

– SVM is the classification algorithm that obtains the smallest number of features, but in terms of precision (see Table 1), it is the one that gets the lowest value of absolute accuracy and assertiveness in the type of target.

– Unanimously, the accident registration and ergonomic workplace analysis databases have the greatest number of significant features for predicting the causes of accidents used.

Once again, the difference between RF and NB is minimal, just one feature differentiating them. Thus, it can be mentioned that using the combination of FFS-RF and FFS-NB leads to a good prediction of accidents. However, taking into account what was intended with the present study (identifying the best features for accident prediction), the FFS-NB pair presents better results since it obtained a smaller number of features, it was more assertive in accident identification (> TP) and was faster in getting the results than FFS-RF. Although FFS-RF had a higher total precision, it was less assertive in predicting the accident, which does not fit with what is intended.

6 Conclusions and Future Work

This work presented a comparative study in which some of the standard methods of feature selection (χ^2 and FFS) and learning/classification (SVM, RF, and NB) were applied to accident prediction. These methods were used to collect five databases (obtained from a leading retail company in Portugal) and compared their performance in predicting the causes of accidents. The pre-processing used, One Hot Encoding, was crucial for the correct functioning of the learning algorithms since it made the training data more expressive and machine-readable and, consequently, improved the accuracy of the classification algorithms compared to the database without treatment.

Regarding feature selection methods, it was possible to observe that, since the FFS is a non-deterministic method, it can obtain different output results at each execution. This obstacle was overcome by repeating the FFS multiple times. In contrast, χ^2 presents a consistent result and lower computational cost but ignores the type of learning model in which the data would be used.

Comparing different attribute selection methods combined with prediction algorithms shows that the FFS-NB pair obtained fewer features, was more assertive in identifying the accident (> TP), and was faster in getting results than FFS-RF. However, the FFS-RF presented better performance considering absolute precision. Thus, considering the study's purpose, the FFS-RF pair is the most suitable. Nevertheless, both pairs can be considered good options, so it will be necessary to confirm this information through new tests, like sensitivity testing and validation.

Taking into account the selection of features, it can be stated that it is unanimous that the features with the most impact on accident prediction belong to the databases: accident records and ergonomic workplace analysis.

In summary, the results revealed how an integrated feature selection and classification algorithm could be a viable tool in identifying the most likely causes of an accident in the workplace, in addition to the fact that the pre-treatment of the databases and the choice of algorithms and rating metrics can substantially influence the rating success.

Further studies include new databases with a potential relationship with the occurrence of accidents, the improvement of the algorithms presented in this work, and the possible testing of new algorithms.

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) 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. Antão, P., Calderón, M., Puig, M., Michail, A., Wooldridge, C., Darbra, M.R.: Identification of occupational health, safety, security (ohhs) and enviromental performance indicators in port areas. Safety Science **85**, 266–275 (2016)
2. Beus, J.M., McCord, A.M., Zohar, D.: Workplace safety: A review and research synthesis. Organizational psychology review **4**, 352–381 (2016)
3. Capodaglio, E.M.: Occupational risk and prolonged standing work in apparel sales assistants. International Journal of Industrial Ergonomics **60**, 53–59 (2017)
4. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A.: A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing **408**, 189–215 (9 2020)
5. Cioni, M., Sabioli, M.: A survey on semi-supervised feature selection methods. Work Employment and Society **30**, 858–875 (2016)
6. Commission, E.: Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions (2012), https://www.eea.europa.eu/policy-documents/communication-from-the-commission-to-1, last accessed 20 January 2022
7. Dukart, J.: Basic concepts of image classification algorithms applied to study neurodegenerative diseases. Brain Mapping pp. 641–646 (2015). https://doi.org/10.1016/B978-0-12-397025-1.00072-5
8. Encarnação, J.: Identificação de perigos e avaliação de riscos nas operações de carga e descarga numa empresa de tratamento e valorização de resíduos. Ph.D. thesis, Escola Superior de Tecnologia do Instituto Politécnico de Setúbal (2014)
9. Explained, E.S.: Accidents at work statistics, https://ec.europa.eu/
10. Garcia-Herrero, S., Mariscal, M.A., Garcia-Rodrigues, J., Ritzel, O.D.: Working conditions, psychological/physical symptoms and occupational accidents. bayesian network models. Safety Science **50**, 1760–1774 (2012)
11. Kang, K., Ryu, H.: Predicting types of occupational accidents at construction sites in korea using random forest model. Safety Science **120**, 226–236 (2019)
12. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2** (2002)
13. Loske, D., Klumpp, M., Keil, M., Neukirchen, T.: Logistics work, ergonomics, and social sustainability: Empirical musculoskeletal system strain assessment in retail intralogistics. Logistics **5**, 89 (2021)

14. López-García, J.R., Garcia-Herrero, S., Gutiérrez, J.M., Mariscal, M.A.: Psychosocial and ergonomic conditions at work: influence on the probability of a workplace accident. Safety Science **5** (2019)
15. Martins, D.M.D., Silva, F.G., Sena, I., Lima, L.A., Fernandes, F.P., Pacheco, M.F., Vaz, C.B., Lima, J., Pereira, A.I.: Dynamic extraction of holiday data for use in a predictive model for workplace accidents. In: Second Symposium of Applied Science for Young Researchers - SASYR (*In Press*) (2022)
16. Matías, J.M., Rivas, T., Martin, J.E., Taboada, J.: Workplace safety: A review and research synthesis. International Journal of Computer Mathematics **85**, 559–578 (2008)
17. Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C., Mohammed, I.A.: Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. SN Computer Science **2**, 11 (2 2021)
18. Pordata: Acidentes de trabalho: total e por sector de actividade económica., https://www.pordata.pt
19. Rivas, T., Paz, M., Martin, J.E., Matías, J.M., García, J.F., Taboada, J.: Explaining and predicting workplace accidents using data-mining techniques. Reliability Engineering & System Safety **96**, 739–747 (2011)
20. Silva, F., Sena, I., Lima, L., Fernandes, F.P., Pacheco, M.F., Vaz, C., Lima, J., Pereira, A.I.: External climate data extraction using the forward feature selection method in the context of occupational safety. In: Computational Science and Its Applications - ICCSA 2022. *In* Lecture Notes in Computer Science (*In Press*). Springer (2022)
21. Trivedi, S.K.: A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in Society **63**, 101413 (11 2020)
22. Wang, Y., Jin, Z., Deng, C., Guo, S., Wang, X., Wang, X.: Establishment of safety structure theory. Safety Science **115**, 265–277 (2019)