Labor Accidents Classification Using Machine Learning

Enádio da Silva Barbosa[1][a], Yandre M. G. Costa[1][b], Juliano H. Foleis[2] and Diego Bertolini[2][c]
[1] *Department of Information Tchnology, State University of Maringá — DIN/UEM, Paraná, Brazil*
[2] *DACOM, Universidade Tecnológica Federal do Paraná (UTFPR), Paraná, Brazil*

Keywords: Labor Accidents, Work Accidents, Occupational Accidents, Machine Learning, Classification.

Abstract: The application of artificial intelligence is increasingly growing in all public and private industry fields. In this work, we propose applying machine learning techniques to perform work accident classification according to Brazilian laws. The type of accident is part of the communication of occupational accidents (CAT) database held by the National Institute of Social Security. In Brazil, that communication can come from different sources. Because of this, some of them lack the type of work accident. This information is crucial to allow labor authorities to understand better the circumstances surrounding the accidents and to help plan and create more specific strategies to prevent them. In this sense, we perform data cleaning, and we use feature engineering techniques to treat data from CAT database. Following, we use machine learning algorithms aiming to perform the classification according to the type of accident. For this, we investigate a strategy to identify the type of labor accident when this information is missing using algorithms based on trees or gradient boosting trees. Preliminary results showed promising performance, where the algorithms achieved the following weighted average F1-score for labor accident types classification: XGboost 0.94, CAtboost 0.94, Lightgbm 0.94, and Random Forest 0.91. As far as we know, work accident type classification using machine learning, considering Brazilian labor legislation and a huge governmental dataset is addressed for the first time in this work.

1 INTRODUCTION

The International Labour Organization (ILO) (ILO, 2021) estimates that around 2.39 million workers die annually around the world due to labor-related accidents. Approximately 340 million labor accidents[1] happen worldwide yearly, meaning a great social and financial cost for the respective countries. In Brazil, according to SmartLab Observatory of Occupational Safety and Health of Public Labour Prosecutor (MPT) (MPT, 2020), between 2018 and 2020, the average of labor accidents reached 569,998 per year. This shows the seriousness of the situation and how important is to build better statistical information to understand the problem and to propose solutions that can be effective to deal with it.

Labor accidents may have many different causes, most of which could be avoided (Alli, 2008). The causes vary from insufficient protective measures in the working environment, exhausting working schedules, lack of adequate rest, and many workers being exposed to excessive extra working hours. According to Brazilian law number 8,213 of 1991 (PBPS, 1991), labour accidents can be classified into three types:

- Typical - represents the labor accident that occurs while the worker performs their duties.
- Displacement - represents the labor accident that occurs while the worker is going to or leaving the workplace.
- Labour Illness - all illnesses suffered by workers caused or related to the tasks they perform while working.

Still, according to the same law, labor accidents must be communicated to public authorities. This communication is essential to help the federal government and the Labour Inspectorate, in particular, build more knowledge about the overall picture of work accidents in the country and how they occur. It is also important as a prof for workers to request governmental assistance.

[a] https://orcid.org/0000-0002-3597-1961
[b] https://orcid.org/
[c] https://orcid.org/0000-0002-6196-4538
[1] In this work, ‘labor accidents’, ‘work accidents’, and ‘occupational accidents’ refer to the same concept.

Barbosa, E., Costa, Y., Foleis, J. and Bertolini, D.
Labor Accidents Classification Using Machine Learning.
DOI: 10.5220/0011856500003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 509-516
ISBN: 978-989-758-648-4; ISSN: 2184-4992

In Brazil, the labor accident is communicated through a form known as CAT (acronym for "Comunicação de Acidente de Trabalho", in Portuguese). It can be sent by different entities (e.g., people or institutions). For instance, it can be sent by the company where the accident happened, by the respective employee's union, by any public authority, or even by the worker himself or his family. The possibility of many sources sending CATs helps increase the number of communications sent. However, it also causes a problem of missing information. For example, many CATs have the type of accident missing. It makes it difficult for the labor authorities to understand the accidents fully.

As it is difficult to deal with the consequences of work accidents, prevention is crucial because it can avoid all related suffering and financial costs. Beyond reducing the risks of damage to the worker's health, which is by far the most important one, prevention also allows the companies' activities to keep going without interruptions. Another reason is that avoiding accidents from happening prevents the companies from suffering sanctions from the government, whether it becomes evident that the accident was caused because the company failed to regard the safety and security measures instated by law. Therefore, it is crucial to have all information, especially the type of accident, to understand the whole picture and move straight toward building better strategies and concrete actions to minimize labor accidents.

Machine Learning (ML) is an area of artificial intelligence that allows computers to learn from data (Abu-Mostafa et al., 2012). It is part of an interdisciplinary scientific field that combines computer science, data science, and statistics through complex data analysis, searching for patterns in the data. There are four types of ML algorithms(Burkov, 2019): Supervised, semi-supervised, unsupervised, and reinforcement.

Supervised learning algorithms perform prediction or classification for new data samples based on knowledge obtained from examples previously presented to the algorithm in the training phase. These algorithms can improve themselves as much data are analyzed over time, and more predictions and classifications are made. This happens because the algorithms feed themselves with information about the process to help correct their own mistakes. In this work, the hypothesis is that ML can use data related to labor accidents to classify the types of labor accidents. Therefore, this work proposes to use Machine Learning to create a model that uses supervised learning algorithms to predict the missing type of accident in the CATs. It can help the Brazilian labor inspectorate improve its service to the general public. Thus, the following research question (RQ) is raised in this scenario: Is it possible to apply ML algorithms to classify the types of labor accidents using the CAT database as input data?

The remaining of this paper is organized as follows: Section 2 presents some works in any way close to this one available in the literature. Section 3 describes the Machine Learning techniques that support the development of this research. Section 4 introduces the CAT database and the study design. Section 5 presents results and discussions. And finally, we describe concluding remarks and future works.

2 RELATED WORKS

(Di Noia et al., 2020) used machine learning strategies in the context of work accidents, predicting occupational disease risks using computational intelligence and pattern recognition techniques. They used real data about the worker and the workplace from the Italian Health Authority (ASL). Three machine learning algorithms were experimented: K-means, Support Vector Machines, and K-Nearest Neighbors. In summary, the authors obtained encouraging results using artificial intelligence approaches to create an alternative for occupational disease risk prediction.

(Shkanov, 2019) investigated Multiclass Classifiers for Processing Archives of Accidents in Manufacturing. This work compared the best techniques for preprocessing labor accident data. The data comprised 1,600 acts from the archive of accidents in Russia's metallurgy industry. Each act of accident is presented in free textual form and includes three parts: Event description, analysis of reasons, and recommendations. They chose and made adjustments to the best classification methods for these data and used two classifiers to group 19 classes related to the causes of accidents and another to group 39 classes of recommendations. The preprocessing was performed in three steps: text normalization, filtering, and parametrization. They used the following algorithms in the classification phase: Logistic Regression, Naive Bayes, Random Forest e Gradient Boasting. The results classifying the reasons for accidents showed an accuracy of 79% for Random Forest, 82% for Gradient Boosting, and 84% for Logistic Regression. In comparison, the results for classifying the recommendations showed an accuracy of 63% for Random Forest, 64% for Gradient Boosting, and 66% for Logistic regression.

Another related work uses the Random Forest model to predict occupational accidents at construction sites in Korea. In this work, (Kang and Ryu, 2019) analyzed the question of labor accidents in the Korean construction industry. The data is composed from Korea Occupational Safety and Health Agency (KOSHA). The KOSHA's dataset includes 9,796 accident reports in construction sites from 2008 to 2014. They have 55 input variables such as age, occupation injury, occupational accident, and others. The occupation accident types were classified into nine classes and set up as output values. They also gathered weather data from Korea Meteorological Agency (KMA) to include, in the accident dataset, temperature, humidity, wind speed, and precipitation. Initially, they generated and analyzed derived variables comparing the occurrence of accidents against temperature, season, and wind chill changes. They used the technique of determining feature importance that permits creating a ranking of features regarding their contribution to the overall result of the model. To classify the types of labor accidents, they used the Random Forest CART (Classification and Regression Tree) algorithm that resulted in an accuracy of approximately 71%.

(Suárez Sánchez et al., 2011) investigated the prediction of work-related accidents according to working conditions using the Support Vector Machines classifier. The database was composed of the responses to a 78 variables questionnaire applied to a total of 11,054 workers in Spain, carried out between December 2006 and April 2007. The target was to determine if a worker has suffered or not an accident in the last year. Then, they employed SPPCA (Semi-parametric principal component analysis) to reduce the dimensions of the feature vectors. The algorithm MARS (Multivariate adaptive regression splines) was evaluated, and it was able to reduce the number of features from 78 to 18. The features pointed with the most significant importance by MARS were used as input to feed the SVM classifier. Hence, based on work conditions, they were able to classify the workers that suffered and those that did not suffer occupational accidents in twelve months with an accuracy of 99.77%.

In (Suárez Sánchez et al., 2016), the authors applied K-Nearest Neighbor to classify workers according to their risk of musculoskeletal disorders. In that work, they dealt with the binary classification of workers developing or not musculoskeletal disorders caused by occupational tasks during work. The input database was composed of the responses to a 78 variables questionnaire applied to a total of 11,054 workers in Spain, carried out between December 2006 and April 2007.

(Sarkar et al., 2019) dealt with the prediction of occupational accidents and extracting decision-making rules from labor accident data. The database had 3308 incident records and 16 features (15 categorical and one textual). The dataset generated after a preprocessing step had 1500 instances and 13 attributes. The authors used SVM and ANN (Artificial Neural Network) to perform classification. For parameters optimization, they used GA (genetic algorithm) and PSO (particle swarm optimization) algorithms to reach the model's higher accuracy and robustness. The SVM algorithm achieved a higher performance with higher accuracy and robustness. Hence, a set of nine rules were extracted to identify the root causes of wounds, circumstances where workers were barely hit, and property damage. To achieve these results, they used data from a steel plant.

3 GRADIENT BOOSTING TREES (GBT)

In this section, we briefly describe Gradient Boosting Trees (GBT) algorithms, used for the development of the experiments presented in this work.

In this work, we deal with tabular data as inputs for the models. The literature shows that GBT is among the most used methods for modeling discrete or tabular data (Feng et al., 2018). (Shkanov, 2019) also corroborates that, as GBT was also successfully used in that work.

As pointed by (Freund and R, 1997), Gradient Boosting algorithms combine weak learners into strong learners in an iterative way. The objective of gradient boosting is to find an approximation function $F(x)$ that can map instances of x to their output values, by minimizing a given loss function for a specific set of training data $TD = (x_i, y_i)_1^N$. It makes global convergence of the algorithm by following the direction of the negative gradient. The Gradient boosting builds an additive approximation of $F(x)$ as a weighted sum of functions as shown in Equation 1.

$$F_m(x) = F_{m-1} + p_m h_m(x) \tag{1}$$

In this case, p_m is the weight of the m^{th} function, $h_m(x)$. These functions are the models of the ensemble (i.e. decision trees), and the approximation is constructed iteratively. As stated in (Zhou, 2012), ensemble learning is a technique that tries to construct a set of learners and combine them by boosting weak learners that are just slightly better than random. The ensembled learner has better generalization ability and can make very accurate predictions.

GBT algorithms are based on trees and, as explained, use some methods that try to create a robust predictor based on the combination of less efficient ones (Yuichiro, 2012; James et al., 2013; Kubat, 2017). These methods are well known as committees. When considering the result of the committee of algorithms, the prediction tends to compensate for the individual errors of each predictor (Kubat, 2017). The first adaptative boosting algorithm proposed in the literature was AdaBoost (Freund, 1997). Since this moment, they evolved up to the arrival of Gradient boosting algorithms (Bentejac, 2020).

GBT produces models sequentially in a form of a linear combination of decision trees, working in an infinite dimensional optimization problem (Biau, 2019). Boosting is an ensemble strategy that works by dividing the training data and using each part to train different models or one model with different setups. Then the results are combined together using majority vote (Daoud, 2019). They use a stage-wise approach and the loss function to avoid overfitting. It happens by training learners based on minimizing the differential loss function of a weak learner using a gradient descent optimization process.

Concerning the GBT algorithms, in this work, we experiment and compare the results of XGboost, Lightgbm, and CATboost. Each one of them holds some peculiarities that will be explored below as explained by (Daoud, 2019).

XGboost was developed by (Chen and Guestrin, 2016) as a scalable machine learning system. It differentiates itself, mainly from the other gradient boosting, because it adds a new term to the loss function to deal with overfitting.

According to (Ke, 2017), Lightgbm uses XGBoost as a baseline, but executes the classification problem through the combined application of the following two techniques:

- Gradient-based One-Side Sampling - The model omits the majority of examples where the Gradient weight is expected to be smaller, helping going into branches with more importance for the information gain.
- Exclusive Feature Bundling - Reduces features sparsity by bundling them together and reducing their total number, helping decrease training time.

As stated in (Prokhorenkova, 2019), Catboost introduces two new functions, one for handling categorical features and the other is an ordered permutation-driven boosting. It handles the issue of exponential feature combination growth by using a greedy method at every new split of the current tree.

The growing development, utilization, and flexibility of ML puts this technique in a good position to solve the specific problem of this work. In the case of this research, machine Learning supervised classification algorithms will be used to determine the types of labor accidents that are missing in the available CATs. Among a great number of algorithms, GBT shows in the literature a good performance using data with the same characteristics as the CAT's data. As an example, we can cite (Shkanov, 2019), in which Gradient Boosting and Linear Regression obtained the best performances when compared to Random Forest and Naïve Bayes.

4 MATERIALS AND METHODS

This section introduces the main characteristics of the dataset used for developing this work. In addition, we also describe the methodological design of the proposed solution.

4.1 CAT Database

In this section, we describe the fields selected to compose the CAT database and some details to help understand its particularities and usefulness in the development of this work. As part of the work, we selected the attributes to compose the database also taking into account tax secrecy concerns. In addition, we have discarded some columns, originally present in the data, in which we noticed too many null or spurious values. The database produced contains 76,017 instances and a total of 30 attributes. This data was collected from 2019 to 2022.

Table 1 shows the fields of the CAT database used in the experiments described here. As we can see, the data basically contains relevant information associated with the accident occurrences.

4.2 Study Design

In this section, we describe the study design flow, according to the steps shown in Figure 1, which start from selecting the data to be used as input for the classification model, followed by the classification phase, and final results.

Firstly, we analyzed the CAT database to better understand the overall quality of the data. In this stage, the aim was to determine if the data had missing, null, or spurious values that could affect the models results. Therefore, we performed data cleaning to eliminate those values from the database to preserve the data integrity. In this direction, we eliminated blanks and dots in some column names.

The next stage was to perform feature engineering. Firstly, we executed feature transformation by

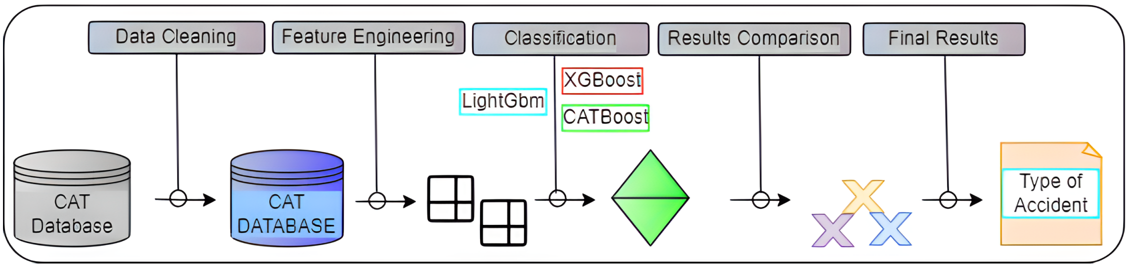

Figure 1: Graphical Abstract - Study Design Flow.

calculating the age of each worker using their date of birth. As the database had some categorical values, we executed the encoding method to transform all categorical columns into numeric ones. This method creates a new column for every different value a specific categorical column holds and attributes the value 1 to it.

After performing feature engineering, based on expert knowledge, to define the data composition, we have chosen some of the most used and promising classifiers algorithms described in the ML literature to deal with tabular data. The comparison between these different algorithms in the task investigated here is one of the main objectives of this work.

The classification algorithms selected for this work are based on gradient boosting trees. They were chosen based on the literature considering the dominant method for modeling discrete or tabular data (Feng et al., 2018).

In the following step, we observed the feature importance built in XGBoost. The importance is calculated by the amount that each feature split point improves the performance, weighted by the number of observations the node is responsible for. This creates a ranking that informs how much each feature improves the overall results. Figure 2 shows the Top-10 features that most contribute to the results.

An important step for improving the boosting-based tree algorithms results is finding the best set of hyperparameters. For this, we performed 5-fold cross-validation. It works by exhaustively searching subsets of hyperparameters space of the targeted algorithm to find the best combination that can improve the outcome. For the algorithms used in this work, we describe in Table 2 the hyperparameters found using grid-search.

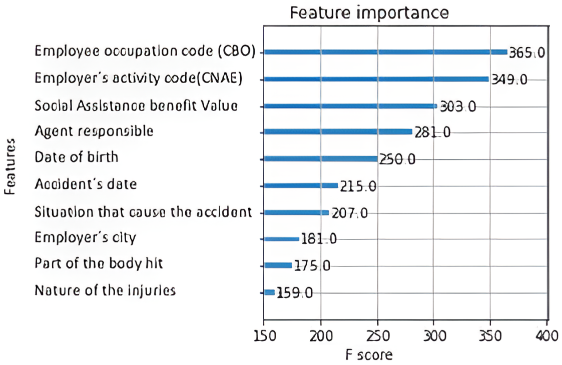

Figure 2: Top-10 Feature Importance.

The next stage was conducted by applying those hyperparameters to tune the machine learning algorithms to perform the classification of the type of accident. Therefore, all algorithms were set up with their respective hyperparameters in order to perform the classification. The final stage was to plot a confusion matrix for each one of the algorithms and compare the results of them, demonstrating which one performed better.

5 RESULTS AND DISCUSSIONS

After the hyperparameters were set up for each algorithm, the classification task was performed. The training and test subset were defined as 80-20, respectively. Table 3 shows the number of accidents per type. As we can see, the data is imbalanced.

These experiments aimed to improve the initial classification to achieve better and more robust results. Furthermore, the aim is also to compare different classifiers' performance and demonstrate which achieves the best results. Therefore, all classifiers received the same data to make it possible to compare their results.

We used some metrics to analyse the performance of the models. Among them we used macro-averaged F1 score (or macro F1 score). It is computed using the arithmetic mean (aka unweighted mean) of all the per-class F1 scores where all classes are treated equally regardless of their support values. Another metric is weighted-averaged F1 score where the calculation is done via mean of all per-class F1 scores while considering each class's support. This Support refers to the number of actual occurrences of the class in the dataset while the 'weight' refers to the proportion of each class's support relative to the sum of all support

Table 1: CAT Database.

#	Field	Description
1	Accident's Date	Date of the accident
2	Accident's estate code	Estate code of the accident
3	CBO	Employee occupation code
4	Employer's CNAE	Employer's activity code
5	Contrator's CNAE	Contractor's activity code
6	SEX	Employee's gender
7	CID-10	International disease code
8	Date of birth	Employee's Date of birth
9	CAT's type	Type of CAT
10	Employer's city	Employer´s city
11	Employer's Estate code	Employer's estate code
12	Employee marital status	Employee marital status
13	Pensioner's Activity	Pensioner's Activity
14	Pensioner's benefit value	Pensioner's benefit value
15	Pensioner's city	Pensioner's city
16	Pensioner's state code	Pensioner's state code
17	Accident local type	Accident local type
18	Part of the body hit	Part of the body hit
19	Agent responsible	Agent responsible for the accident
20	Accident's situation	Situation that cause the accident
21	Nature of the injury	Nature of the injuries
22	Retirement indicator	Retirement indicator
23	Leave Indicator	Leave Indicator
24	Police complaint indicator	Police complaint indicator
25	Hospitalization indicator	Hospitalization indicator
26	Lack of CAT indicator	Lack of CAT indicator
27	CAT issuance indicator	CAT issuance indicator
28	CAT issuance delay indicator	CAT issuance delay indicator
29	Medical leave solicitation indicator	Medical leave solicitation indicator
30	Death indicator	Death indicator

values. The first above mentioned metric is affected by the fact that CAT´s database is unbalanced while the second metric is not affected.

In the first experiment, we used CatBoost Classifier. As done for the experiment previously shown, this model had also to be trained, It was performed with 5-fold cross-validation and hyperparameters set up accordingly with grid-search execution. The hyperparameters were set up as follows: CATBoost - learningrate: 0.2, maxdepth: 5, nestimators: 300.

CatBoost works similarly to XGBoost. In this case, iterations or estimators control the maximum number of trees that the model will have and the depth parameter represents how big the tree is. Table 4 shows the results, the first part of the table shows the classification rates obtained with different metrics for each class, while the second part of the table shows the overall performance.

Another algorithm experimented was LightGbm Classifier. As performed in the previous experimentation, this model had also to be trained. Again, it was performed with 5-fold cross-validation and the parameters were tuned accordingly with grid-search. The parameters were setup as follows: boostingtype': 'gbdt', colsamplebytree: 0.65, 'learningrate': 0.01, nestimators: 8, numclass: 3, numleaves: 6, objective: multiclass, regalpha': 1, reglambda: 1, seed: 500, subsample: 0.7.

LightGbm works, in general, similarly as the previous algorithms. But, performing a Leaf-wise tree growth. In this case, iterations or estimators control the maximum number of trees that the model will have and the depth parameter represents how big the tree is.

Table 5 shows the results for this model. The first part of the table shows the classification metrics for each class while the second part of the table shows the overall performance.

The following algorithm experimented was Random Forest Classifier. As occurred in the previous experimentation, this model had also to be trained and its parameters tuned accordingly with grid-search. The parameters were set up as follows:

Random Forest algorithm also works with a number of decision trees working ensembled as a committee. The fundamental concept of this model is that the trees are relatively uncorrelated, and consequently, the trees may correct the errors each other.

Table 6 shows the results for the model. The first part of the table shows the classification metrics for each class while the second part of the table shows the overall performance.

The last experimentation used XGBoost Classifier, where the best results were achieved. The algorithm was run with 5-fold cross-validation and tuned with the hyperparameters that resulted from grid-search execution. The better hyperparameters found were: Subsample: 0.5; Num classes: 3, nestimators: 100, maxdepth: 6, learningrate: 0.2, colsamplebytree: 0.5, colsamplebylevel: 0.5.

XGBoost works by creating and adding trees level-wise (Daoud, 2019; Chen and Guestrin, 2016) to the model sequentially, in order to correct the residual

Table 2: Hyperparameters.

XGBOOST	CATBoost	Lightgbm	Random Forest
subsample: 0.5	nestimators: 300	subsample: 0.7	nestimators:200
numclass: 3	maxdepth: 5	numclass: 3	maxfeatures: 7
nestimators: 100	learningrate: 0.2	nestimators: 8	minsamples_leaf: 1
maxdepth: 6		boostingtype': gbdt	minsamples_split: 2
learningrate: 0.2		learningrate': 0.01	njobs: 1
colsamplebytree: 0.5		colsamplebytree: 0.65	
colsamplebylevel: 0.5		numleaves: 6	
		objective: multiclass	
		regalpha': 1	
		reglambda: 1	
		seed: 500	
		subsample: 0.7	

Table 3: CAT - Work Accidents per TYPE.

Typical	Displacement	Illness
63,602	10,788	1,627

Table 4: CatBoost Classifier results.

Class	Precision	Recall	F1-score
0 - Illness	0.96	0.97	0.97
1 - Displacement	0.89	0.68	0.77
2 - Typical	0.81	0.80	0.81
Macro avg	0.89	0.82	0.85
Weighted avg	0.94	0.94	0.94

Table 5: LightGBm Classifier results.

Class	Precision	Recall	F1-score
0 - Illness	0.96	0.97	0.97
1 - Displacement	0.90	0.66	0.76
2 - Typical	0.84	0.82	0.83
Macro avg	0.90	0.82	0.85
Weighted avg	0.94	0.94	0.94

Table 6: Random Forest Classifier results.

Class	Precision	Recall	F1-score
0 - Illness	0.92	0.98	0.95
1 - Displacement	0.97	0.50	0.66
2 - Typical	0.84	0.58	0.68
Macro avg	0.91	0.68	0.76
Weighted avg	0.91	0.91	0.91

errors in the predictions from the existing sequence of trees. As the trees grow, the learning rate or shrinkage factor represents how fast the model will learn, meaning how many corrections will be made for each new tree added. The parameter n_estimators stands for the number of estimators or iterations and represents the total number of trees that the model will have and the depth parameter represents how high is the tree.

As shown in Table 7, the XGBoost algorithm presented the best overall performance. Thus, we also present in Figure 3 the confusion matrix obtained using it. The first part of Table 7 shows the rates obtained with different classification metrics for each class while the second part of the table shows the overall performance.

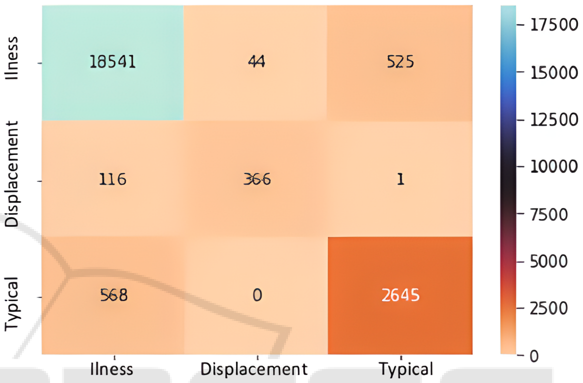

Figure 3: XGBOOST Confusion Matrix.

Table 7: XGBoost Classifier results.

Class	Precision	Recall	F1-score
0 - Illness	0.96	0.97	0.97
1 - Displacement	0.89	0.76	0.82
2 - Typical	0.83	0.82	0.83
Macro avg	0.90	0.85	0.87
Weighted avg	0.94	0.95	0.94

Finally, taking into account the results obtained, we can conclude that the RQ raised in the introduction was positively answered.

6 CONCLUDING REMARKS AND FUTURE WORK

The drama of occupational accidents victims workers causes a negative impact on companies' activities and on the whole of Brazil's economy as well. It also means a greater expenditure of public funds. Hence, there is a huge urgency to search for alternatives to support the creation of preventive actions to reduce the occurrence of occupational accidents.

Although the initiatives from companies aiming at reducing work accidents can be extremely useful, the Brazilian labor inspectorate is definitely the institution capable of proposing alternatives that can be used in the entire country and across all industries.

In this case, machine learning presents itself as a tool that, applied across the communication of labor accidents, has the capacity to automatically classify the types of labor accidents in cases this information is missing. This classification can help the labor inspectorate to create educational and fiscal actions to reduce the problem. However, It is important to notice that this is a very complex problem and several initiatives has to be implemented simultaneously to have a wider impact. Therefore, the initiative proposed in this research can be one of them.

Experiments accomplished on the CAT database showed that XGBoost achieved the best performance for the classification of labor accident type, obtaining 0.87 of Macro avg F1-score, and 0.94 of Weighted avg F1-score.

Future research could focus on other aspects of work accidents. There are many possibilities where machine learning can be used, for instance, to predict work illness and work accidents with fatal outcomes. It is clear that this subject is very important and the development of new researches are welcome to contribute to reducing labor accidents and, therefore, to help create a safer environment for workers across industries and across the globe.

ACKNOWLEDGEMENTS

This research has been partly supported by the Brazilian agencies National Council for Scientific and Technological Development (CNPq) and Brazilian Labor Ministry.

REFERENCES

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York, NY, USA:, New York, USA.

Alli, B. O. (2008). Fundamental principles of occupational health and safety second edition. *Geneva, International Labour Organization*, 15.

Bentejac, C. e. a. (2020). A comparative analysis of gradient boosting algorithms.

Biau, G. e. a. (2019). Accelerated gradient boosting.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov, Canada, 1 edition.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system.

Daoud, E. A. (2019). Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1):6 – 10.

Di Noia, A., Martino, A., Montanari, P., and Rizzi, A. (2020). Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Computing*, 24(6):4393–4406.

Feng, J., Yu, Y., and Zhou, Z.-H. (2018). Multi-layered gradient boosting decision trees. *Advances in neural information processing systems*, 31.

Freund, Y. and R, S. (1997). A short introduction to boosting.

Freund, Y. e. a. (1997). A decision-theoretic generalization of on-line learning and an application to boosting.

ILO (2021). International labour organization world statistic. Available on 25th May 2021.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kang, K. and Ryu, H. (2019). Predicting types of occupational accidents at construction sites in korea using random forest model. *Safety Science*, 120:226–236.

Ke, G. e. a. (2017). Lightgbm: A highly efficient gradient boosting decision tree.

Kubat, M. (2017). *An introduction to machine learning*, volume 2. Springer, Zurich, Switzerland.

MPT (2020). Observatory of occupational safety and health of the public labor prosecutor of brazil (MPT). Available on 25th October 2020.

PBPS (1991). Brazilian labor law nº 8,213, from 24th July 1991. *Brazilian official journal of the union*.

Prokhorenkova, L. e. a. (2019). Catboost: unbiased boosting with categorical features.

Sarkar, S., Vinay, S., Raj, R., Maiti, J., and Mitra, P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, 106:210–224.

Shkanov, B. e. a. (2019). Multiclass classifiers for processing archives of accidents in manufacturing. In *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 1, pages 187–190.

Suárez Sánchez, A., Iglesias-Rodríguez, F., Riesgo Fernández, P., and de Cos Juez, F. (2016). Applying the k-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders. *International Journal of Industrial Ergonomics*, 52:92–99.

Suárez Sánchez, A., Riesgo Fernández, P., Sánchez Lasheras, F., de Cos Juez, F., and García Nieto, P. (2011). Prediction of work-related accidents according to working conditions using support vector machines. *Applied Mathematics and Computation*, 218(7):3539–3552.

Yuichiro, A. (2012). *Pattern recognition and machine learning*. Elsevier.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC.